

ATELIER

QUALITÉ DES DONNÉES ET DES
CONNAISSANCES

2010

JA, LBE, SG

QDC 2010

Actes du 6e Atelier

Qualité des Données et des Connaissances

En conjonction avec EGC 2010

26 Janvier 2010
Hammamet, Tunisie

**Organisé par
Jérôme Azé, Laure Berti Equille et Sylvie Guillaume**

Sixième Atelier

Qualité des Données et des Connaissances

26 janvier 2010, Hammamet, Tunisie

Préface

Après le succès des quatre premières éditions de l'atelier Qualité des Données et des Connaissances en conjonction avec la conférence EGC - 2005 à Paris, 2006 à Lille, 2007 à Namur, 2008 à Nice et 2009 à Strasbourg - nous organisons la sixième édition de l'atelier à l'occasion des journées EGC 2010 à Hammamet en Tunisie.

Cet atelier se concentre sur les méthodes et techniques d'analyse et d'évaluation de qualité au sens large, tant en fouille de données qu'en gestion des connaissances :

- qualité des données (nettoyage, méthodologies de prétraitement, métriques d'évaluation et approches algorithmiques, prise en compte de l'hétérogénéité des données),
- qualité des modèles et de leurs résultats en fouille de données (évaluation des méthodes et algorithmes, études des mesures d'intérêt, agrégation, post-traitement des résultats),
- qualité des connaissances (ontologies, alignements, typologie, visualisation, usages).

La découverte de connaissances et la prise de décision à partir de données de qualité médiocre (*c'est-à-dire contenant des erreurs, doublons, incohérences, valeurs manquantes, ...*) ont des conséquences directes et significatives pour tous les utilisateurs, quelque soit le domaine d'application, gouvernemental, commercial, industriel ou scientifique. Pour cela, le thème de la qualité des données et des connaissances est devenu un des sujets d'intérêt tout à la fois émergent dans le domaine de la recherche et critique dans les entreprises.

Toutes les applications dédiées à l'analyse des données (*telles que la fouille de données textuelles par exemple*) requièrent différentes formes de préparation des données avec de nombreuses techniques de traitement, afin que les données passées en entrée aux algorithmes de fouille se conforment à des distributions relativement "sympathiques", ne contenant pas d'incohérences, de doublons, de valeurs manquantes ou incorrectes. Seulement, entre la réalité des données disponibles et toute la machinerie permettant leur analyse, un assez vaste fossé demeure.

In fine, l'évaluation des résultats issus du processus de traitement des données, est généralement effectuée par un spécialiste (*expert, analyste, ...*). Cette tâche de post-traitement est souvent très lourde et un moyen de la faciliter consiste à aider le spécialiste en lui fournissant des critères de décision sous la forme de mesures de qualité ou d'intérêt des résultats. Ces mesures doivent être conçues afin de combiner deux dimensions : l'une objective liée à la qualité des données, l'autre subjective liée aux intérêts du spécialiste. Bien que les techniques utilisées en fouille de données et en gestion des connaissances soient très différentes, elles partagent

l'objectif de produire des modèles de connaissances pertinents pour les décideurs, avec une préoccupation commune d'évaluation de la qualité des modèles produits.

Cet atelier concerne donc tous les domaines qui participent à la chaîne de production des connaissances : données, méthodes de fouille et gestion des connaissances.

Nous encourageons la soumission d'articles de recherche et/ou d'études de cas industriels liés à tous les aspects de la qualité des données, des méthodes de fouille et de gestion des connaissances au sens large. La durée de l'atelier est d'une demi-journée dédiée à des présentations d'articles dans les thèmes d'intérêt indiqués ci-après :

- Métriques de qualité des données
- Techniques de nettoyage et préparation intelligente des données ; détection de données contradictoires, de données isolées, de doublons, d'incohérences, bruit
- Fouille et découverte de *patterns* de non-qualité ou de qualité médiocre
- Transformations, réconciliation, consolidation des données
- Correction d'erreurs
- Métriques de qualité pour les résultats de fouille ou d'analyse
- Métriques de qualité centrées utilisateurs, mesures subjectives et objectives, mesure d'intérêt des règles
- Validation de modèles de fouille de données
- Post-traitement des résultats
- Qualité de modèles de représentation de connaissances et d'ontologies
- Identification d'objets
- Appariement et alignement d'ontologies
- Application à tout type de données (XML, données transactionnelles, numériques, catégorielles, multimédia, ontologies OWL) dans différents contextes d'application (Bioinformatique, Marketing, e-Commerce, etc)

Jérôme Azé, Laure Berti Equille et Sylvie Guillaume
Organisateurs de QDC 2010

Comités

Comité d'organisation

- Jérôme Azé, LRI, Université Paris-Sud 11
- Laure Berti Equille, IRISA, Université Rennes I
- Sylvie Guillaume, LIMOS, Université d'Auvergne

Comité de programme

- Alexandre Aussem, LIESP, Université Lyon 1
- Jérôme Azé, LRI, Université Paris-Sud 11
- Laure Berti-Equille, IRISA - Université de Rennes 1
- Julien Blanchard, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Marc Boullé, Orange Labs, TEACH/EASY
- Martine Cadot, LORIA, Nancy
- Jean Diatta, IREMA, Université de la Réunion
- Thanh-Nghi Do, Telecom Bretagne
- Sylvie Guillaume, LIMOS, Université d'Auvergne
- Fabrice Guillet, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Ali Khenchaf, Laboratoire E3I2-EA3876, Brest
- Yves Kodratoff, CNRS, LRI, Université Paris-Sud 11
- Stéphane Lallich, ERIC, Université Lumière - Lyon 2
- Ludovic Lebart, CNRS, Département SES, ENST
- Alain Léger, Orange - France Telecom R&D
- Vincent Lemaire, Orange Labs, TEACH/EASY
- Philippe Lenca, Telecom Bretagne
- Israel-César Lerman, IRISA - Université de Rennes 1
- Engelbert Mephu Nguifo, LIMOS, Université Blaise Pascal
- Patrick Meyer, Telecom Bretagne
- Amédéo Napoli, LORIA, Nancy
- Nathalie Pernelle, LRI, Université Paris-Sud 11
- Pascal Poncelet, LIRMM - Université de Montpellier
- Ricco Rakotomalala, ERIC, Université Lumière - Lyon 2
- Chantal Reynaud, LRI, Université Paris-Sud 11
- Ansaif Salieb-Aouissi, Columbia University, Center for Computational Learning Systems
- André Totohasina, Université Nord Madagascar Antsiranana
- Benoit Vaillant, Cohéris - SPAD

Table des matières

Vers un usage éclairé de la donnée géographique Christelle Pierkot	1
Propriété des mesures d'intérêt pour l'extraction des règles Sylvie Guillaume, Dhouha Grissa, Engelbert Mephu Nguifo	15
Mesure de la robustesse de règles d'associations Yannick Le Bras, Patrick Meyer, Philippe Lenca, Stéphane Lallich	29
Une approche basée agrégation pour une meilleure détection d'intrusions Emna Bahri, Harbi Nouria	41
Nouvelle représentation concise exacte des motifs corrélés fréquents basée sur une exploration simultanée des espaces de recherche conjonctif et disjonctif Nassima Ben Younes, Tarek Hamrouni, Sadok Ben Yahia	53
Index des auteurs	65

Vers un usage éclairé de la donnée géographique

Christelle Pierkot

LIRMM, 161 rue ADA, 34392 Montpellier cedex 5
christelle.pierkot@lirmm.fr,
<http://www.lirmm.fr>

Résumé. Nous nous situons dans le domaine applicatif de l'information géoréférencée. L'accès à celle-ci, outre le fait qu'il se soit largement démocratisé, a été largement facilité par l'apparition du Web et des services associés. L'utilisation des données spatiales est de plus en plus fréquente de nos jours mais elle concerne des utilisateurs d'horizons très variés dont les compétences sont diverses et dont les objectifs vont depuis un usage naturel jusqu'à des applications sophistiquées (cartographie en ligne et aide à la décision).

Afin d'utiliser au mieux l'information, ces *consommateurs* de données spatiales doivent pouvoir estimer la qualité des données mises à disposition. Une des voies les plus répandues pour évaluer cette qualité consiste à s'appuyer sur les métadonnées. La norme ISO 19115, spécifique au domaine, prévoit des champs liés à la description de la qualité des données mais ceux-ci ont été spécifiés en prenant en compte essentiellement le point de vue du producteur de données. Hors, la qualité souhaitée des données diffère en fonction des utilisateurs car elle dépend du besoin, de l'application et de l'usage final.

Nous présentons dans ce papier, tout d'abord brièvement, la norme ISO 19115 et les divers critères de qualité supportés, puis nous proposons à partir de réflexions existantes, quelques pistes pour spécifier des métadonnées spécifiques à la notion de qualité externe et qui permettrait une évaluation adaptée et en adéquation au besoin de l'utilisateur (fitness for use).

1 Introduction

L'information géographique est de plus en plus utilisée de nos jours. L'accès aux données spatiales ayant été considérablement facilité par l'apparition des services web, celle-ci est aujourd'hui disponible auprès d'un grand nombre d'utilisateurs provenant d'horizons variés et ayant des objectifs et des compétences différents. De ce fait, le type de consommateur de cette information géographique est très variable et va du simple utilisateur qui souhaite obtenir une information géospatiale simple (par exemple un itinéraire routier), à des experts qui utilisent les données pour mener à bien une étude (par exemple, un écologue dans le cadre de l'observation de l'évolution d'un écosystème) ou qui souhaitent diffuser une information élaborée à partir de celles-ci (par exemple, un géographe qui souhaite présenter une carte élaborée d'indicateurs sur un flux de migration). Tous ces utilisateurs emploient finalement les données spatiales à des fins de prise de décision, quelque soit l'importance de la décision à prendre.

Cependant, comme le souligne, (Harding, 2005), dans un contexte de résolution de problèmes ou de prise de décision, la fiabilité des résultats est en partie dépendante de l'utilisation prévue des données ainsi que de leur interopérabilité avec d'autres sources de données. La disponibilité d'un grand nombre de données n'assure pas forcément que celles-ci soient compatibles avec l'environnement technique de l'utilisateur, ni qu'elles soient conformes à la qualité attendue par l'utilisateur. Il est cependant, évident que l'usage de données de mauvaise qualité conduit presque toujours à des résultats incohérents voire erronés, ce qui est préjudiciable dans un processus de prise de décision. .

Avant d'utiliser des données spatiales, un utilisateur devrait donc toujours se poser un certain nombre de questions, comme par exemple, est ce que les données qu'on me propose correspondent à mon besoin ? sont-elles fiables ? sont-elles d'actualité ? sont-elles précises ? ... Le problème sous-jacent à cette disponibilité accrue des données spatiales est donc bien **l'évaluation de leur qualité.**

Dans le domaine de l'information géographique, depuis des décennies, la qualité préoccupe les organismes de normalisation et ceci notamment dans le contexte de la normalisation des métadonnées (ex. FGDC, OGC, CEN, ISO TC/211). Le concept de métadonnées avait été suggéré (par les producteurs essentiellement institutionnels) car les métadonnées complétaient les informations afférentes aux données produites et constituaient une première étape permettant au futur utilisateur d'orienter ses choix . Le standard de métadonnées actuel prédominant pour l'information géographique est l'ISO 19115 (ISO19115, 2006). La norme prévoit un grand nombre d'éléments et notamment des champs pour considérer la qualité des données mais ceux ci ont été spécifiés du point de vue du producteur de données. Progressivement, l'approche pour la qualité s'est développée autour d'une définition basée sur le « fitness for use » , c'est-à-dire autour du constat que celle-ci doit aussi être exprimée en fonction de l'adéquation à l'usage qu'en prévoit l'utilisateur final (Devillers et Jeansoulin, 2005), (Veregin, 1999). Force est de constater que cela implique un changement de point de vue. Le producteur de données qui était responsable de la définition (et par là même des tests d'évaluation) de la qualité doit céder une part de cette responsabilité aux utilisateurs de celle-ci. Avec l'approche « fitness for use » , (Chrisman, 2005) déclare, « le producteur ne porte aucun jugement, mais révèle simplement les résultats de certains tests. L'utilisateur potentiel évalue alors ces résultats en fonction de l'usage spécifique qui est prévu » .

Dans ce papier, nous définissons en section 2, la qualité telle qu'elle a été initialement perçue dans le domaine de l'information géographique et nous listons les critères habituellement utilisés pour la spécifier. Nous relatons ensuite les critères qui ont été retenus et définis dans la norme de métadonnées ISO 19115, en relevant les manques vis à vis de l'objectif adéquation aux usages. Dans la section 3, nous abordons la question difficile de la qualité d'une donnée du point de vue de son utilisateur final. Cela relève de l'évaluation de la qualité par celui-ci et en fonction de l'usage qu'il désire en faire. Cette évaluation est évidemment contextuelle et dépend de l'usage et du type d'utilisateur (simple consommateur ou expert). Nous proposons des pistes pour spécifier de nouveaux éléments de métadonnées selon le point de vue de l'utilisateur, et abordons comment, à partir de ceux-ci, l'utilisateur peut réaliser l'évaluation des données en fonction du contexte applicatif et de l'usage attendu de celles-ci.

2 La qualité des données en information géographique

Historiquement, la qualité des données spatiales était limitée à l'exactitude et la précision des données (Chrisman, 2005). Puis l'introduction de la notion de *terrain nominal* a permis de préciser la notion d'exactitude. Pour (Chrisman, 2005), le concept central de « terrain nominal » indique qu'un test d'exactitude ne peut pas être effectué de façon naïve. Le test n'est pas effectué dans le monde « réel », mais dans un monde « nominal » dans lequel les objets sont définis en fonction de spécifications et de techniques de mesure définies. Pour illustrer notre propos, l'Institut Géographique National (IGN) produit pour tous les lots de données correspondant à leur production (produits BD Carto, BD Topo, etc.) des spécifications textuelles qui définissent le terrain nominal de celle-ci. Le terrain nominal est donc un filtre sur le monde réel effectué par le producteur et dont le futur utilisateur doit avoir connaissance.

2.1 Critères de la qualité

Comme le souligne (Bel-Hadj-Ali, 2001) dans ses travaux, la qualité des données géographiques est tellement complexe qu'il est impossible d'utiliser une mesure globale et qu'il faut par conséquent recourir à plusieurs composantes pour la déterminer. Divers travaux de recherche ont proposé des critères (quantitatifs et qualitatifs) pour évaluer la qualité d'un jeu de données spatial. (Moellering, 1987) a initialement spécifié cinq critères : la généalogie, la précision géométrique ou précision de position, la précision sémantique ou précision des attributs, l'exhaustivité ou la complétude, et enfin la cohérence logique.

- **La généalogie** décrit l'histoire du jeu de données. C'est un critère qualitatif qui retrace la vie du jeu de données, depuis sa création jusqu'à la mise à disposition de l'utilisateur. Elle fournit des informations telles que l'historique des données, les indications sur les sources, les opérations de saisie, les transformations effectuées sur les données. Ces informations sont très utiles pour les utilisateurs car elles permettent d'indiquer à partir de quelle version, de quelles données de référence le jeu de données a été créé, quel traitement le jeu a-t-il subi, quelle est la personne ou l'organisme à contacter pour obtenir des informations ou compléments sur ces données...
- **La précision géométrique** (ou exactitude spatiale) donne les écarts de position entre les objets de la base et ceux du monde réel. C'est une valeur chiffrée qui se décompose en deux types :
 - La précision de position : L'objet est plus ou moins bien positionné sur la carte.
 - La précision de forme : la forme de l'objet est plus ou moins juste sur la carte.
- **La précision sémantique** (ou exactitude des attributs) est la différence entre la valeur d'un attribut du jeu de données et sa valeur dans le monde nominal. C'est un critère quantitatif qui porte sur la classification des objets, la codification des attributs et les relations entre objets.
- **L'exhaustivité** indique si les objets du terrain nominal sont tous représentés dans le jeu de données. C'est un critère quantitatif qui permet de répondre aux questions suivantes :
 - La zone est elle couverte complètement ?
 - Le nombre d'objets modélisés est-il égal au nombre d'objets sur le terrain ?
 - Est-ce que les objets modélisés ont le bon nombre d'attributs ?
 - Tous les objets présents dans le terrain nominal sont-ils représentés ?

Vers un usage éclairé de la donnée géographique

- **La cohérence logique** définit le degré de cohérence interne des données selon les règles de spécifications et de modélisation du jeu de données. Elle inclut la cohérence géométrique et la cohérence topologique des données spatiales. C'est un critère quantitatif qui permet de vérifier :
 - si les objets de la base de données géographique respectent les spécifications,
 - si les relations entre objets sont respectées et si elles sont conformes aux spécifications,
 - si la topologie est représentée,
 - si les variables utilisées respectent les valeurs prédéfinies.

Des composantes supplémentaires ont été ajoutées :

- **L'actualité** (ou précision temporelle) détermine les dates de la dernière mise à jour et de la validité des données (Guptill, 1995). C'est un critère qualitatif qui permet de répondre aux questions du type : mes données sont-elles à jour ? et qui renseigne en quelque sorte la « fraîcheur » des données.
- **La fidélité textuelle** est une mesure de l'exactitude de l'orthographe des informations écrites. C'est un critère quantitatif.
- **La cohérence sémantique** est un critère qualitatif qui fait référence à la qualité avec laquelle les objets géographiques sont décrits (Salgé, 1995)

Tous ces critères ont été largement expérimentés et ont été utilisés dans de nombreux groupes de normalisation européens et mondiaux. Les travaux de normalisation se sont focalisés sur des propositions étroitement liées à la réalisation d'objectifs qui dépassent le cadre stricto sensu de la qualité pour tenter de résoudre les problèmes plus larges d'interopérabilité, de partage et réutilisation. Ainsi, le comité européen de normalisation (CEN, 1998) a établi une norme expérimentale pour manipuler plus efficacement l'information géographique (CEN/TC287). Aux Etats-Unis, le standard FGDC/CSDGM (Content Standard for Digital Geospatial Metadata) a été développé par le comité fédéral des données géographiques, dans le but d'être utilisé par une infrastructure de données spatiale nationale (NSDI) (FGDC, 1998). Plus récemment, le comité international de normalisation (ISO) a lui aussi proposé une norme pour la gestion et l'échange des données spatiales (ISO19115, 2006).

2.2 Les métadonnées pour l'information géographique

Les métadonnées (données sur les données) forment un ensemble formel de propriétés descriptives relatives aux données qui peuvent être partagées par une communauté. L'introduction des métadonnées, fort ancienne au demeurant, a le double mérite de permettre la diffusion d'inventaires de données par les producteurs et de permettre une identification plus aisée par les utilisateurs.

En information géographique, les métadonnées sont plus complexes que celles utilisées dans les autres domaines du fait de la nature spécifique des informations qu'elles renseignent. En effet, elles sont composées à la fois d'éléments descriptifs relatifs aux dimensions thématiques et d'éléments spatiaux. Elles apportent différents renseignements tels que l'identification, l'étendue, la qualité, les schémas spatiaux et temporels ou encore la distribution des données spatiales (Danko, 2005).

La référence internationale en matière de métadonnées, l'ISO 19115 (ISO19115, 2006)

se décompose en plusieurs **sections** qui contiennent une ou plusieurs **entités** de métadonnées, elles mêmes constituées d'**éléments** de métadonnées. Le standard est présenté grâce au langage de modélisation UML. Les sections sont représentées avec des packages, les entités par des classes et les éléments par des attributs.

Les principales sections permettent l'identification des données (description, version, dimension spatiale, mots clés, etc), des informations relatives aux contacts (production, commercialisation, etc.) et enfin apportent des renseignements sur leur qualité. Le standard ISO 19115 définit volontairement un ensemble volumineux d'éléments de métadonnées afin de diversifier les exploitations. Cependant, un sous ensemble de la norme suffit généralement à une communauté spécifique mais une quantité minimale de métadonnées, contenue dans ce que l'ISO dénomme noyau de la norme, doit être maintenue. Ces éléments permettent de répondre aux questions les plus fréquemment posées par des utilisateurs de données géographiques (quoi ? où ? quand ? qui ?).

Une communauté spécifique d'utilisateurs de données géographiques n'utilise donc généralement qu'une partie des métadonnées définies dans la norme et a contrario a souvent besoin d'ajouter des métadonnées qui ne sont pas spécifiées dans le standard. ISO 19115 permet cela grâce à la définition de profils communautaires. Un profil permet de restreindre la norme à un sous ensemble d'éléments qui doivent être obligatoires (le noyau) et de l'étendre en ajoutant des sections, entités et éléments manquants.

Les métadonnées de qualité sont accessibles via la section `Data quality information` qui permet l'évaluation générale de jeux de données ou d'éléments du jeux de données via un ensemble d'éléments relatifs à une ou plusieurs instances de la classe `DQ_DataQuality` (Cf. figure 1).

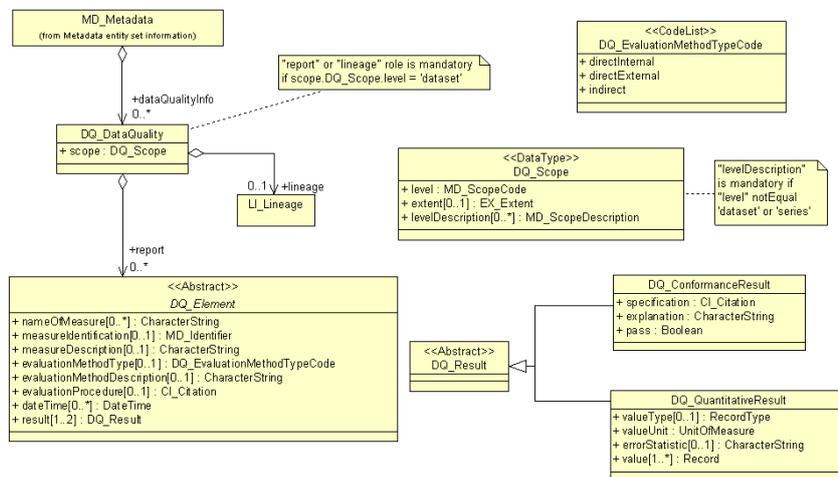


FIG. 1 – Informations de qualité dans ISO 19115.

Chaque instance de la classe `DQ_DataQuality` est caractérisée par un champ d'application (attribut `scope` de type `DQ_Scope`) qui spécifie la nature des données cibles, en particulier

Vers un usage éclairé de la donnée géographique

le niveau d'application des métadonnées (attribut `level` dont les valeurs possibles sont fournies dans la liste de code `MD_ScopeCode`) et la zone géographique concernée (attribut `extent` de type `EX_Extent`).

La classe `DQ_DataQuality` permet donc d'accéder aux informations de qualité. Cette classe est composée de deux classes dédiées l'une aux informations de généalogie (`LI_Lineage`), et l'autre aux informations quantitatives sur la qualité telles que la précision ou encore la cohérence des données (`DQ_Element`) (Cf. figure 2).

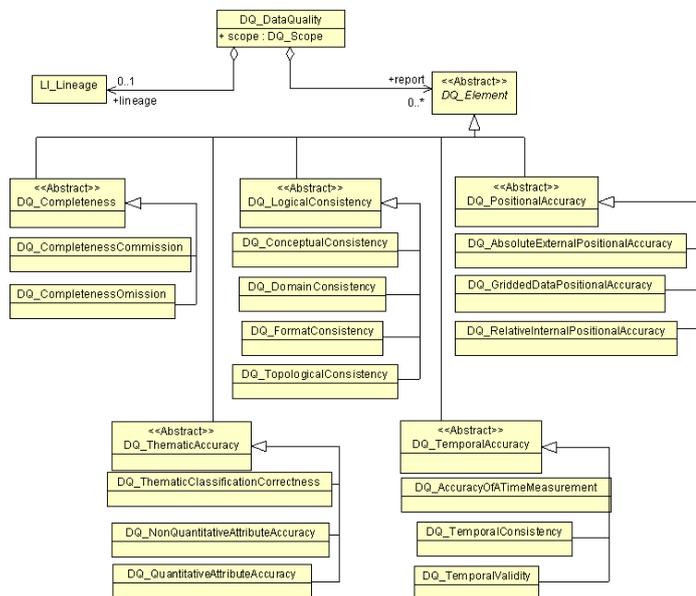


FIG. 2 – Classes pour la représentation de qualité dans ISO 19115.

La classe `LI_Lineage` renseigne sur la nature des données sources (`LI_Source`) et sur le processus de production (`LI_ProcessStep`) ayant conduit à la création du jeu de données.

La classe `DQ_Element` fournit un certain nombre d'informations sur les tests effectués pour mesurer la qualité, celles-ci portent notamment sur la méthode utilisée, la période pendant laquelle le test a été fait et surtout sur les résultats obtenus.

Elle est spécialisée en cinq sous-hiérarchies correspondant chacune à un critère de qualité :

- `DQ_Completeness` qualifie l'exhaustivité du jeu de données. Elle se spécialise en deux sous-classes :
 - `DQ_CompletenessCommission` qui décrit les données excédentaires du jeu de données.

- DQ_CompletenessOmission qui décrit les données manquantes du jeu de données.
- DQ_ThematicAccuracy fournit l'information sur la qualité des attributs. Elle se spécialise en trois sous-classes :
 - DQ_ThematicClassificationCorrectness qui décrit la cohérence des attributs,
 - DQ_NonQuantitativeAttributeAccuracy qui décrit la justesse des attributs non-quantitatifs,
 - DQ_QuantitativeAttributeAccuracy qui décrit la précision des attributs quantitatifs
- DQ_LogicalConsistency donne le degré d'adhésion aux règles logiques. Elle se spécialise en quatre sous-classes :
 - DQ_ConceptualConsistency qui indique la conformité par rapport au schéma conceptuel,
 - DQ_DomainConsistency qui indique la conformité par rapport au domaine de valeurs,
 - DQ_FormatConsistency qui indique la conformité par rapport au format,
 - DQ_TopologicalConsistency qui informe sur le degré de cohérence topologique.
- DQ_TemporalAccuracy qui fournit l'information sur la précision temporelle. Elle se spécialise en trois sous-classes :
 - DQ_AccuracyOfTimeMeasurement qui donne la précision d'une mesure temporelle,
 - DQ_TemporalConsistency qui définit le degré de cohérence temporelle,
 - DQ_TemporalValidity qui définit la validité temporelle.
- DQ_PositionalAccuracy qui fournit l'information sur la précision de position. Elle se spécialise en trois sous-classes :
 - DQ_AbsoluteExternalPositionalAccuracy qui indique la précision absolue,
 - DQ_GridDEDDataPositionalAccuracy qui indique la précision absolue pour le cas spécifique des données maillées,
 - DQ_RelativeInternalPositionalAccuracy qui indique la précision relative

Cependant, ces champs ont été spécifiés du point de vue du producteur de données et non du point de vue de l'utilisateur final. Hors, tout le monde s'accorde aujourd'hui pour dire que la qualité des données doit aussi être exprimée en fonction de l'adéquation à l'usage (Devillers et Jeansoulin, 2005), (Veregin, 1999).

3 Vers la qualité pour l'adéquation aux besoins

Les premières réflexions menées pour rendre compte du point des vues des utilisateurs, ont permis de définir les notions de qualité interne et externe (David et Fasquel, 1997).

3.1 Qualité interne et qualité externe

- La qualité interne est l'ensemble des propriétés et caractéristiques d'un produit ou service qui lui confère l'aptitude à satisfaire aux spécifications de contenu de ce produit ou de ce service. Elle se mesure par la différence entre les données qui devraient être produites et les données qui ont effectivement été produites. Elle est liée aux spécifications du terrain nominal (et en particulier aux erreurs qui peuvent être commises lors de la production des données) et est évaluée en fonction du producteur.
- La qualité externe est définie comme étant l'adéquation des spécifications aux besoins de l'utilisateur. Elle se mesure quant à elle, par la différence entre les données souhaitées par l'utilisateur et les données effectivement produites. Elle est liée aux besoins des utilisateurs et varie donc d'un utilisateur à l'autre.

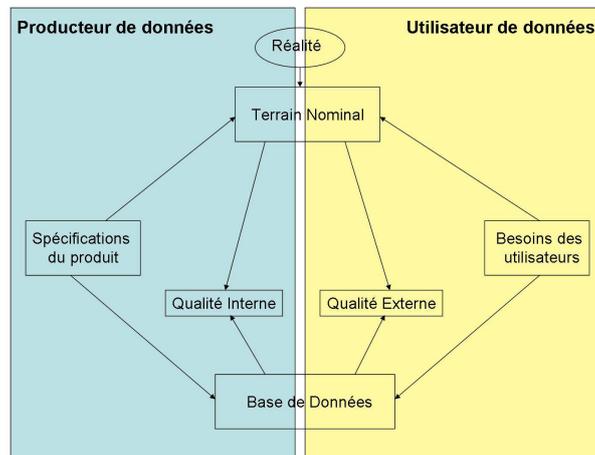


FIG. 3 – *Qualités interne et externe des jeux de données spatiaux.*

3.2 Les dimensions de la qualité externe

Concernant l'adéquation aux besoins, (Wang et Strong, 1996) ont défini formellement quatre dimensions pour prendre en compte la qualité externe des données spatiales :

- La qualité **intrinsèque** détermine la crédibilité, la précision, l'objectivité et la réputation que l'on peut accorder aux données.
- La qualité **contextuelle** s'attache plutôt à vérifier si les données sont appropriées (pertinence, valeur ajoutée) et suffisantes (complétude, volume de données) pour l'usage qui doit en être fait.
- La qualité **représentationnelle** aborde les notions d'interopérabilité et de compréhension des données.
- Enfin, le dernier critère concerne l'accessibilité et la sécurité liées aux données.

Le tableau 1 fait une synthèse des différents critères qui sont utilisés dans le domaine de l'information géographique pour définir la qualité des données spatiales. Nous montrons les correspondances entre ces éléments et les notions de qualité interne et externe vue précédemment et précisons pour chacun d'entre eux, s'ils sont de nature quantitative ou qualitative.

	Qualité Interne	Qualité Externe	Quantitatif	Qualitatif
Cohérence logique	X		X	
Cohérence sémantique	X			X
Fidélité textuelle	X		X	
Généalogie	X	X		X
Précision géométrique	X	X	X	
Précision sémantique	X	X	X	
Exhaustivité	X	X	X	
Actualité	X	X		X
Fiabilité		X	X	X
Réputation		X	X	X
Accessibilité		X		X
Pertinence		X	X	
Interprétabilité		X		X
Coût		X	X	

TAB. 1 – *Classement des critères de qualité.*

3.3 Quelques pistes pour prendre en compte l'adéquation aux besoins

(Devilleers et Jeansoulin, 2005) souligne que pour définir correctement la qualité des données, il faut disposer d'informations sur les données utilisées mais également sur les besoins des utilisateurs. D'ailleurs, la norme ISO 9000 définit la qualité comme étant « L'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites » (ISO9000, 2000).

Dans ce contexte, (Gutiérrez et Servigne, 2009) définissent la qualité comme étant la proximité entre les caractéristiques des données et les besoins d'un utilisateur pour une application donnée à un instant donné. En partant de ce constat, ils définissent deux catégories de métadonnées : les métadonnées génériques et les métadonnées applicatives (Gutiérrez et Servigne, 2009).

Les métadonnées génériques décrivent l'information minimale requise pour identifier et définir un ensemble de données. Elles représentent un socle commun à l'ensemble des applications. Les initiatives relatives à la mise en place des plateformes d'infrastructures de données spatiales (Spatial Data Infrastructure) comme la directive INSPIRE (Infrastructure for Spatial Information in the European Community), reprennent cette idée. Dans ce contexte, les producteurs fournissent des métadonnées conformes à ISO 19115 et les utilisateurs potentiels peuvent donc identifier et localiser les données à partir des critères relatifs à la qualité interne.

Les métadonnées applicatives décrivent quant à elles l'information spécifique à chaque appli-

cation, par rapport au type ou domaine d'application et à leurs caractéristiques. Les producteurs (notamment institutionnels) ciblent un type applicatif ou des usages prédéfinis mais cette information reste implicite. Nous atteignons ici le point délicat car pour décrire celles-ci, il faudrait avoir recensé les diverses catégories d'applications possibles. Cela relève de l'expertise métier des usagers de l'information spatiale. Pour que ces métadonnées soient renseignées, cela demande une analyse (qui commence à être menée notamment par les organisations généralistes comme l'US Census Bureau, ou plus proche de nous le CNIG (Conseil National de l'information géographique). Cette analyse demande de revisiter la qualité des données à partir d'un degré de satisfaction des exigences, exprimées par un ou plusieurs processus « métier », et par leur coût minimal de production.

La figure 4 représente une piste qui permettrait de définir la qualité du point de vue de l'utilisateur. Dans ce schéma, nous distinguons les contraintes applicatives auxquelles un utilisateur est soumis et les besoins que celui-ci est en mesure d'exprimer. Les contraintes sont données par le contexte d'application (type d'utilisateur, matériel mis à disposition, moyens logiciels, ...), alors que les besoins dépendent du ou des usages prévus pour les données (utilisation, exploitation, visualisation, traitements, ...). Les contraintes sont fixées pour une application donnée et n'évoluent pas a priori dans le temps. En revanche, les besoins peuvent évoluer régulièrement, par exemple dans un contexte de prise de décision, lorsque les usages prévus dépendent de la dernière décision prise.

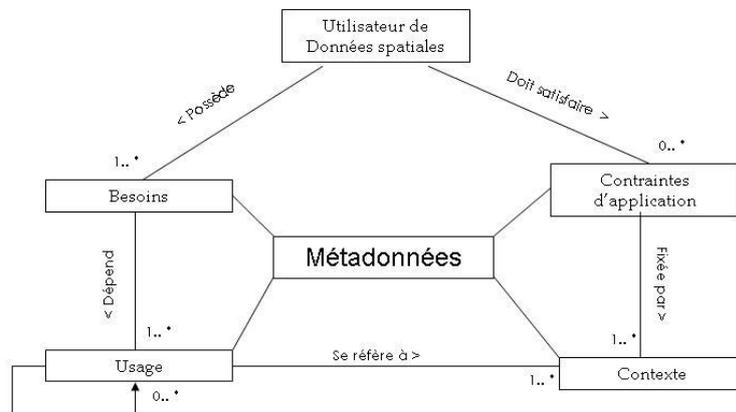


FIG. 4 – Une piste pour modéliser la qualité externe.

Le schéma proposé dans la figure 4 est le point de départ de notre réflexion concernant la qualité externe des données spatiales. Il va nous permettre de spécifier les métadonnées qui permettront de déterminer la qualité du point de vue de l'utilisateur, indépendamment du processus de production comme cela est le cas aujourd'hui dans la norme ISO 19115. Ces métadonnées pourraient ensuite être évaluées afin que l'utilisateur soit capable de dire si une donnée qu'on lui propose satisfera son besoin et si cette donnée sera appropriée à l'usage souhaité.

En ce qui concerne le contexte, celui-ci peut être décrit par l'utilisateur lui-même (voire découvert automatiquement). Les travaux relatifs aux annotations sur les services peuvent largement contribuer à la description du contexte applicatif. La partie plus complexe, relève de la description des besoins et usages. Plusieurs propositions se font jour, d'une part autour de la conception de moteurs de recherche et d'autre part autour de la description des traitements ou des chaînes de traitements à partir des connaissances utilisateurs.

En ce qui concerne les moteurs de recherche, de nombreuses occurrences existent¹ (M3cat, GEONetwork, MdWeb, etc.) toutes basées sur les métadonnées ISO. La précision des réponses aux requêtes des utilisateurs dépend de la prise en compte de la sémantique (description des connaissances via des thesaurus, des ontologies). Le projet européen SPIRIT² s'est quant à lui attaché à améliorer la distance sémantique qualitative au sein des recherches en construisant une ontologie géographique. Ce projet a aussi initié une réflexion autour de l'expression des besoins afin de dégager une liste de fonctionnalités attendues par les utilisateurs visés par les partenaires du projet (Bucher et al., 2004).

En ce qui concerne la description des traitements, la communauté des services a déjà défini des métadonnées qui correspondent grosso-modo aux signatures des traitements (sémantique du traitement et type des entrées/sorties). Nous pensons, qu'il faut établir un lien entre la description des traitements et la sémantique des fonctionnalités attendues, c'est-à-dire construire un modèle de métadonnées liant ressources logicielles et fonctions. Le problème est d'envergure, car il recouvre toute la difficulté de la prise en compte de points de vue différents sur des traitements pré-établis.

4 Conclusions et perspectives

Nous avons tout d'abord défini la notion de qualité des données dans le domaine de l'information géographique. Nous avons ensuite exposé les différents critères permettant de définir la qualité et avons montré comment ces éléments étaient modélisés dans la norme ISO 19115. Nous avons ensuite souligné les manques inhérents face à la nécessité de prendre en compte le point de vue de l'utilisateur et ceci dans le but de dégager quels éléments complémentaires peuvent aider celui-ci à évaluer la qualité des données spatiales. Nous avons proposé une première piste pour prendre en compte la notion de qualité externe.

La proposition s'appuie sur un modèle de métadonnées décrivant le contexte physique du poste ou de l'application dont l'utilisateur dispose (qui peuvent être découvertes ou bien décrites par celui-ci) et sur une catégorisation des usages établie par référence à l'expertise des usagers c'est-à-dire sur des métadonnées extraites d'ontologies métier. Le travail est de longue haleine, nous comptons construire rapidement des hiérarchies d'usages à partir d'expériences menées dans les unités de recherche environnementales. Les perspectives consistent à envisager la construction de règles et algorithmes d'appariement entre catégorie d'usage et type de contexte qui permettraient l'évaluation de la notion de distance entre catégorie d'usage et type de contexte.

1. une première analyse peut être trouvée sur les sites <http://www.fgdc.gov/metadata/geospatial-metadata-tools> et <http://www.inspire-geoportal.eu/>

2. <http://www.geo-spirit.org/>

Références

- Bel-Hadj-Ali, A. (2001). *Qualité géométrique des entités géographiques surfaciques : Application à l'appariement et définition d'une typologie des écarts géométriques*. Ph. D. thesis, Université de Marne la Vallée.
- Bucher, Balley, Levin, Syed, Petrelli, Weibel, Abdelmoky, Bailieu, et Heinzle (2004). User requirements specification reassessment. Technical report. IS2001-35047 deliverable.
- CEN (1998). *Geographic Information European Prestandards, Euro-norme Voluntaire for Geographic Information Data description Metadata*. European Committee for Standardization CEN/TC287.
- Chrisman, N. (2005). Traitement de la qualité : Perspective historique. In R. Devillers et R. Jeansoulin (Eds.), *Qualité de l'information géographique*, pp. 25–35. Traités IGAT, Hermès Sciences, Lavoisier.
- Danko, D. (2005). Metadata and related standards: Overview / demonstration. In *ISO Standard Workshop in the 22th International Cartographic Conference*.
- David, B. et P. Fasquel (1997). Qualité d'une base de données géographique : concepts et terminologie. Technical report, IGN. Bulletin d'information n.67.
- Devillers, R. et R. Jeansoulin (2005). *Qualité de l'information géographique*. Traités en Information Géographique et Aménagement du Territoire, IGAT, Hermès Sciences, Lavoisier.
- FGDC (1998). *Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-SDT-001-1998*. Federal Geographic Data Committee, Metadata Ad Hoc Working Group.
- Guptill, S. (1995). Temporal information. In *Elements of Spatial Data Quality*, pp. 153–165. S.C.Guptill et J.L.Morisson. Oxford : Elsevier.
- Gutiérrez, C. et S. Servigne (2009). Métadonnées et qualité pour les systèmes de surveillance en temps-réel. *Revue Internationale de Géomatique* 19/2, pp. 151–168.
- Harding, J. (2005). Qualité des données vectorielles : perspective d'un producteur de données. In *Qualité de l'information géographique*, pp. 171–192. Traités IGAT, Hermès Sciences, Lavoisier. ISBN 2-7462-1097-5.
- ISO19115 (2006). *Geographic Information : Metadata*. ISO/TC 211.
- ISO9000 (2000). *Quality management systems*. International Organization for Standardization (ISO).
- Moellering, H. (1987). A draft proposed standard for digital cartographic data, national committee for digital cartographic standards. Technical report, American Congress on Surveying and Mapping.
- Salgé, F. (1995). Semantic accuracy. In *Elements of Spatial Data Quality*, pp. 139–152. S.C.Guptill et J.L.Morisson. Oxford : Elsevier.
- Veregin, H. (1999). Data quality parameters. *Geographical Information Systems Principles and Technical Issues*, pp. 177–189.
- Wang, R. W. et D. M. Strong (1996). Beyond accuracy : what data quality means to data consumers. *Journal of Management Information systems* 12(4), p5–34.

Summary

The access to geographic information has been facilitated by the emergence of web services, the use of spatial data is increasingly frequent nowadays by users of different backgrounds, having different objectives and competences.

To use information as well as possible, these consumers of spatial data must be able to estimate the quality of the available data. An opportunity to assess the quality is to use metadata. The standard ISO 19115 provides fields to consider the quality of data, but they have been specified according to the producer view. However, the desired quality differs according to users because it depends on the need, of the application and end the final use.

In this paper, we present the existing quality criteria defined in ISO 19115, and we propose ways to take into account the quality according to the application context, the user needs and the fitness for use.

Propriétés des mesures d'intérêt pour l'extraction des règles

Sylvie Guillaume¹, Dhouha Grissa² et Engelbert Mephu Nguifo²

Laboratoire LIMOS, UMR 6158 CNRS, Université d'Auvergne¹ et Université Blaise Pascal²
Complexe scientifique des Cézeaux, 63177 Aubière Cedex - France
sylvie.guillaume@isima.fr, dgrissa@gmail.com, mephu@isima.fr

Résumé. La recherche de règles d'association intéressantes est un domaine de recherche important et actif en fouille de données. Les algorithmes de la famille *Apriori* reposent sur deux mesures pour extraire les règles, le support et la confiance. Bien que ces deux mesures possèdent des vertus algorithmiques accélératrices, elles génèrent un nombre prohibitif de règles dont la plupart sont redondantes et sans intérêt. Il est donc nécessaire de disposer d'autres mesures filtrant les règles inintéressantes. Cet article synthétise les différents travaux réalisés pour dégager les "bonnes" propriétés des mesures d'extraction des règles afin de retenir celles qui sont intéressantes pour l'utilisateur. Toutes ces propriétés sont ensuite formalisées et évaluées sur soixante-neuf mesures, étendant de manière significative les précédents travaux de la littérature. Cette synthèse est le point de départ pour une catégorisation des mesures ainsi que pour la détection des propriétés redondantes.

1 Introduction

Les algorithmes d'extraction de règles d'association (Agrawal et Srikant 1994), fondés sur les mesures *support* et *confiance*, ont tendance à générer un nombre important de règles et plus particulièrement lorsque les données sont denses ou lorsque les items (*ou variables ou attributs*) sont fortement corrélés entre eux. Ce phénomène s'amplifie encore dans les deux situations suivantes :

1. lorsqu'on diminue le seuil du support, diminution souvent indispensable pour découvrir les spécificités des données (*c'est-à-dire les règles non générales*) que l'on nomme pépites de connaissances (*c'est-à-dire les règles ayant un faible support et une forte confiance*), ce qui a pour conséquence d'augmenter le nombre de motifs fréquents et donc le nombre de règles ;

2. lorsqu'on est en présence de variables numériques où une étape de transformation des données est nécessaire (*étape de discrétisation suivie d'un codage disjonctif complet*), ce qui a pour conséquence d'augmenter le nombre d'items et donc le nombre de règles extraites.

Le couple (*support, confiance*) utilisé par ces algorithmes n'est pas suffisant pour extraire uniquement les connaissances réellement intéressantes et a été remis en cause dans de nombreux travaux comme par exemple (Sese et Morishita 2002). Il est ainsi nécessaire de recourir à une étape supplémentaire d'analyse des règles extraites afin de ne retenir que celles qui sont réellement intéressantes.

Pour répondre à ces deux problèmes, le premier de nature quantitatif (*beaucoup de règles extraites*) et le second de nature qualitatif (*beaucoup de règles redondantes et non pertinentes*), différentes solutions ont été proposées. Une première solution consiste à restituer facilement et de façon synthétique l'information extraite grâce à des techniques de

représentation visuelle (Hofmann et Wilhelm 2001). Une seconde voie consiste à réduire le nombre de règles extraites. Certains auteurs (Zaki 2000, Zaman Ashrafi et al., 2004) éliminent les règles redondantes, d'autres évaluent et ordonnent les règles grâce à d'autres mesures d'intérêt (Lenca et al., 2003b).

Comme l'intérêt dépend à la fois des préférences de l'utilisateur et des données, les mesures ont été répertoriées en deux catégories (Freitas 1999) : les mesures subjectives (*orientées utilisateur*) et les mesures objectives (*orientées données*). Les mesures subjectives prennent en compte les objectifs de l'utilisateur, ses connaissances et ses croyances a priori sur le domaine étudié (Padmanabhan et Tuzhilin 1998). Les mesures objectives prennent en compte la structure des données et plus particulièrement les effectifs liés à la contingence des données (Tan et al. 2002, Lallich et Teytaud 2004).

Dans cet article, nous nous limitons aux mesures objectives car nous souhaitons une approche générique, indépendante du contexte applicatif et des connaissances a priori de l'utilisateur sur le domaine.

De nombreux travaux de synthèse ont comparé ces différentes mesures objectives. Ces synthèses abordent cette comparaison selon deux points de vue qui sont complémentaires et essentiels. Le premier point de vue recherche les propriétés sous-jacentes à une "bonne" mesure d'intérêt (Tan et al. 2002, Lallich et Teytaud 2004, Geng et Hamilton 2007, Feno 2007 et Vaillant 2007). Le second point de vue fait une étude comparative de leur comportement sur différents jeux de données. B. Vaillant dans (Vaillant, 2002) propose un premier outil d'expérimentation *HERBS* dédié à la post-analyse des règles, et Huynh et al. dans (Huynh et al., 2006) propose la plateforme *ARQAT* (*Association Rule Quality Analysis Tool*) afin d'étudier le comportement spécifique des mesures dans le contexte d'un jeu de règles.

Dans cet article, nous nous focalisons sur le premier point de vue c'est-à-dire la recherche de propriétés sous-jacentes à une "bonne" mesure d'intérêt.

L'objectif de cet article est double. Tout d'abord, il synthétise les "bonnes" propriétés pour une mesure rencontrées dans la littérature. En aucun cas, notre objectif est l'explication de l'intérêt de ces propriétés, explication déjà présente dans la littérature, mais notre objectif est la formalisation de ces propriétés permettant ainsi de lever toute ambiguïté sur celles-ci. De plus, nous avons généralisé ces propriétés toutes les fois que c'était possible. L'autre objectif est un recensement des mesures présentes dans la littérature, mesures qui sont ensuite évaluées sur les propriétés mises en évidence précédemment. Ce travail va conduire à la construction d'une matrice (*69 mesures × 19 propriétés*) qui est une tâche préliminaire et indispensable à une catégorisation des mesures afin d'aider l'utilisateur dans le choix de ces mesures ainsi qu'à la recherche de propriétés redondantes.

L'article s'organise donc de la façon suivante. La *section 2* expose les travaux effectués dans le domaine de la recherche de "bonnes" propriétés ainsi que les différentes mesures rencontrées dans la littérature. La *section 3* synthétise et formalise ces différentes propriétés et la *section 4* évalue ces propriétés sur soixante-neuf mesures objectives.

2 Règles d'association, propriétés et mesures de qualité

Dans cette section, nous commençons par définir les règles d'association et les notations utilisées dans cet article. Ensuite, nous exposons les travaux réalisés dans le domaine de la recherche de "bonnes" propriétés pour les mesures objectives rencontrées dans la littérature.

◆ Définition des règles d'association et notations utilisées

Une règle d'association (Agrawal et Srikant 1994) est une implication de la forme $X \rightarrow Y$ où X et Y sont des motifs (ou conjonctions de variables binaires) disjoints.

Une règle d'association est évaluée classiquement grâce à deux mesures : le *support* et la *confiance*. Pour simplifier les notations, nous noterons par XY la conjonction $X \wedge Y$ des motifs X et Y .

Le support $support(X \rightarrow Y)$ de la règle $X \rightarrow Y$ est la proportion d'individus vérifiant simultanément X et Y , c'est-à-dire $support(X \rightarrow Y) = P(XY) = \frac{|(XY)_{e_i \in \Omega}|}{|\Omega|} = \frac{n_{XY}}{n}$.

$(XY)_{e_i \in \Omega}$ est l'ensemble des individus e_i de l'ensemble d'apprentissage Ω vérifiant simultanément les motifs X et Y . La notation $|(XY)_{e_i \in \Omega}|$ ou n_{XY} indique le nombre d'individus vérifiant à la fois X et Y et la notation $|\Omega|$ ou n correspond au nombre d'individus dans l'ensemble d'apprentissage Ω . Pour finir, $P(XY)$ est la probabilité d'apparition du motif XY .

La confiance $confiance(X \rightarrow Y)$ de la règle $X \rightarrow Y$ est la proportion d'individus vérifiant Y parmi ceux qui vérifient X , c'est-à-dire $confiance(X \rightarrow Y) = P(Y|X) = \frac{|(XY)_{e_i \in \Omega}|}{|(X)_{e_i \in \Omega}|} = \frac{n_{XY}}{n_X}$.

Soient 2 seuils définis par l'utilisateur : le support minimum min_{sup} et la confiance minimum min_{conf} .

Une règle est dite valide si elle vérifie les contraintes de support et de confiance suivantes : $support(X \rightarrow Y) \geq min_{sup}$ et $confiance(X \rightarrow Y) \geq min_{conf}$.

Pour la suite, on entend par \bar{X} le motif ne vérifiant pas X .

Après avoir défini les règles d'association et les notations utilisées dans cet article, nous exposons les travaux réalisés dans la recherche de "bonnes" propriétés pour les mesures objectives.

◆ Travaux de Piatetsky-Shapiro (Piatetsky-Shapiro, 1991)

Piatetsky-Shapiro a proposé trois propriétés (S_1 à S_3) que doivent vérifier les mesures m d'intérêt :

S_1 : la valeur de la mesure m doit être nulle dans le cas de l'indépendance c'est-à-dire lorsque $P(XY) = P(X)P(Y)$ ou encore lorsque $P(Y|X) = P(Y)$.

S_2 : la mesure m doit être croissante en fonction du nombre n_{XY} d'exemples lorsque la taille n_X de la prémisse X et la taille n_Y de la conclusion Y restent constantes,

S_3 : la mesure m doit être décroissante en fonction de la taille n_X de la prémisse lorsque le nombre n_{XY} d'exemples et la taille n_Y de la conclusion restent constantes (ou encore en fonction de la taille de la conclusion lorsque le nombre d'exemples et la taille de la prémisse restent constantes).

De ces trois propriétés, il en déduit deux implications intéressantes qui sont les suivantes :

I_1 : les valeurs des mesures doivent être positives en cas d'attraction entre X et Y c'est-à-dire quand la réalisation de X augmente les chances d'apparition de Y , autrement dit lorsque $P(Y|X) > P(Y)$,

I_2 : les valeurs des mesures doivent être négatives en cas de répulsion entre X et Y c'est-à-dire lorsque la réalisation de X diminue les chances d'apparition de Y , autrement dit lorsque $P(Y|X) < P(Y)$.

Propriétés des mesures d'intérêt

◆ Travaux de Tan et al. (Tan et al., 2002)

Tan et al. ont réalisé une étude sur 21 mesures (*coefficient de corrélation, confiance, conviction, cosinus, facteur de certitude, force collective, Gini, Goodman-Kruskal, information mutuelle, intérêt, Jaccard, J-mesure, Laplace, Piatetsky-Shapiro, Q de Yule, Kappa, Kloggen, ratio des chances, support, valeur ajoutée, Y de Yule*) et les ont étudiées à travers 8 propriétés.

Les 3 premières propriétés se rapportent aux propriétés S_1 à S_3 de Piatetsky-Shapiro énoncées précédemment, et les cinq autres propriétés (T_1 à T_5) sont les suivantes :

T_1 : la mesure m est symétrique,

T_2 : la mesure m est invariante dans les deux cas suivants :

1- lorsqu'on multiplie les effectifs des ensembles $(XY)_{e_i \in \Omega}$ et $(X\bar{Y})_{e_i \in \Omega}$ par une constante positive k_1 et les effectifs des ensembles $(\bar{X}Y)_{e_i \in \Omega}$ et $(\bar{X}\bar{Y})_{e_i \in \Omega}$ par une constante k_2 positive,

2- lorsqu'on multiplie les effectifs des ensembles $(X\bar{Y})_{e_i \in \Omega}$ et $(\bar{X}\bar{Y})_{e_i \in \Omega}$ par une constante positive k_1 et les effectifs des ensembles $(XY)_{e_i \in \Omega}$ et $(\bar{X}Y)_{e_i \in \Omega}$ par une constante k_2 positive.

T_3 : la mesure m doit vérifier les relations suivantes :

$$m(\bar{X} \rightarrow Y) = -m(X \rightarrow Y) \quad \text{et} \quad m(X \rightarrow \bar{Y}) = -m(X \rightarrow Y) .$$

T_4 : la mesure m doit vérifier la relation $m(\bar{X} \rightarrow \bar{Y}) = m(X \rightarrow Y)$.

T_5 : la mesure m doit être invariante lorsque la taille n de l'ensemble d'apprentissage Ω augmente et que tous les autres effectifs (n_X , n_Y et n_{XY}) restent constants.

◆ Travaux de Lallich et Teytaud (Lallich et Teytaud, 2004)

Quant à S. Lallich et O. Teytaud, ils montrent les avantages et les inconvénients pour une mesure objective de posséder différentes propriétés. Cette discussion porte sur essentiellement 15 mesures (*coefficient de corrélation, confiance, confiance centrée, conviction, facteur bayésien, indice d'implication, intensité d'implication, intérêt, J-mesure, Loevinger, moindre contradiction, Pearl, Piatetsky-Shapiro, Sebag-Schoenauer et Zhang*) et sur 13 propriétés qui sont les suivantes :

- L_1 : Intelligibilité ou compréhensibilité de la mesure (Lenca et al., 2003a)

La mesure doit être intelligible pour pouvoir communiquer et expliquer les résultats obtenus.

- L_2 : Mesure non symétrique au sens de la négation de la conclusion

Une mesure doit pouvoir faire la distinction entre $X \rightarrow Y$ et $X \rightarrow \bar{Y}$.

- L_3 : Mesure non symétrique

Il est préférable d'avoir des mesures qui évaluent différemment les règles $X \rightarrow Y$ et $Y \rightarrow X$ puisque l'antécédent et la conclusion jouent des rôles différents (Freitas, 1999).

- L_4 : Mesure évaluant de la même façon $X \rightarrow Y$ et $\bar{Y} \rightarrow \bar{X}$ dans le cas de l'implication logique.

- L_5 : Taille de la prémisse fixe ou aléatoire

La taille de la prémisse est aléatoire lorsque la mesure est fondée sur un modèle probabiliste.

- L_6 : Mesure décroissante en fonction du nombre de contre-exemples

C'est une propriété équivalente à celle de Piatetsky-Shapiro, à savoir la propriété S_2 .

- L_7 : Valeurs fixes pour l'indépendance (*généralisation de S_1*)

- L_8 : Valeurs fixes pour l'implication (Lenca et al., 2003a)

- L_9 : Tolérance aux premiers contre-exemples (Gras et al., 2001)
- L_{10} : Mesure croissante en fonction de la rareté du conséquent
C'est une propriété équivalente à celle de Piatetsky-Shapiro, à savoir la propriété S_3 .
- L_{11} : Mesure descriptive ou statistique
Une mesure est descriptive lorsque sa valeur est invariante en cas de dilatation des données, c'est-à-dire lorsque tous les effectifs sont multipliés par un même coefficient k . Dans le cas contraire, elle est statistique.
- L_{12} : Mesure discriminante : c'est une mesure qui permet de discerner les règles, même lorsque l'ensemble d'apprentissage est volumineux.
- L_{13} : Facilité à fixer un seuil d'acceptation de la règle (Lenca et al., 2003b)

◆ **Travaux de Blanchard et al. (Blanchard et al. 2005a)**

Les auteurs préconisent que les mesures de qualité doivent tenir compte de l'état caractéristique "équilibre" ou "indétermination", c'est-à-dire l'état où le nombre d'exemples et de contre-exemples est identique. Pour cela, ils suggèrent que les mesures aient une valeur constante pour l'équilibre, d'où la propriété B_1 suivante.

B_1 : valeur fixe dans le cas de l'équilibre.

◆ **Travaux de Geng et Hamilton (Geng et Hamilton, 2007)**

L. Geng et H.J. Hamilton ont fait une étude de 38 mesures (*coefficient de corrélation, confiance, conviction, cosinus, couverture, dépendance pondérée d'intérêt de Gray et Orłowska, facteur bayésien, facteur de certitude, force collective, gain informationnel, Gini, Goodman-Kruskal, information mutuelle, intérêt, Jaccard, J-mesure, Klosgen, Laplace, Leverage, Loevinger, moindre contradiction, Piatetsky-Shapiro, précision, prévalence, Q de Yule, rappel, ratio des chances, risque relatif, Sebag-Schoenauer, spécificité, support, support à sens unique, support à double sens, taux d'exemples et de contre-exemples, valeur ajoutée, variation du support à double sens, Y de Yule, Zhang*) selon 11 propriétés. Les propriétés étudiées sont les suivantes. Tout d'abord, les propriétés S_1 à S_3 de Piatetsky-Shapiro puis les 5 propriétés T_1 à T_5 de Tan et al. et pour finir, les propriétés L_7 et L_9 . La dernière propriété étudiée est la suivante :

Le_1 : la mesure est croissante en fonction de la taille de l'ensemble d'apprentissage (Lallich, 2002).

◆ **Travaux de Feno (Feno, 2007)**

D. Feno a étudié 15 mesures (*coefficient de corrélation, confiance, confiance centrée, conviction, J-mesure, indice d'implication, intérêt, Loevinger, moindre contradiction, nouveauté, Pearl, Piatetsky-Shapiro, rappel, Sebag-Schoenauer, support*) selon 13 propriétés. Les propriétés étudiées sont les suivantes : $S_1, S_2, S_3, I_1, I_2, L_1, L_2, L_3, L_4, L_9, L_{13}, Le_1$ et B_1 .

◆ **Travaux de Vaillant (Vaillant, 2007)**

B. Vaillant a étudié 20 mesures (*coefficient de corrélation, Cohen, confiance, confiance centrée, conviction, facteur bayésien, indice d'implication, indice probabiliste discriminant, intensité d'implication, intensité d'implication entropique tronquée, intérêt, gain informationnel, Laplace, Loevinger, moindre contradiction, Piatetsky-Shapiro, Sebag-Schoenauer, support, taux d'exemples et de contre-exemples, Zhang*) selon 9 propriétés. Les propriétés étudiées sont les suivantes : $L_1, L_3, L_7, L_8, L_9, L_{10}, L_{13}, B_1$ et Le_1 .

Après avoir exposé les travaux réalisés dans la recherche de "bonnes" propriétés pour les mesures d'extraction des règles, la section suivante va les formaliser.

3 Synthèse et formalisation des propriétés

Dans cette section, nous synthétisons et formalisons donc les différentes propriétés rencontrées dans la littérature et qui ont été exposées dans la *section 2*. Le titre des 21 propriétés répertoriées est, de préférence, la propriété souhaitée pour la mesure m d'extraction de règles.

Propriété 1 : Intelligibilité ou compréhensibilité de la mesure

Cette propriété a été reprise de la propriété L_1 et nous avons conservé les 3 niveaux d'intelligibilité définis par (Vaillant, 2007).

$P_1(m) = 0$ si l'interprétation de la mesure est difficile,
 $P_1(m) = 1$ si la mesure se ramène à des quantités usuelles,
 $P_1(m) = 2$ si la mesure peut s'expliquer par une phrase.

Propriété 2 : Facilité à fixer un seuil d'acceptation de la règle. Cette propriété a été reprise de la propriété L_{13} .

$P_2(m) = 0$ si la détermination du seuil est problématique,
 $P_2(m) = 1$ si la détermination du seuil est immédiate.

Propriété 3 : Mesure non symétrique. Cette propriété a été reprise des propriétés T_1 et L_3 .

$P_3(m) = 0$ si m est symétrique i.e. si $\forall X \rightarrow Y m(X \rightarrow Y) = m(Y \rightarrow X)$
 $P_3(m) = 1$ si m est non symétrique i.e. si $\exists X \rightarrow Y / m(X \rightarrow Y) \neq m(Y \rightarrow X)$

Propriété 4 : Mesure non symétrique au sens de la négation de la conclusion

Cette propriété a été reprise de la propriété L_2 ($P_4(m) = 1$). Tan et al. ont exprimé une idée similaire avec la deuxième partie de la propriété T_3 , la partie indiquant que quelque soit les règles $X \rightarrow Y$, nous devons avoir $m(X \rightarrow \bar{Y}) = -m(X \rightarrow Y)$. C'est un cas particulier de $P_4(m) = 1$.

$P_4(m) = 0$ si m est *nc*-symétrique i.e. si $\forall X \rightarrow Y m(X \rightarrow Y) = m(X \rightarrow \bar{Y})$
 $P_4(m) = 1$ si m est non *nc*-symétrique i.e. si $\exists X \rightarrow Y / m(X \rightarrow Y) \neq m(X \rightarrow \bar{Y})$

Propriété 5 : Mesure évaluant de la même façon $X \rightarrow Y$ et $\bar{Y} \rightarrow \bar{X}$ dans le cas de l'implication logique. Cette propriété a été reprise de la propriété L_4 ($P_5(m) = 1$).

$P_5(m) = 0$ si $\exists X \rightarrow Y / P(Y/X) = 1$ et $m(X \rightarrow Y) \neq m(\bar{Y} \rightarrow \bar{X})$
 $P_5(m) = 1$ si $\forall X \rightarrow Y P(Y/X) = 1 \Rightarrow m(X \rightarrow Y) = m(\bar{Y} \rightarrow \bar{X})$

Propriété 6 : Mesure croissante en fonction du nombre d'exemples ou décroissante en fonction du nombre de contre-exemples. Cette propriété a été reprise des propriétés S_2 et L_6 .

$P_6(m) = 0$ si m n'est pas croissante en fonction de n_{XY} i.e. si $\exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2 /$
 $n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $(n_{X_1 Y_1} < n_{X_2 Y_2}$ ou $n_{X_1 \bar{Y}_1} > n_{X_2 \bar{Y}_2})$ et $m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2)$
 $P_6(m) = 1$ si m est croissante en fonction de n_{XY} i.e. si $\forall X_1 \rightarrow Y_1, \forall X_2 \rightarrow Y_2$
 $[n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $(n_{X_1 Y_1} < n_{X_2 Y_2}$ ou $n_{X_1 \bar{Y}_1} > n_{X_2 \bar{Y}_2})] \Rightarrow m(X_1 \rightarrow Y_1) \leq m(X_2 \rightarrow Y_2)$ et
 $\exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2 / n_{X_1} = n_{X_2}$ et $n_{Y_1} = n_{Y_2}$ et $(n_{X_1 Y_1} < n_{X_2 Y_2}$ ou $n_{X_1 \bar{Y}_1} > n_{X_2 \bar{Y}_2})$
et $m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2)$

Propriété 7 : Mesure croissante en fonction de la taille de l'ensemble d'apprentissage

Cette propriété a été reprise de la propriété Le_1 . Elle a également été reprise en partie de la propriété T_5 ($P_7(m) = 0$) puisque Tan et al. préconisent une invariance lorsque la taille de l'ensemble d'apprentissage augmente, c'est-à-dire lorsque $\forall X_1 \rightarrow Y_1 (\Omega_1), \forall X_2 \rightarrow Y_2 (\Omega_2)$,

$$[n_{X_1} = n_{X_2} \text{ et } n_{Y_1} = n_{Y_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_1 < n_2] \Rightarrow m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2).$$

C'est donc un cas particulier de $P_7(m) = 0$. Soient Ω_1 et Ω_2 deux ensembles d'apprentissage. Soient n_1 la taille du premier ensemble d'apprentissage Ω_1 et n_2 la taille du second ensemble d'apprentissage Ω_2 . On entend par la notation $X_1 \rightarrow Y_1 (\Omega_1)$, la règle $X_1 \rightarrow Y_1$ extraite dans l'ensemble d'apprentissage Ω_1 .

$$\begin{aligned} P_7(m) = 0 \text{ (} m \text{ pas croissante en fonction de } n \text{)} & \text{ si } \exists (\Omega_1, \Omega_2) \exists X_1 \rightarrow Y_1 (\Omega_1), \exists X_2 \rightarrow Y_2 (\Omega_2) \\ & / n_{X_1} = n_{X_2} \text{ et } n_{Y_1} = n_{Y_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_1 < n_2 \text{ et } m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2) \\ P_7(m) = 1 \text{ (} m \text{ croissante en fonction de } n \text{)} & \text{ si } \forall \Omega_1, \forall \Omega_2, \forall X_1 \rightarrow Y_1 (\Omega_1), \forall X_2 \rightarrow Y_2 (\Omega_2), \\ & (n_{X_1} = n_{X_2} \text{ et } n_{Y_1} = n_{Y_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_1 < n_2) \Rightarrow m(X_1 \rightarrow Y_1) \leq m(X_2 \rightarrow Y_2) \text{ et} \\ & \exists \Omega_1, \exists \Omega_2, \exists X_1 \rightarrow Y_1 (\Omega_1), \exists X_2 \rightarrow Y_2 (\Omega_2) / \\ & n_{X_1} = n_{X_2} \text{ et } n_{Y_1} = n_{Y_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_1 < n_2 \text{ et } m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2) \end{aligned}$$

Propriété 8 : Mesure décroissante en fonction de la taille du conséquent ou de la taille de la prémisse. Cette propriété a été reprise de la propriété L_{10} ($P_8(m) = 1$) et de la propriété S_3 .

$$\begin{aligned} P_8(m) = 0 \text{ si } m \text{ n'est pas décroissante en fonction de } n_Y \text{ i.e. si } & \exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2 / \\ & n_{X_1} = n_{X_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_{Y_1} < n_{Y_2} \text{ et } m(X_1 \rightarrow Y_1) < m(X_2 \rightarrow Y_2) \\ P_8(m) = 1 \text{ si } m \text{ est décroissante en fonction de } n_Y \text{ i.e. si } & \forall X_1 \rightarrow Y_1, \forall X_2 \rightarrow Y_2 \\ & (n_{X_1} = n_{X_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_{Y_1} < n_{Y_2}) \Rightarrow m(X_1 \rightarrow Y_1) \geq m(X_2 \rightarrow Y_2) \text{ et} \\ & \exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2 / n_{X_1} = n_{X_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_{Y_1} < n_{Y_2} \text{ et } m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2) \end{aligned}$$

Si nous considérons la taille de la prémisse, la propriété $P_8(m) = 1$ s'écrit aussi :

$$\begin{aligned} P_8(m) = 1 \text{ si } m \text{ est décroissante en fonction de } n_X \text{ i.e. lorsque} \\ \forall X_1 \rightarrow Y_1, \forall X_2 \rightarrow Y_2 (n_{Y_1} = n_{Y_2} \text{ et } n_{X_1 Y_1} = n_{X_2 Y_2} \text{ et } n_{X_1} < n_{X_2}) \Rightarrow m(X_1 \rightarrow Y_1) > m(X_2 \rightarrow Y_2) \end{aligned}$$

Propriété 9 : Valeur fixe a dans le cas de l'indépendance

Cette propriété a été reprise de la propriété L_7 et S_1 (cas où $a = 0$).

$$\begin{aligned} P_9(m) = 0 & \text{ si } \forall a \in \mathfrak{R} \exists X \rightarrow Y / P(Y/X) = P(Y) \text{ et } m(X \rightarrow Y) \neq a \\ P_9(m) = 1 \text{ (valeur fixe)} & \text{ si } \exists a \in \mathfrak{R} / \forall X \rightarrow Y P(Y/X) = P(Y) \Rightarrow m(X \rightarrow Y) = a \end{aligned}$$

Propriété 10 : Valeur fixe b dans le cas de l'implication logique. Cette propriété a été reprise des propriétés L_8 .

$$\begin{aligned} P_{10}(m) = 0 & \text{ si } \forall b \in \mathfrak{R} \exists X \rightarrow Y / P(Y/X) = 1 \text{ et } m(X \rightarrow Y) \neq b \\ P_{10}(m) = 1 \text{ (valeur fixe)} & \text{ si } \exists b \in \mathfrak{R} / \forall X \rightarrow Y P(Y/X) = 1 \Rightarrow m(X \rightarrow Y) = b \end{aligned}$$

Propriété 11 : Valeur fixe c dans le cas de l'équilibre

Cette propriété a été reprise de la propriété B_1 .

$$\begin{aligned} P_{11}(m) = 0 & \text{ si } \forall c \in \mathfrak{R} \exists X \rightarrow Y / P(Y/X) = P(X)/2 \text{ et } m(X \rightarrow Y) \neq c \\ P_{11}(m) = 1 \text{ (valeur fixe)} & \text{ si } \exists c \in \mathfrak{R} / \forall X \rightarrow Y P(Y/X) = P(X)/2 \Rightarrow m(X \rightarrow Y) = c \end{aligned}$$

Propriétés des mesures d'intérêt

Propriété 12 : Valeurs identifiables en cas d'attraction entre X et Y

Cette propriété a été reprise et généralisée de l'implication I_1 de Piatetsky-Shapiro.

$$\begin{aligned} P_{12}(m) = 0 & \quad \text{si } \forall a \in \mathfrak{R} \exists X \rightarrow Y / P(Y/X) > P(Y) \text{ et } m(X \rightarrow Y) \leq a \\ P_{12}(m) = 1 \text{ (valeurs identifiables)} & \text{ si } \exists a \in \mathfrak{R} / \forall X \rightarrow Y P(Y/X) > P(Y) \Rightarrow m(X \rightarrow Y) > a \end{aligned}$$

La valeur a indiquée dans la *propriété 12* correspond à la valeur fixe dans le cas de l'indépendance. De cette remarque, nous en déduisons que si la *propriété 9* n'est pas vérifiée alors la *propriété 12* ne le sera pas non plus : si $P_9(m) = 0$ alors $P_{12}(m) = 0$. Cette remarque est aussi valable pour la *propriété 13*.

Propriété 13 : Valeurs identifiables en cas de répulsion entre X et Y

Cette propriété a été reprise et généralisée de l'implication I_2 de Piatetsky-Shapiro.

$$\begin{aligned} P_{13}(m) = 0 & \quad \text{si } \forall a \in \mathfrak{R} \exists X \rightarrow Y / P(Y/X) < P(Y) \text{ et } m(X \rightarrow Y) \geq a \\ P_{13}(m) = 1 \text{ (valeurs identifiables)} & \text{ si } \exists a \in \mathfrak{R} / \forall X \rightarrow Y P(Y/X) < P(Y) \Rightarrow m(X \rightarrow Y) < a \end{aligned}$$

Propriété 14 : Tolérance aux premiers contre-exemples

Cette propriété a été reprise de la propriété L_9 et nous avons retenu les trois modalités proposées par B. Vaillant dans (Vaillant, 2007).

$$\begin{aligned} P_{14}(m) = 0 \text{ si rejet donc convexe, } & \exists \min_{conf} \in [0, 1] / \forall X_1 \rightarrow Y_1 \forall X_2 \rightarrow Y_2 \forall \lambda \in [0, 1] \\ & n_{X_1Y_1} \geq \min_{conf} n_{X_1} \text{ et } n_{X_2Y_2} \geq \min_{conf} n_{X_2} \\ & \Rightarrow f_{m,n_{XY}}(\lambda n_{X_1Y_1} + (1-\lambda)n_{X_2Y_2}) \leq \lambda f_{m,n_{XY}}(n_{X_1Y_1}) + (1-\lambda) f_{m,n_{XY}}(n_{X_2Y_2}) \\ P_{14}(m) = 1 \text{ si indifférence donc notamment linéaire i.e. } & P_{14}(m) \neq 0 \text{ et } P_{14}(m) \neq 2 \\ P_{14}(m) = 2 \text{ si tolérance donc concave } & \exists \min_{conf} \in [0, 1] / \forall X_1 \rightarrow Y_1 \forall X_2 \rightarrow Y_2 \forall \lambda \in [0, 1] \\ & n_{X_1Y_1} \geq \min_{conf} n_{X_1} \text{ et } n_{X_2Y_2} \geq \min_{conf} n_{X_2} \\ & \Rightarrow f_{m,n_{XY}}(\lambda n_{X_1Y_1} + (1-\lambda)n_{X_2Y_2}) \geq \lambda f_{m,n_{XY}}(n_{X_1Y_1}) + (1-\lambda) f_{m,n_{XY}}(n_{X_2Y_2}) \end{aligned}$$

La notation $f_{m,n_{XY}}$ correspond à la fonction d'évolution de la mesure m en fonction de n_{XY} lorsque les effectifs n_X , n_Y et n restent constants.

Propriété 15 : Invariance en cas de dilatation de certains effectifs

Cette propriété a été reprise de la propriété T_2 .

$$\begin{aligned} P_{15}(m) = 0 \text{ (variance) si } & \exists (k_1, k_2) \in \mathbb{N}^{*2}, \exists X_1 \rightarrow Y_1, \exists X_2 \rightarrow Y_2 / \\ [& n_{X_1Y_1} = k_1 n_{X_2Y_2} \text{ et } n_{X_1\bar{Y}_1} = k_1 n_{X_2\bar{Y}_2} \text{ et } n_{\bar{X}_1Y_1} = k_2 n_{\bar{X}_2Y_2} \text{ et } n_{\bar{X}_1\bar{Y}_1} = k_2 n_{\bar{X}_2\bar{Y}_2} \text{ et} \\ & m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)] \text{ ou } [n_{X_1\bar{Y}_1} = k_1 n_{X_2\bar{Y}_2} \text{ et } n_{\bar{X}_1Y_1} = k_1 n_{\bar{X}_2Y_2} \text{ et } n_{X_1Y_1} = k_2 n_{X_2Y_2} \text{ et} \\ & n_{\bar{X}_1\bar{Y}_1} = k_2 n_{\bar{X}_2\bar{Y}_2} \text{ et } m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)] \\ P_{15}(m) = 1 \text{ (invariance) si } & \forall (k_1, k_2) \in \mathbb{N}^{*2}, \forall X_1 \rightarrow Y_1, \forall X_2 \rightarrow Y_2 \\ [& (n_{X_1Y_1} = k_1 n_{X_2Y_2} \text{ et } n_{X_1\bar{Y}_1} = k_1 n_{X_2\bar{Y}_2} \text{ et } n_{\bar{X}_1Y_1} = k_2 n_{\bar{X}_2Y_2} \text{ et } n_{\bar{X}_1\bar{Y}_1} = k_2 n_{\bar{X}_2\bar{Y}_2}) \\ \Rightarrow & m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)] \text{ et } [(n_{X_1\bar{Y}_1} = k_1 n_{X_2\bar{Y}_2} \text{ et } n_{\bar{X}_1Y_1} = k_1 n_{\bar{X}_2Y_2} \text{ et } n_{X_1Y_1} = k_2 n_{X_2Y_2} \\ \text{et } & n_{\bar{X}_1\bar{Y}_1} = k_2 n_{\bar{X}_2\bar{Y}_2}) \Rightarrow m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)] \end{aligned}$$

Il est à noter que la formalisation de cette propriété par (Tan et al., 2002) à l'aide de matrices est plus compacte que ce que nous vous présentons mais nous avons recherché dans cet article une formalisation de même type pour toutes les propriétés.

Propriété 16 : Relation souhaitée entre les règles $X \rightarrow Y$ et $\bar{X} \rightarrow Y$

Cette propriété a été reprise de la première partie de la propriété T_3 .

$$P_{16}(m) = 0 \text{ si } \exists X \rightarrow Y / m(\bar{X} \rightarrow Y) \neq -m(X \rightarrow Y)$$

$$P_{16}(m) = 1 \text{ si } \forall X \rightarrow Y \quad m(\bar{X} \rightarrow Y) = -m(X \rightarrow Y)$$

Propriété 17 : Relation souhaitée entre les règles antinomiques $X \rightarrow Y$ et $X \rightarrow \bar{Y}$

Cette propriété a été reprise de la deuxième partie de la propriété T_3 .

$$P_{17}(m) = 0 \text{ si } \exists X \rightarrow Y / m(X \rightarrow \bar{Y}) \neq -m(X \rightarrow Y)$$

$$P_{17}(m) = 1 \text{ si } \forall X \rightarrow Y \quad m(X \rightarrow \bar{Y}) = -m(X \rightarrow Y)$$

Nous avons la liaison suivante avec la propriété 4 : si $P_4(m) = 0$ alors $P_{17}(m) = 0$.

Propriété 18 : Relation souhaitée entre les règles $X \rightarrow Y$ et $\bar{X} \rightarrow \bar{Y}$

Cette propriété a été reprise de la propriété T_4 .

$$P_{18}(m) = 0 \text{ si } \exists X \rightarrow Y / m(\bar{X} \rightarrow \bar{Y}) \neq m(X \rightarrow Y)$$

$$P_{18}(m) = 1 \text{ si } \forall X \rightarrow Y \quad m(\bar{X} \rightarrow \bar{Y}) = m(X \rightarrow Y)$$

Propriété 19 : Taille de la prémisse fixe ou aléatoire. Cette propriété est issue de la propriété L_5 .

$$P_{19}(m) = 0 \text{ (taille fixe)} \quad \text{si } m \text{ n'est pas fondée sur un modèle probabiliste}$$

$$P_{19}(m) = 1 \text{ (taille aléatoire)} \quad \text{si } m \text{ est fondée sur un modèle probabiliste}$$

Propriété 20 : Mesure descriptive ou statistique. Cette propriété est issue de la propriété L_{11} .

$P_{20}(m) = 0$ (descriptive ou invariante) si $\forall k \in \mathbb{N}^*$, $\forall X_1 \rightarrow Y_1$, $\forall X_2 \rightarrow Y_2$, ($n_{X_1 Y_1} = k n_{X_2 Y_2}$ et $n_{X_1 \bar{Y}_1} = k n_{X_2 \bar{Y}_2}$ et $n_{\bar{X}_1 Y_1} = k n_{\bar{X}_2 Y_2}$ et $n_{\bar{X}_1 \bar{Y}_1} = k n_{\bar{X}_2 \bar{Y}_2}$) $\Rightarrow m(X_1 \rightarrow Y_1) = m(X_2 \rightarrow Y_2)$

$P_{20}(m) = 1$ (statistique) si $\exists k \in \mathbb{N}^*$, $\exists X_1 \rightarrow Y_1$, $\exists X_2 \rightarrow Y_2$ / $n_{X_1 Y_1} = k n_{X_2 Y_2}$ et $n_{X_1 \bar{Y}_1} = k n_{X_2 \bar{Y}_2}$ et $n_{\bar{X}_1 Y_1} = k n_{\bar{X}_2 Y_2}$ et $n_{\bar{X}_1 \bar{Y}_1} = k n_{\bar{X}_2 \bar{Y}_2}$ et $m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)$

Propriété 21 : Mesure discriminante. Cette propriété est issue de la propriété L_{12} .

$P_{20}(m) = 0$ (non discriminante) si $\exists \eta \in \mathbb{N}^* / \forall n > \eta \quad \forall X_1 \rightarrow Y_1 \quad \forall X_2 \rightarrow Y_2$

$$[P(Y_1/X_1) > P(Y_1) \text{ et } P(Y_2/X_2) > P(Y_2)] \Rightarrow m(X_1 \rightarrow Y_1) \approx m(X_2 \rightarrow Y_2)$$

$P_{20}(m) = 1$ (discriminante) si $\forall \eta \in \mathbb{N}^* \exists n > \eta \quad \exists X_1 \rightarrow Y_1 \quad \exists X_2 \rightarrow Y_2$ /

$$P(Y_1/X_1) > P(Y_1) \text{ et } P(Y_2/X_2) > P(Y_2) \text{ et } m(X_1 \rightarrow Y_1) \neq m(X_2 \rightarrow Y_2)$$

Après avoir formalisé les propriétés, nous allons les étudier sur différentes mesures objectives.

4 Étude des propriétés des mesures objectives d'intérêt

Cette section va donc étudier, pour différentes mesures d'intérêt objectives, la présence ou l'absence des propriétés mises en évidence dans la *section 2* et formalisées dans la *section 3*. Ce travail va déboucher sur la construction d'une matrice, matrice que l'on peut visualiser dans le *tableau 1*.

Cette matrice étudie 69 mesures dont 46 proviennent des travaux de synthèse (Piatetsky-Shapiro 1991, Tan et al. 2002, Lallich et Teytaud 2004, Geng et Hamilton 2007, Feno 2007

et Vaillant 2007) exposés dans la *section 2*. Neuf mesures extraites de (Huynh et al., 2006) ont également été étudiées. Ces mesures sont les suivantes : *confiance causale*, *confiance confirmée causale*, *confiance confirmée descriptive*, *confirmation causale*, *confirmation descriptive*, *dépendance*, *dépendance causale estimée*, *Pavillon et support causal*.

Pour finir, les mesures restantes sont les suivantes : *Czekanowski-Dice* (Czekanowski, 1913), *Fukuda* (Fukuda et al., 1996), *Ganascia* (Ganascia, 1987), *indice probabiliste d'écart à l'équilibre* (Blanchard et al., 2005a), *indice probabiliste d'écart à l'équilibre entropique* (Blanchard et al., 2005b), *intensité d'implication entropique* (Gras et al., 2001), *indice de vraisemblance du lien* (Lerman, 1981), *Kappa* (Cohen, 1960), *Kulczynski* (Kulczynski, 1928), M_{GK} (Guillaume, 2000), *Ochiai* (Ochiai, 1957), *satisfaction* (Lavrac et al., 1999) et *VT100* (Morineau et Rakotomalala, 2006).

Quant aux 21 propriétés mises en évidence dans la *section 3*, 19 sont étudiées. En effet, les propriétés " P_1 : intelligibilité ou compréhensibilité de la mesure" et " P_2 : facilité à fixer un seuil d'acceptation de la règle" n'ont pas été retenues dans cette étude car nous pensons qu'elles sont subjectives et fonction de la connaissance qu'a l'utilisateur en statistique. B. Vaillant (Vaillant, 2007) a également souligné ce point dans sa thèse.

Après avoir rempli cette matrice, nous indiquons quelques remarques.

- Tout d'abord, si la mesure est strictement croissante en fonction du nombre n_{XY} d'exemples et qu'il y a une valeur fixe a dans le cas de l'indépendance ($P_9(m) = 1$), alors nous aurons des valeurs identifiables dans le cas de l'attraction ($P_{12}(m) = 1$) et de la répulsion ($P_{13}(m) = 1$). Nous ne pouvons pas dégager de règle avec les propriétés car la *propriété 6* ne vérifie pas la croissance stricte.

- Pour finir, nous avons étudié la *propriété 5*, cas où la mesure évalue de la même façon $X \rightarrow Y$ et $\bar{Y} \rightarrow \bar{X}$ dans le cas de l'implication logique. Peut-être serait-il intéressant de créer un troisième cas : cas où la mesure évalue de la même façon $X \rightarrow Y$ et $\bar{Y} \rightarrow \bar{X}$ dans toutes les situations et non pas uniquement dans le cas de l'implication logique. Nous avons plusieurs mesures vérifiant cette propriété (*coefficient de corrélation*, *Cohen*, *confiance causale*, *confirmation causale*, *conviction*, *facteur de certitude*, *force collective*, *Goodman-Kruskal*, *Kappa*, *loevinger*, *nouveauté*, *Pearl*, *Piatetsky-Shapiro*, *précision*, *Q*, *ratio des chances*, *support causal*, *variation de support*, *Y*).

5 Conclusion et perspectives

Cet article synthétise et propose une formalisation des "bonnes" propriétés des mesures d'intérêt rencontrées dans la littérature. Ces propriétés ont ensuite été étudiées sur 69 mesures objectives afin d'en fournir une caractérisation. Cette étude a permis la construction d'une matrice évaluant chaque mesure sur 19 propriétés, étendant de manière considérable de précédents travaux de la littérature.

La formalisation de ces propriétés nous paraît essentielle afin d'éliminer toute interprétation possible de celles-ci, comme par exemple considérer une croissance stricte ou non pour les *propriétés 6*, *7* et *8*, pouvant conduire à la construction de matrices différentes.

Ce travail est le point de départ pour une catégorisation des mesures afin d'aider l'utilisateur dans le choix de ses mesures en fonction de ses objectifs, du domaine d'étude et des données avec lesquelles il travaille, notamment dans la phase de post-traitement en fouille de données. De plus, cette étude va permettre de rechercher s'il y a des propriétés redondantes parmi celles qui ont été mises en évidence.

Summary

Finding interesting association rules is an important and active research field in data mining. The algorithms of the *Apriori* family are based on two measures to extract the rules, support and confidence. Although these two measures have accelerators algorithmic virtues, they generate a prohibitive number of rules most of which are redundant and irrelevant. It is therefore a need for further measures filtering uninteresting rules. This article synthesizes different reported works to identify the "good" measures properties for extraction rules to retain those who are interesting for the user. All these properties are then formalized and assessed on more than sixty measures, significantly extending the previous work of literature. This synthesis is the starting point for measures categorization as well as for the detection of redundant properties.

Références

- Agrawal R. et Srikant R. (1994), Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, Santiago, Chile.
- Blanchard J., Guillet F., Briand H. and Gras R. (2005a). IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. Dans *l'Atelier Qualité des Données et des Connaissances*, p. 26–34.
- Blanchard J., Guillet F., Briand H. and Gras R. (2005b), Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. In *Troisièmes rencontres internationales de l'Analyse Statistique Implicative (ASI 05)*.
- Cohen, J. (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20 :37–46.
- Czekanowski J. (1913). *Zarys metod statystycznych (Die Grundzuge der statischen Methoden)*, Warsaw.
- Feno D.J. (2007). Mesures de qualité des règles d'association : normalisation et caractérisation des bases, PhD thesis, Université de La Réunion.
- Freitas, A. (1999). On rule interestingness measures. *Knowledge-Based Systems journal*, pages 309–315.
- Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T. (1996). Data mining using two-dimensional optimized association rules : scheme, algorithms, and visualization. In *ACM SIGMOD International Conference on Management of Data*, pages 13–23.
- Ganascia, J.-G. (1987). Charade : A rule system learning system. In *10th International Joint Conference on Artificial Intelligence*, Milan.
- Geng L. and Hamilton H.J. (2007), Choosing the Right Lens : Finding What is Interesting in Data Mining. *Quality Measures in Data Mining*, pages 3–24, ISBN 978-3-540-44911-9.
- Gras R., Kuntz P., Couturier R. and Guillet F. (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des connaissances et apprentissage (EGC 2001)*, 1(1-2) :69–80.
- Guillaume, S. (2000), Traitement des données volumineuses : mesures et algorithmes d'extraction de règles d'association et règles ordinales, PhD thesis, Université de Nantes.
- Hofmann, H. and Wilhelm, A. (2001). Visual comparison of association rules. *Computational Statistics*, 16(3) :399–415.

Propriétés des mesures d'intérêt

- Huynh, X.-H., Guillet, F., and Briand, H. (2006). Arqat : plateforme exploratoire pour la qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (EGC : Etat et perspectives)*, (RNTI-E-5).
- Kulczynski, S. (1928). Die Pflanzenassoziationen der Pieninen, *Bull. Int. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat.*, Ser. B, Suppl. II (1927), pp. 57-203.
- Lallich, S. (2002). *Mesure et validation en extraction des connaissances à partir des données*, Habilitation à diriger les recherches, Université Lyon 2.
- Lallich, S. et Teytaud, O. (2004). Evaluation et validation de mesures d'intérêt des règles d'association, *RNTI-E-1*, numéro spécial, p.193-217.
- Lavrac N., Flach P., and Zupan B. (1999), Rule evaluation measures : A unifying view. In G. Mineau and B. Ganter, editors, *Ninth International workshop on Inductive Logic Programming*, volume 1634, p. 174–185.
- Lenca, P., Meyer, P., Picouet, P., Vaillant, B., and Lallich, S. (2003a). Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de données)*, (1) :123–134.
- Lenca P., Meyer P., Vaillant B., and Picouet P. (2003b), Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. *RSTI-RIA (Extraction et Gestion des Connaissances 2003)*, 1(17) :271–282.
- Lerman I.C. (1981). Classification et analyse ordinale des données, Dunod.
- Morineau A. and Rakotomalala R. (2006). Critère vt100 de sélection des règles d'association. In Cépaduès, editor, *Actes de Extraction et Gestion de Connaissances, EGC'2006*, pages 581–592.
- Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, *Bull. Jpn. Soc. Sci. Fish.*, 22, pp. 526-530.
- Padmanabhan, B. and Tuzhilin, A. (1998). A belief-driven method for discovering unexpected patterns. In *ACM International Conference on Knowledge Discovery and Data Mining*, p. 94–100. ACM Press.
- Piatetsky-Shapiro (1991) G. Piatetsky-Shapiro, "Discovery, Analysis and Presentation of Strong Rules", In G. Piatetsky-Shapiro & W.J. Frawley, editors, *Knowledge Discovery in Databases*, AAAI Press, pages 229-248.
- Sese, J. and Morishita, S. (2002). Answering the most correlated n association rules efficiently. In *Proceedings of the 6th European Conf on Principles of Data Mining and Knowledge Discovery*, pages 410–422. Springer-Verlag.
- Tan, P.N., Kumar, V., Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 32-41.
- Vaillant, B. (2002). Evaluation de connaissances : le problème du choix d'une mesure de qualité en extraction de connaissances à partir des données. Master's thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne.
- Vaillant, B. (2007). Mesurer la qualité des règles d'association : études formelles et expérimentales, PhD thesis, ENST Bretagne.
- Zaki, M. (2000). Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pages 34–43.
- Zaman Ashrafi, M., Taniar, D., and Smith, K. (2004). A new approach of eliminating redundant association rules. In Galindo, F. and Takizawa, M. Traunmüller, R., editors, *Database and Expert Systems Applications*, volume 3180 of *Lecture Notes in Computer Science*, pages 465–474, Zaragoza, Spain. Springer.

Mesure	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁	P ₁₂	P ₁₃	P ₁₄	P ₁₅	P ₁₆	P ₁₇	P ₁₈	P ₁₉	P ₂₀	P ₂₁
coefficient corrélation	0	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1
Cohen	0	1	1	1	1	1	1	0	0	1	1	1	0	0	0	1	0	0	1
confiance	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1
confiance causale	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1
confiance centrée	1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	1
confiance confirmée descriptive	1	1	1	1	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1
confiance confirmée causale	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1
confirmation causale	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1
confirmation descriptive	1	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1
conviction	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1
cosinus	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
couverture	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Czekanowski	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
dépendance	0	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
dépendance causale	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1
dépendance pondérée	1	1	0	0	1	0	0	0	0	0	0	I ¹	0	0	0	0	0	0	1
facteur bayésien	1	1	0	1	1	1	1	0	0	1	1	0	1	0	0	0	0	0	1
facteur certitude	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	1
fiabilité négative	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1
F-mesure	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
force collective	0	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	1
Fukuda	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
gain informationnel	0	1	0	1	1	1	1	0	0	1	1	2	0	0	0	0	0	0	1
Ganascia	1	1	1	1	0	0	0	1	1	0	0	1	0	0	1	0	0	0	1
Gini	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Goodman	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1
indice d'implication	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1
IPEE	1	1	1	1	0	0	0	0	1	0	0	2	0	0	0	0	1	1	1
IP3E	1	1	1	1	0	0	0	0	1	0	0	2	0	0	0	0	1	1	1
IPD	0	1	0	1	I	1	0	0	0	0	0	2	1	0	0	0	1	1	1
information mutuelle	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
II	1	1	1	1	1	1	1	0	0	1	1	2	1	0	0	0	1	1	0
IIE	1	1	1	1	1	0	0	0	0	0	0	2	0	0	0	0	1	1	1
IHER	1	1	1	1	1	0	1	0	0	1	1	2	0	0	0	0	1	1	1
IVL	0	1	0	1	1	1	1	0	0	1	1	2	0	0	0	0	1	1	0
intérêt	0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1

¹ I : indéterminé car fonction des paramètres k et m .

Propriétés des mesures d'intérêt

Mesure	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁	P ₁₂	P ₁₃	P ₁₄	P ₁₅	P ₁₆	P ₁₇	P ₁₈	P ₁₉	P ₂₀	P ₂₁
Jaccard	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
J-mesure	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
Kappa	0	1	1	1	1	1	1	0	0	1	1	1	0	0	0	1	0	0	1
Kloggen	1	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1
Kulezynski	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Laplace	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
Leverage	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1
Loevinger	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	1
M _{GK}	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	0	0	1
moindre contradiction	1	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
nouveauté	0	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1
Ochiai	0	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
Pavillon	1	1	0	1	1	1	1	0	0	1	1	1	0	0	1	0	0	0	1
Pearl	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1
Piatetsky-Shapiro	0	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	1	1
précision	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1
prévalence	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Q de Yule	0	1	1	1	1	0	1	1	0	1	1	2	1	1	1	1	0	0	1
rappel	1	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
ratio des chances	0	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	1
risque relatif	1	1	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1
satisfaction	1	1	0	1	1	0	1	0	0	1	1	1	0	0	0	0	0	0	1
Sebag	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
spécificité	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1
support	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
support causal	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	1	0	0	1
support sens unique	1	1	0	0	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1
support double sens	0	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1
taux exemples	1	1	0	1	0	0	0	1	1	0	0	2	0	0	0	0	0	0	0
VT100	0	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	1
variation support	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1
Y de Yule	0	1	1	1	1	0	1	1	0	1	1	0	1	1	1	1	0	0	1
Zhang	1	1	0	1	1	0	1	1	0	1	1	2	1	0	1	0	0	0	1

Tab 1 : Matrice des propriétés.

Mesure de la robustesse de 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 choisi par l'utilisateur. 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 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

Une méthode très populaire pour évaluer l'intérêt des règles d'association consiste à quantifier cet intérêt à l'aide de mesures objectives. 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) et Suzuki (2008).

On rappelle qu'une règle d'association $A \rightarrow B$, extraite d'une base de données \mathcal{B} , est déclarée de qualité, selon la mesure μ et le seuil σ_μ (fixé par l'utilisateur), si $\mu(A \rightarrow B) \geq \sigma_\mu$. 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 encore si le seuil d'acceptation avait été

augmenté, même légèrement (inversement 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 : 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.

A notre connaissance, très peu de travaux se sont concentrés sur la robustesse des règles d'association. 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, 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.

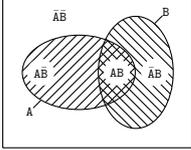
2 Robustesse

2.1 Règles d'association et mesures d'intérêt

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 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 . Notre approche est fondée sur ce cadre, dont nous rappelons ici les notions principales. Notons $r : A \rightarrow B$ 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 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 x .

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. Dans nos précédents travaux (Le Bras et al., 2009b), nous avons 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

	B	\bar{B}	
A	p_{ab}	$p_{a\bar{b}}$	p_a
\bar{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 $r : A \rightarrow B$

de bien définir le domaine de définition. Ce domaine dépend de la paramétrisation choisie : à l'aide des exemples, des contre-exemples, ou encore de la confiance. Guillaume (2000), Vaillant et al. (2006) et Lenca et al. (2008) ont souligné l'importance du nombre de contre-exemples dans l'évaluation de l'intérêt des règles. Par conséquent, nous nous intéresserons ici au comportement des mesures vis-à-vis de la variation des contre-exemples des règles, c'est-à-dire qu'une règle d'association $r : A \rightarrow B$ 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 (Le Bras et al. (2009a)) :

$$\mathcal{D} = \left\{ (x, y, z) \mid \begin{array}{l} 0 < y < 1 \\ 0 < z < 1 \\ \max(0, y - z) < x < \min(y, 1 - z) \end{array} \right\}$$

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} grâce à une mesure d'intérêt objective μ . L'utilisateur aura fixé un seuil, μ_{\min} , au dessus duquel les règles sont jugées intéressantes. Les règles ainsi sélectionnées sont dépendantes :

- du seuil μ_{\min} : l'utilisateur peut à tout moment modifier le seuil et faire ainsi apparaître, ou disparaître, un grand nombre de règles ;
- du 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. Cependant, les auteurs n'ont pas fourni de modèle général agrégeant leurs différentes propositions.

Notre vision de la robustesse est différente et s'appuie sur la notion de règle limite, qui peuvent être abstraites, au sens où elles ne sont pas nécessairement réalisées dans la base \mathcal{B} :

Définition 1 (Règle limite). Une règle limite est une règle d'association r_{\min} , éventuellement abstraite (voir ci-dessous), telle que $\mu(r_{\min}) = \mu_{\min}$. Soit r une règle d'association, on note r^* une règle limite qui minimise $\|r - r_{\min}\|_2$ dans \mathbb{R}^3 .

Mesure de robustesse de règles d'association

Ce sont des règles qui, si elles étaient réalisées, seraient sélectionnées de justesse (par rapport au seuil μ_{\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 $(x_{\min}, y_{\min}, z_{\min})$ est nécessairement un élément de \mathcal{D} . Ainsi, $\|r - r^*\|_2$ n'est pas simplement la distance de r à la surface \mathcal{S} d'équation $\mu = \mu_{\min}$, mais la distance à $\mathcal{S} \cap \mathcal{D}$.

Définition 2 (Robustesse d'une règle). Soit μ une mesure d'intérêt des règles d'association et μ_{\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 μ_{\min} par :

$$\text{rob}_\mu(r, \mu_{\min}) = \frac{\|r - r^*\|_2}{\sqrt{3}}$$

Le facteur important est le numérateur $\|r - r^*\|_2$, 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 $r : A \rightarrow B$ 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 :

$$|p'_{a\bar{b}} - p_{a\bar{b}}| \leq \frac{\|r - r^*\|_2}{\sqrt{3}} ; |p'_a - p_a| \leq \frac{\|r - r^*\|_2}{\sqrt{3}} ; |p'_b - p_b| \leq \frac{\|r - r^*\|_2}{\sqrt{3}}$$

Ainsi, $\|r - r'\|_2 = \sqrt{|p'_{a\bar{b}} - p_{a\bar{b}}|^2 + |p'_a - p_a|^2 + |p'_b - p_b|^2} \leq \|r - r^*\|_2$ et donc, par définition de r^* , $\mu(r') > \mu_{\min}$. $\text{rob}(r)$ traduit donc la quantité de bruit acceptée par la règle pour rester de qualité. C'est une notion de sécurité, qui permet d'affirmer que si le bruit est suffisamment contrôlé, la règle restera intéressante. Notons cependant qu'une règle peu robuste pourra évoluer de manière à devenir plus robuste et rester intéressante.

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 $\text{rob}(r)\%$ des transactions, la règle r restera intéressante par rapport à μ_{\min} . 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.

Propriété 1. La robustesse $\text{rob}(r)$ 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 seront robustes et plus il sera important d'avoir des données fiables. D'autre part, deux règles dont les projetés sont proches auront 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 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. Klossgen, 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, précision, recall, spécificité. Dans ce cas, 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. L'algorithme de calcul de la robustesse dans le cas d'une mesure plane est donc le suivant :

- Trouver r^\perp , projection orthogonale de r sur \mathcal{P} ;
- Si $r^\perp \in \mathcal{D}$, $r^* = r^\perp$ et renvoyer $\|r - r^*\|_2$;
- Sinon, renvoyer la distance au périmètre du polygone intersection.

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 : $(1 + \mu_0)x - y = 0$;
- taux d'exemples-contre-exemples : $(2 - \mu_0)x - (1 - \mu_0)y = 0$;
- mesure de Jaccard : $(1 + \mu_0)x - y + \mu_0z = 0$.

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 , μ_{\min} et de μ . En pratique dans nos expériences nous n'avons pas rencontré cette valeur.

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

$$d = y_1 \frac{\mu(r) - \mu_{min}}{\sqrt{1 + (1 - \mu_{min})^2}} \quad (1)$$

La robustesse dépend donc, à μ_{min} fixé, de deux paramètres : y_1 , le support de l'antécédent ; et $\mu(r)$, 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 être de confiances variées. 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.

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 notée par les mesures, mais de robustesse très faible et une règle moins bien notée, 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. On peut distinguer quatre comportements types. S'il est facile d'arbitrer entre deux règles (robuste/intéressante) et (fragile/peu intéressante), la tâche est moins évidente entre deux règles (robuste/peu intéressante) et (fragile/intéressante). Vaut-il mieux avoir une règle très intéressante, mais très dépendante du bruit dans les données, ou bien est-il mieux 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 évidemment de la situation pratique et de la confiance que l'utilisateur a dans la qualité des données.

Nous allons le voir dans la suite, les graphiques robustesse/mesure font apparaître un grand nombre de règles robustes, mais dominées en terme de mesure par des règles moins robustes.

3 Expériences

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 spécificité. Le tableau 1 rappelle leur écriture en fonction des contre-exemples, ainsi que le plan qu'elles définissent.

nom	formule	plan	seuil
confiance	$\frac{p_a - p_{a\bar{b}}}{p_a}$	$x - (1 - \mu_0)y = 0$	0.984
Jaccard	$\frac{p_a - p_{a\bar{b}}}{p_b + p_{a\bar{b}}}$	$(1 + \mu_0)x - y + \mu_0z = 0$	0.05
Sebag-Shoenauer	$\frac{p_{a\bar{b}}}{p_a - p_{a\bar{b}}}$	$(1 + \mu_0)x - y = 0$	10
spécificité	$\frac{1 - p_b - p_{a\bar{b}}}{1 - p_a}$	$x - \mu_0y + z = 1 - \mu_0$	0.5
taux exemples-contre-exemples	$1 - \frac{p_{a\bar{b}}}{p_a - p_{a\bar{b}}}$	$(2 - \mu_0)x - (1 - \mu_0)y = 0$	0.95

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

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 positif, 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.

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, 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

Mesure de robustesse de règles d'association

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.

indiquées table 1. 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. 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 2 propose un échantillon représentatif des résultats, dans le sens où l'allure des graphiques est sensiblement la même pour toutes les bases, pour une mesure donnée. Plusieurs points peuvent être relevés.

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 supportera 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, la Spécificité semble présenter des courbes convexes. Traitons le cas particulier des courbes relatives à la

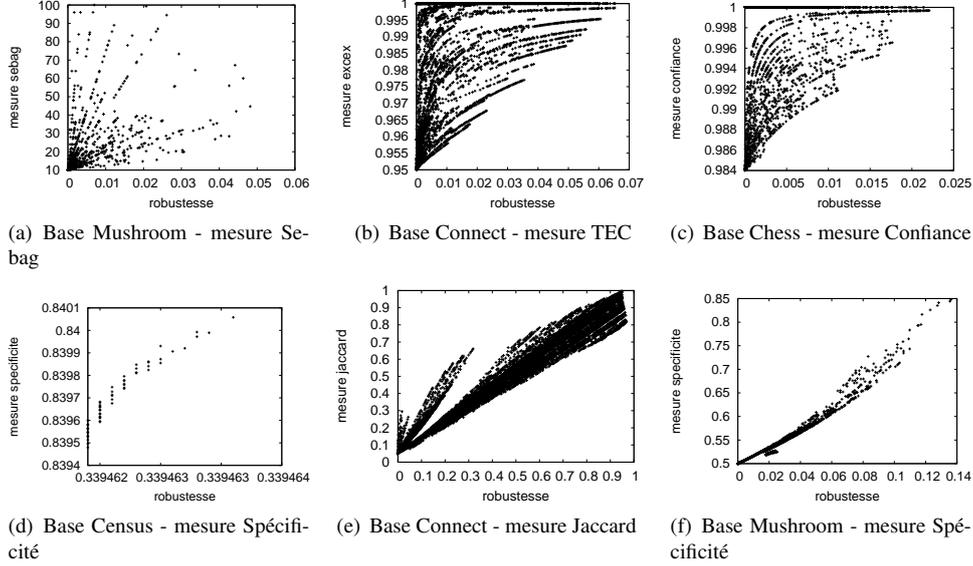


FIG. 2 – Valeur de la mesure en fonction de la robustesse pour différents couples base/mesure.

confiance. Une telle démonstration peut se faire pour les autres mesures. L'équation (1) montre l'écriture de la robustesse en fonction de la mesure, où y représente p_a . Puisque $p_a = \frac{p_{a\bar{b}}}{1-\mu(r)}$, on peut écrire $\mu(r)$ en fonction de d :

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

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 quelque 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.

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 (cf. 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 3 montre les résultats obtenus

Mesure de robustesse de règles d'association

(a) mesure TEC			(b) mesure de Sebag			(c) mesure de spécificité		
base	disparues	conservées	base	disparues	conservées	base	disparues	conservées
census	0.83e-6	0.79e-6	census	1.53e-6	1.53e-6	census	0	0.19
chess	1.16e-3	0.96e-2	chess	1.63e-3	1.72e-2	chess	7.23e-5	8.76e-2
connect	5.26e-4	7.72e-3	connect	8.38e-4	1.42e-2	connect	0	1.2e-1
mushroom	9.4e-5	6.6e-4	mushroom	1.28e-4	1.22e-3	mushroom	2.85e-4	1.37e-2

(d) mesure de confiance			(e) mesure de Jaccard		
base	disparues	conservées	base	disparues	conservées
census	2.61e-7	2.61e-7	census	0	0
chess	5.59e-4	3.77e-3	chess	3.2e-4	1.69e-1
connect	2.16e-4	2.73e-3	connect	1.94e-3	1.43e-1
mushroom	5.51e-5	2.34e-4	mushroom	3.20e-4	1.90e-2

TAB. 3 – Comparaison entre les robustesses moyennes des règles disparues et conservées pour les différentes mesures

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 présentées par la base de données Census sont de l'ordre de 10^{-6} et sont donc vulnérables à 5% de bruit, donc 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

- Asuncion, A. et D. Newman (2007). UCI machine learning repository.
- Azé, J., S. Guillaume, et P. Castagliola (2003). Evaluation de la résistance au bruit de quelques mesures quantitatives. *n° 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.
- Borgelt, C. et R. Kruse (2002). Induction of association rules : APRIORI implementation. In *15th Conference on Computational Statistics, Berlin, Germany*, pp. 395–400.
- 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.
- 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° 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.
- 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.
- Lallich, S. et O. Teytaud (2004). évaluation et validation de l’intérêt des règles d’association. *n° 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.
- 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 conférence, 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° 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° 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.
- 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.

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 and 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.

Une approche basée sur l'agrégation pour une meilleure détection d'intrusions

Nouria Harbi*, Emna Bahri*

*Laboratoire ERIC

5, avenue Pierre Mendès-France, 69500 Bron

nouria.harbi | emna.bahri@univ-lyon2.fr,

<http://eric.univ-lyon2.fr>

Résumé. Actuellement, les recherches en fouille de données et plus précisément en apprentissage automatique se sont focalisées sur le domaine de détection d'intrusions dans les systèmes informatiques. En effet, la détection des attaques et des anomalies dans les réseaux informatiques est considérée comme un problème de classification de données d'où l'utilisation des techniques du domaine de la fouille de données. Plusieurs approches d'apprentissage automatique sont utilisées sur des masses de données complexes et dynamiques afin de construire un système de détection d'intrusions efficaces (IDS). Cependant, ces approches se trouvent confrontées à des problèmes de précision et surtout à l'augmentation des fausses alertes "faux positifs", face à des données complexes et déséquilibrées avec un trafic réseaux de grande vitesse. Afin d'améliorer la détection de ces attaques et par conséquent diminuer le taux d'erreur en classification, nous proposons dans cet article une nouvelle approche basée sur le principe d'agrégation de classifieurs ayant pour but d'améliorer la performance d'un classifieur par une méthode de vote. En fait, notre approche est hybride, puisqu'elle classe à la fois les connexions normales et les attaques. Elle utilise une version de boosting adapté aux données réseaux que nous proposons pour réduire principalement les "faux positifs".

1 Introduction

Le développement rapide des technologies et des services mobiles est issu des récentes innovations techniques et de la demande grandissante des utilisateurs. Cette évolution se traduit par l'intégration massive de technologie communicante. Cette émergence des technologies de la mobilité autorise le personnel à exercer ses fonctions dans des espaces autres que celui de l'entreprise et dans des plages horaires différentes de la normale, nous parlerons alors de nomadisme. De plus, la généralisation des liaisons haut débit et la multiplication des accès distants (extranet, intranet, télétravail, cybercafé) laissent l'information de l'entreprise accessible à partir de n'importe quel endroit et à chaque instant grâce aux réseaux virtuels privés (VNP). De ce fait, chaque connexion augmente la vulnérabilité du réseau par rapport aux agressions. Parallèlement, le nombre de problèmes de sécurité, en particulier les intrusions par l'Internet,

a augmenté de manière très importante ces dernières années, et cette courbe ascendante ne devrait malheureusement pas s'infléchir. Il est donc nécessaire de se protéger.

En effet, la sécurité des systèmes informatiques constitue un enjeu crucial pour la survie de l'entreprise, surtout face à des politiques de sécurité non efficaces. Ces limitations justifient le recours à des mécanismes de détection d'intrusions (*Intrusion Detection Systems*, IDS) qui reposent essentiellement sur l'analyse du contenu des données réseaux (trames), à la recherche de traces d'attaques connues. Actuellement, les IDS deviennent une étape principale des dispositifs de sécurité. Afin de détecter les intrusions, plusieurs algorithmes d'apprentissage supervisé tels que les réseaux de neurones [2], les Supports Vecteurs Machine [16] et les algorithmes génétiques [18] sont souvent utilisés sur des masses de données complexes pour classifier des attaques connues et non connues. En fait, les IDS ont recours à ces algorithmes pour résoudre le problème d'analyse de données volumineuses en améliorant la performance de détection d'intrusions qui se trouve face à une évolution exponentielle de nouvelles attaques.

Dans cet article, nous présentons une nouvelle approche de détection d'intrusions basée sur le principe d'agrégation de décisions adapté aux données réseaux afin d'améliorer la fiabilité du système de détection d'intrusions. En fait, l'agrégation de décisions, utilisée dans l'analyseur de l'IDS, réduit le nombre de fausses alertes "faux positifs" et surtout le nombre d'attaques non détectées "faux négatifs". Cette approche hérite à la fois des bases de l'approche comportementale et de l'approche par scénarios.

Cet article est organisé comme suit. En section 2, nous présentons les systèmes de détection d'intrusions actuellement utilisés ainsi que leurs architectures et les approches déjà utilisées. Dans la section 3, nous présentons notre approche hybride en détaillant le principe d'agrégation ainsi que l'algorithme utilisé. Les données expérimentales ainsi que les résultats obtenus sont présentés en section 4. Enfin, nous terminons par une conclusion et des perspectives.

2 Les systèmes de détection d'intrusions : IDS

La détection d'intrusions a pour objectif de déceler toute violation de la politique de sécurité en vigueur sur un système informatique [6] [3]. Il n'est donc pas suffisant d'agir préventivement, c'est-à-dire de définir une politique de sécurité (en termes de confidentialité, d'intégrité, de disponibilité des données et ressources du système à protéger) et de mettre en oeuvre des mécanismes implantant cette politique. Il faut aussi être capable de détecter toute tentative de violation de la politique de sécurité, c'est-à-dire toute intrusion. La détection d'intrusions a été introduite en 1980 par J.P Anderson qui a été le premier à montrer l'importance de l'audit de sécurité [9] dans le but de détecter les éventuelles violations de la sécurité d'un système.

Architecture d'un système de détection d'intrusions DIS

Un système de détection d'intrusions est constitué classiquement de trois composants [3]. La figure 1 illustre les interactions entre ces trois composants. Un capteur est chargé de collecter des informations sur l'évolution de l'état du système et de fournir une séquence d'événements qui renseignent sur l'évolution de l'état du système. Un analyseur détermine si un sous-ensemble des événements produits par le capteur est caractéristique d'une activité malveillante. Un manager collecte les alertes produites par le capteur, les met en forme et les présente à l'opérateur. Éventuellement, le manager est chargé de la réaction à adopter.

Nous détaillons dans cet article seulement l'analyseur puisqu'il constitue la partie essentielle de la détection.

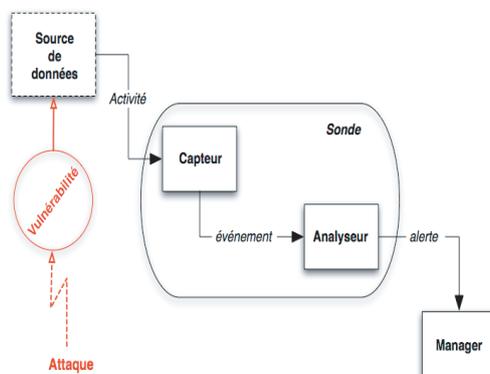


FIG. 1: Architecture d'un IDS

Analysateur

L'objectif de l'analyseur est de déterminer si le flux d'événements fourni par le capteur contient des éléments caractéristiques d'une activité malveillante. Deux grandes approches ont été proposées : l'approche comportementale (*anomaly detection*) et l'approche par scénarios (*misuse detection*);

- Dans l'approche comportementale, une attaque est qualifiée par la mesure d'une déviation sensible du système surveillé par rapport à un comportement de référence, réputé sain et défini auparavant.
- Dans l'approche par scénarios, le système de détection possède une base de signatures qui modélisent les différents scénarios, c'est-à-dire les différentes attaques connues. L'analyse consiste à rechercher l'occurrence d'un motif caractéristique d'une attaque dans le flux d'événements.

L'analyseur doit détecter de manière automatique les violations de la politique de sécurité, qu'on appelle intrusions. Dans la pratique, les outils actuels ne sont pas configurés directement par les instances de sécurité. Ainsi, s'ils détectent certaines intrusions, ils détectent aussi des tentatives d'intrusions infructueuses, ce qui n'est pas souhaitable. En outre, la relative naïveté des algorithmes de détection conduit à un nombre élevé d'alertes, dont une part significative est en fait constituée de fausses alertes (faux positifs). Enfin, certaines intrusions peuvent ne pas être détectées (faux négatifs), figure 2.

Nous détaillons par la suite les deux approches utilisées actuellement et surtout leurs limites en nous référant notamment à [12].

Approche comportementale

Cette approche consiste à modéliser des comportements normaux pour détecter les comportements interdits.

Plusieurs méthodes ont été utilisées afin de construire ces comportements (profils) telles que :

- *Méthodes statistiques* : le profil est calculé à partir de variables considérées comme aléatoires et échantillonnées à intervalles réguliers. La distribution de chaque variable est modélisée pour mesurer le caractère inattendu du comportement courant.

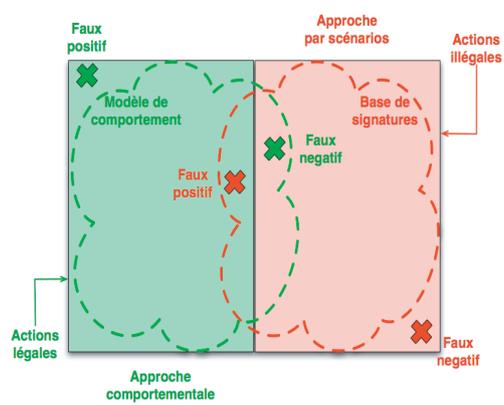


FIG. 2: Problèmes de la fiabilité des IDS

- *Immunologie* : cette analogie avec l’immunologie biologique consiste à construire un modèle de comportement normal des services (et non des utilisateurs) au travers de courtes séquences d’appels système qui sont considérées comme représentatives de l’exécution normale des services considérés. Cette approche est utilisée dans [8].
- *Réseaux de neurones et Graphes* : certaines approches comportementales utilisent des modèles à base de graphes pour mettre en évidence des propriétés et des relations entre ces propriétés comme des modèles dérivés des modèles de Markov ou les réseaux de neurones [7] et [14]
- *Approche bayésienne* : les réseaux Bayésiens permettent de modéliser des situations dans lesquelles la causalité joue un rôle, mais où la connaissance de l’ensemble des relations entre les phénomènes est incomplète, de telle sorte qu’il est nécessaire de les décrire de manière probabiliste [11].
- Plusieurs mécanismes ont également été proposés afin de localiser les signatures d’attaques dans les traces d’audit (systèmes experts, algorithmes génétiques).

L’approche comportementale est intéressante car, ne faisant aucune hypothèse sur les comportements illégaux, elle permet en théorie de détecter de nouvelles formes d’intrusions (attaques dites *zero-day*). Cependant, la définition du profil de référence (comportement normal) par apprentissage reste délicate, puisque la phase d’apprentissage s’effectue sur des données enregistrées avant la mise en place de l’IDS. En effet, cette phase requiert une base de données à la fois saine et exhaustive par rapport au comportement attendu des utilisateurs dans l’environnement réel. Pour éviter la mise en place d’un profil trop rigide et pour pouvoir s’adapter aux changements de comportement des utilisateurs, certains IDS proposent des phases de réapprentissage au cours de l’utilisation de l’IDS. Il reste tout de même un risque qu’un attaquant arrive à modifier le profil de référence à son avantage par déviation progressive. De manière générale, fixer des seuils peut s’avérer délicat et les résultats peuvent être pénalisés par un sur-apprentissage.

Approche par scénarios

L’approche par scénarios est actuellement la plus commune. Elle s’appuie sur une base de signatures d’attaque. Le système de détection consiste alors à reconnaître la présence de

signatures parmi les traces d'audit fournies par les observateurs. Plusieurs techniques ont été proposées qui reposent en général sur des mécanismes de reconnaissance de motifs (*pattern matching*) [10]. Le *pattern matching* possède l'avantage d'être une méthode fiable car déterministe. Cependant, la difficulté vient de la définition des motifs. En effet, ceux-ci doivent être suffisamment précis pour pouvoir discriminer les différents types d'attaques, mais suffisamment génériques pour pouvoir détecter les différentes variantes d'un même type d'attaque. Une signature trop générique conduira à l'augmentation du nombre de faux positifs, diminuant par là même la fiabilité. La technique de détection par scénarios nécessite en outre une maintenance active du système pour mettre à jour régulièrement la base des signatures. En théorie, cette approche devrait produire peu de faux positifs (une connexion normale détectée comme étant une attaque) car le système utilise une connaissance a priori sur les attaques. Les techniques de ce type restent toutefois faciles et rapides à mettre en oeuvre. Mais, le problème de la fiabilité reste d'actualité concernant les fausses alertes.

Face à ces limites, il nous paraît intéressant de combiner ces deux approches et d'améliorer la fiabilité d'un IDS en réduisant le nombre de faux positifs et les faux négatifs. C'est dans ce cadre que nous nous sommes orientés vers le principe d'agrégation de classifieurs. En effet, de nombreuses recherches ont montré que l'agrégation agit directement sur l'erreur en généralisation à travers la minimisation du biais et/ou de la variance d'un classifieur unique et peut permettre ainsi de réduire le nombre de faux positifs et de faux négatifs. Dans la section suivante, nous détaillons diverses techniques d'agrégation ainsi que notre méthode hybride pour une meilleure détection d'intrusion.

3 Agrégation pour une meilleure détection d'intrusions

Disposer de systèmes de détection d'intrusions efficaces nous apparaît être un enjeu essentiel pour la sécurité des systèmes informatiques. Cette efficacité peut être obtenue par la coopération entre plusieurs techniques de prédiction issues notamment de la statistique et de l'intelligence artificielle.

Plusieurs recherches se sont orientées vers le processus d'extraction de connaissances à partir des données pour assurer des modèles capables de faire face à ces exigences. Notre approche s'inspire de deux approches existantes et améliore leur fiabilité en utilisant une méthode de prédiction et de classification adaptée aux connexions. En fait, non seulement on détecte les anomalies (approche par scénarios) mais aussi les profils normaux (approches comportementales). Par conséquent, au lieu d'utiliser une base de profils normaux d'utilisateur dans la phase d'apprentissage ou une base de signature d'attaques, notre approche utilise dans la phase d'apprentissage une base de données contenant à la fois des comportements normaux et des attaques et par la suite classifie le type de connexion grâce à un algorithme d'agrégation de classifieurs. De cette manière, le problème de la base de signature prédéfinie ainsi que le sur-apprentissage des profils normaux est résolu. Pour traiter le problème de fiabilité du système, nous avons opté pour le principe d'agrégation de classifieurs. En effet, ces dernières années, ce principe est utilisé en apprentissage automatique pour améliorer par des techniques de vote les performances d'un classifieur unique, ce qui a permis d'améliorer l'ajustement par une combinaison ou agrégation d'un grand nombre de modèles tout en évitant un sur-ajustement. Ce type d'agrégation se base sur :

- Des stratégies aléatoires (BAGGING [1]) : Le BAGGING (Bootstrap Aggregating) est une méthode qui combine des classifieurs pour obtenir un classifieur final (le classifieur agrégé). L'idée de cette méthode est de construire K classifieurs de manière indépendante, chacun étant construit sur un échantillon "perturbé" de l'échantillon initial S . La perturbation est réalisée via un bootstrap (tirage aléatoire avec remise). Chaque classifieur est ainsi construit sur un échantillon différent et la prédiction finale d'un exemple x s'obtient par la classe majoritairement prédite par les k classifieurs. Le BAGGING a pour principale action la réduction de la variance du classifieur simple.
- Des stratégies adaptatives (BOOSTING [15]) : L'idée initiale du BOOSTING est d'améliorer les compétences d'un classifieur faible (un classifieur légèrement meilleur que le hasard). L'idée originale de Schapire [15] a été affinée par Freund et Schapire [5] qui ont décrit l'algorithme original ADABOOST (Adaptative boosting) pour la prédiction d'une variable binaire. Le boosting adopte le même principe général que le BAGGING : construction d'une famille de modèles qui sont ensuite agrégés par une moyenne pondérée des estimations ou un vote. Il diffère nettement sur la façon de construire la famille qui est dans ce cas itérative : chaque modèle est une version adaptative du précédent en donnant plus de poids, lors de l'estimation suivante, aux observations mal prédites. Intuitivement, cet algorithme concentre donc ses efforts sur les observations les plus difficiles à ajuster tandis que l'agrégation de l'ensemble des modèles permet d'échapper au sur-ajustement.

3.1 Choix de stratégie d'agrégation pour les données réseaux

Pour choisir la stratégie la mieux adaptée à notre type de données (données réseaux qui sont généralement des données complexes et bruitées), nous avons procédé à une étude comparative entre les deux stratégies détaillées précédemment. Nous avons utilisé 100 itérations pour le BAGGING avec ID3 [13] comme classifieur individuel. AdaBoost a été utilisé avec ID3 [13] comme classifieur faible limité à un seul éclatement (decision stumps) et arrêté lorsque le taux d'erreur en apprentissage est devenu égal à 0. Les données utilisées pour cette étude comparative sont les mêmes données utilisées pour nos expériences (section 4.2) et qui sont décrites dans la section 4.1.

Classification / Réalité														Classes	Total	Précision	Rappel
a	b	c	d	e	f	g	h	i	j	k	l						
3887	0	1	3	0	0	0	0	0	0	0	0	0	a = normal	3891	0,90	0,99	
175	0	4	0	0	0	0	0	0	0	0	0	0	b = rare	179	0	0	
2	0	4286	0	0	0	0	0	0	0	0	0	0	c = neptune	4288	0,97	0,99	
9	0	0	11222	0	0	0	0	0	0	0	0	0	d = smurf	11231	0,99	0,99	
10	0	0	0	0	0	0	0	0	0	0	0	0	e = pod	10	0	0	
39	0	0	0	0	0	0	0	0	0	0	0	0	f = teardrop	39	0	0	
1	0	40	0	0	0	0	0	0	0	0	0	0	g = portsweep	41	0	0	
49	0	0	0	0	0	0	0	0	0	0	0	0	h = ipsweep	49	0	0	
88	0	0	0	0	0	0	0	0	0	0	0	0	i = back	88	0	0	
2	0	61	0	0	0	0	0	0	0	0	0	0	j = satan	63	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	k = nmap	9	0	0	
40	0	0	0	0	0	0	0	0	0	0	0	0	l = warezclient	40	0	0	
4311	0	4392	11225	0	0	0	0	0	0	0	0	0	-	19928	-	-	

TAB. 1: Matrice de Confusion d'Adaboost

Classification / Réalité																
a	b	c	d	e	f	g	h	i	j	k	l	Classes	Total	Précision	Rappel	
0	0	3881	10	0	0	0	0	0	0	0	0	a = normal	3891	0	0	
0	0	179	0	0	0	0	0	0	0	0	0	b = rare	179	0	0	
0	0	4288	0	0	0	0	0	0	0	0	0	c = neptune	4288	0,49	1	
0	0	0	11231	0	0	0	0	0	0	0	0	d = smurf	11231	0,99	1	
0	0	0	10	0	0	0	0	0	0	0	0	e = pod	10	0	0	
0	0	39	0	0	0	0	0	0	0	0	0	f = teardrop	39	0	0	
0	0	41	0	0	0	0	0	0	0	0	0	g = portsweep	41	0	0	
0	0	49	0	0	0	0	0	0	0	0	0	h = ipsweep	49	0	0	
0	0	88	0	0	0	0	0	0	0	0	0	i = back	88	0	0	
0	0	63	0	0	0	0	0	0	0	0	0	j = satan	63	0	0	
0	0	9	0	0	0	0	0	0	0	0	0	k = nmap	9	0	0	
0	0	40	0	0	0	0	0	0	0	0	0	l = warezclient	40	0	0	
0	0	8677	11251	0	0	0	0	0	0	0	0	-	19928	-	-	

TAB. 2: Matrice de Confusion du Bagging

Les résultats montrent la supériorité de la méthode adaptative (Boosting). En effet, le taux de bonne prédiction en généralisation estimé par une 10-validation croisée atteint 97%. La matrice de confusion (Tab1), obtenue dès la sixième itération, montre bien la performance en généralisation. En effet, le modèle construit parvient à différencier les connexions normales des attaques de type Neptune. Cette efficacité est due à la mise à jour adaptative de la distribution des exemples, visant à augmenter le poids de ceux mal appris par le classifieur précédent. De ce fait, à chaque itération les connexions mal interprétées sont re-classifiées. Enfin, nous avons remarqué que sur ces données, le boosting a convergé très vite avant même de terminer les 10 itérations (la convergence correspondant à un taux d'erreur nul en apprentissage). Ce modèle a été construit en 6 itérations. Alors que les résultats trouvés pour le BAGGING nous montrent que le taux d'erreur estimé en généralisation est de 22,12% donc un taux de bonne prédiction de 77,98%. De plus, la matrice de confusion de la validation croisée (Tab2) met en évidence les erreurs commises par le BAGGING. En fait, la plupart des connexions sont classées comme attaque Neptune et ceci est dû à la forte présence de l'attaque Neptune dans les données d'apprentissage. En effet, les enregistrements, ayant la modalité de classe Neptune, représentent 53% des connexions. Cette erreur de classification est due essentiellement à la méthode aléatoire de construction des nouveaux échantillons à partir des données initiales du BAGGING où l'attaque Neptune est fortement représentée. Cette sur-représentation implique que les échantillons bootstrap contiennent une très forte proportion d'individus catégorisés "Neptune" et ne permet pas aux différents classifieurs d'être assez pertinents individuellement.

3.2 Une approche hybride basée sur un *Adaboost adapté*

Les résultats détaillés dans la précédente section nous amènent à choisir le principe du boosting pour la phase d'apprentissage et nous avons cherché à adapter l'algorithme à nos données. En effet, pour améliorer les performances d'*Adaboost* [5] et éviter de le forcer à apprendre des exemples a priori bruités ou des exemples qui deviendraient trop difficiles à apprendre durant le processus du *boosting* telles que les données réseaux, nous proposons une nouvelle approche qui s'inspire du fait qu'*Adaboost* construit, à chaque itération, des hypothèses sur un échantillon bien défini. La mise à jour et le calcul de l'erreur d'apprentissage sont faits à partir des résultats de ces seules hypothèses et n'exploitent pas les résultats fournis par les hypothèses construites aux itérations antérieures sur d'autres échantillons. Cette

approche est appelée "Adaboost adapté" puisque la mise à jour des exemples à l'itération courante prend en compte la nature des données réseaux qui sont complexes et déséquilibrées. En fait, à chaque itération, la mise à jour des données réseaux prend en compte non seulement les résultats de l'itération courante mais aussi ceux des itérations antérieures, Ainsi l'analyseur, à chaque itération, apprend à partir des hypothèses antérieures de profils normaux et d'attaques déjà prédits. Cette mise à jour à chaque itération enrichit la mémoire de l'analyseur.

Pseudo code de Adaboost Adapté

Soit X_0 à prévoir et $S = (x_1, y_1), \dots, (x_n, y_n)$ un ensemble de connexions avec les X_i les attributs et les y_i les types de connexions

- Pour $i = 1, 2 \dots n$, faire
- Initialiser les poids $p_0(x_i) = 1/n$; (Au début, on affecter le même poids aux exemples)
- Fin pour
- $t \leftarrow 0$
- Tant que $t \leq T$ faire (T est le nombre d'itération pour le Boosting)
- Tirer un ensemble de connexions S_t dans S selon les probabilités p_t . (C'est le principe de l'échantillonnage bootstrap qui se base sur les poids des exemples)
- Construire une hypothèse h_t sur S_t par un algorithme d'apprentissage A. (Construction d'un modèle de prédiction)
- Soit ϵ_t l'erreur apparente de h_t sur S avec $\epsilon_t = \sum$ poids des connexions mal classifiées.
- Calculer $\alpha_t = 1/2 \ln((1 - \epsilon_t)/\epsilon_t)$ (Chaque modèle construit est affecté par un poids qui correspond à α_t).
- Pour $i=1, m$ faire (m est le nombre d'exemples existants dans l'échantillon S_t)
- $P_{t+1}(x_i) \leftarrow (p_t(x_i)/Z_t)e^{-\alpha_t}$ si $\text{argmax}(\sum_{i=1}^t \alpha_i h_i(x_i)) = y_i$ (si bien classée aux itérations antérieures)
- $P_{t+1}(x_i) \leftarrow (p_t(x_i)/Z_t)e^{+\alpha_t}$ si $\text{argmax}(\sum_{i=1}^t \alpha_i h_i(x_i)) \neq y_i$ (si mal classée aux itérations antérieures)
- (Z_t est une valeur de normalisation telle que $\sum_{i=1}^n p_t(x_i) = 1$)
- Fin Pour
- $t \leftarrow t + 1$
- Fin tant que
- Fournir en sortie l'hypothèse finale :
- $H(x) = \text{argmax } y \in Y \sum_{t=1}^T \alpha_t$ (La classe prédite correspond à la classe majoritaire du vote)

4 Expérience et résultats

4.1 Présentation des données

Les données utilisées pour nos expériences est un échantillon des données standards qui ont été préparées et contrôlées par le laboratoires MIT Lincoln pour le programme d'évaluation de détection d'intrusion DARPA 1998. Ces données sont aussi utilisées pour le concours de détection d'intrusion de KDD 1999 [17]. Chaque connexion est marquée en tant que normale, ou comme attaque, avec exactement un type spécifique d'attaque. Les attaques trouvées sont classées selon quatre catégories principales : DOS (Deni de service), R2L (Accès non autorisé d'une machine à distance, par exemple devinant le mot de passe), U2R (Accès non autorisé aux

privilèges d'un super-utilisateur tel que buffer overflow) et PROBE (Sondage et surveillance tel que port scanning).

Les différentes attaques existantes classées en catégories sont décrites dans le tableau 3.

Types d'attaques	Catégories d'attaques
back, land, pod, smurf, teardrop	dos
buffer_overflow, loadmodule, perl, rootkit	u2r
ftp_write, guess_passwd, imap, multihop, phf, spy, warezmaster	r2l
ipsweep, nmap, portsweep, satan	probe

TAB. 3: Types et catégories d'attaques existantes

Nom d'attributs	Description	type
Attributs du trafic de connexion pendant une fenêtre de deux secondes		
Count	nombre de connexion à la même machine que la connexion courante durant les dernières deux secondes	continu
Attributs des connexions de même-machine		
Error_rate	nombre de connexions qui ont des erreurs de SYN	continu
Error_rate	nombre de connexions qui ont des erreurs de REJ	Continu
Same_srv_rate	nombre de connexions au même service	Continu
diff_srv_rate	nombre de connexions au service différents	Continu
srv_count	nombre de connexions au même service que le raccordement courant dans les dernières deux secondes	continu
Attributs des connexions de même-service		
srv_error_rate	nombre de connexions qui ont des erreurs de SYN	continu
srv_error_rate	nombre de connexions qui ont des erreurs de REJ	Continu
srv_diff_host_rate	nombre de connexions aux différents hosts machines	Continu
Attributs des connexions TCP individuelles		
durée	longueur (nombre de secondes) de la connexion	continu
protocol_type	type du protocole, par exemple tcp, UDP ...	discret
service	service de réseau pour la destination, par exemple, HTTP, telnet, etc...	discret
src_bytes	nombre de bytes de données de la source à la destination	continu
dst_bytes	nombre de bytes de données de la destination à la source	continu
flag	statut normal ou erreur de la connexion	discret
land	1 si la connexion est from/to même host/port; 0 autrement	discret
wrong_fragment	nombre de " faux " fragments	continu
urgent	nombre de paquets urgent	continu
Attributs de contenu		
hot	Nombre de " hot " indicateurs	continu
num_failed_logins	nombre de tentatives d'ouverture échouées	continu
logged_in	1 si entré avec succès; 0 autrement	discret
num_compromised	le nombre de conditions compromises	continu
root_shell	1 si root shell est obtenue; 0 autrement	discret
su_attempted	1 si la commande su root racine a été essayée; 0 autrement	discret
num_root	nombre d'accès root	continu
num_file_creations	nombre d'opérations de création de dossier	continu
num_shells	Nombre de shell sollicités	continu
num_access_files	nombre d'opérations sur des dossiers de contrôle d'accès	continu
num_outbound_cmds	nombre de commandes venantes d'une session de ftp	continu
is_hot_login	1 si l'ouverture appartient à la hot liste; 0 autrement	discret
is_guest_login	1 si l'ouverture est un guest login; 0 autrement	discret

TAB. 4: Attributs des données réseaux

Pour une meilleure prédiction, nous devons faire une sélection des attributs les plus pertinents des connexions. On a pu identifier plusieurs attributs que nous avons catégorisé selon quatre types. Les attributs d'une même machine décrivent seulement les connexions faites durant les deux dernières secondes et ayant le même destinataire que la connexion courante. Les attributs de même service décrivent seulement les connexions faites durant les deux dernières secondes et ayant le même service que la connexion courante. Les attributs d'une même machine et même service définissent les critères du trafic de connexion faits en une fenêtre de deux secondes. On trouve aussi des attributs de connexions TCP individuelles. Finalement, il existe des attributs qui indiquent le comportement anormale dans les données, telles que

le nombre de tentatives échouées d'ouverture. Ces attributs sont les attributs de contenu. Les variables explicatives des données sont décrites dans le tableau 4.

4.2 Résultats

Les résultats donnés par cette variante adaptée aux données réseaux sont surprenants (Tab 5). En effet, cette variante construit un modèle de prédiction pour la détection d'intrusions avec un taux d'erreur de 0.035% en entraînement et de 0.14% en généralisation. En plus de l'amélioration du taux d'erreur, notre AdaBoost adapté améliore très significativement le rappel et la précision des classes minoritaires par rapport à AdaBoost et au Bagging. Ces deux dernières méthodes d'agrégation sont fortement sensibles au déséquilibre des classes et tendent à prédire les différents types d'attaques comme l'une des trois modalités les plus représentées. La matrice de confusion obtenue au bout de 10 itérations de la procédure montre bien la bonne performance de classification de ce modèle. Sur 3891 connexions normales, 3 seulement sont vues comme des attaques et sur 179 attaques rares seulement 4 sont considérées normales. Les autres connexions sont bien classées.

Classification / Réalité																
a	b	c	d	e	f	g	h	i	j	k	l	Classes	Total	Précision	Rappel	
3884	7	0	0	0	0	0	0	0	0	0	0	a = normal	3891	0,99	0,99	
8	167	1	0	0	0	0	0	0	1	0	2	b = rare	179	0,94	0,93	
0	1	4287	0	0	0	0	0	0	0	0	0	c = neptune	4288	0,99	0,99	
0	0	0	11231	0	0	0	0	0	0	0	0	d = smurf	11231	1	1	
0	0	0	0	10	0	0	0	0	0	0	0	e = pod	10	0,90	1	
0	0	0	0	1	38	0	0	0	0	0	0	f = teardrop	39	1	0,97	
0	0	0	0	0	0	40	0	0	1	0	0	g = portsweep	40	0,93	0,97	
0	0	0	0	0	0	1	48	0	0	0	0	h = ipsweep	49	1	0,97	
0	0	0	0	0	0	0	0	88	0	0	0	i = back	88	1	1	
1	1	0	0	0	0	2	0	0	59	0	0	j = satan	63	0,96	0,93	
1	1	0	0	0	0	0	0	0	0	7	0	k = nmap	9	1	0,77	
0	0	0	0	0	0	0	0	0	0	0	40	l = warezclient	40	0,95	1	
3894	177	4288	11231	11	38	43	48	88	61	7	42	-	19928	-	-	

TAB. 5: Matrice de Confusion du AdaBoost adapté

5 Conclusion et perspectives

Afin d'améliorer la détection des attaques et par conséquent diminuer le taux d'erreur en classification, nous avons proposé dans cet article une nouvelle approche fondée sur le principe d'agrégation de classificateurs. Le but de cette démarche est de perfectionner la performance d'un classificateur par une méthode de vote et agir principalement sur la réduction des alertes de type "faux positifs". En effet, notre méthode classe correctement les connexions normales et les attaques, grâce à une version d'AdaBoost adapté aux données réseaux. Nous constatons après des tests que cette approche permet d'améliorer le taux de détection d'attaques ainsi que les alertes de type "faux positif". Actuellement, nous sommes entrain de comparer notre approche avec des nouvelles approches qui se basent sur le bayes naïf [4]. Nous envisageons de tester notre approche avec des données plus récentes afin de valider cette proposition sur des nouvelles intrusions. Une nouvelle perspective consiste à utiliser la classification à base de règles, à savoir les arbres de décision ou la classification associative. Ce type de classification

est intéressant et assez compréhensible par un utilisateur grâce aux résultats basés sur des règles qui sont plus facilement interprétable. Nous comptons également étudier et utiliser les techniques de fouille de données pour prévenir en plus les nouvelles attaques non identifiées. Enfin, nous envisageons d'étudier de façon plus théorique notre approche de boosting hybride et de la valider sur des données d'autres domaines.

Références

- [1] L. Breiman. Bagging predictors. *Machine Learning*, 26 :123–140, 1996.
- [2] James Cannady. Artificial neural networks for misuse detection. In *NATIONAL INFORMATION SYSTEMS SECURITY CONFERENCE*, pages 443–456, 1998.
- [3] Dacier Marc et Wespi Andreas. Debar Hervé. A revised taxonomy for intrusion detection systems. *Annales des Télécommunications*, 55 :7–8, 2000.
- [4] Dewan Md. Farid and Mohammad Zahidur Rahman. Anomaly network intrusion detection based on improved self adaptive bayesian algorithm. *Journal of Computers*, 5(1), January 2010. Academy Publisher.
- [5] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Machine Learning : Proceedings of the Thirteenth National Conference*, pages 148–156, 1996.
- [6] Rebecca Gurley. Intrusion detection. *MacMillan Technical Publishing*, 2000.
- [7] M. Becker H. Debar and D. Siboni. A neural network component for an intrusion detection system. *Proceedings of the IEEE Symposium of Research in Computer Security and Privacy*, 1992.
- [8] Steven A. Hofmeyr. The implications of immunology for secure systems design. *Computers & Security*, 23(6) :453–455, 2004.
- [9] Anderson (J.P.). Computer security threat monitoring and surveillance. *James P. Anderson Company, Fort Washington, Pennsylvania*, 1980.
- [10] Sandeep Kumar and Eugene H. Spafford. Ia pattern matching model for misuse intrusion detection. *Proceedings of the 17th National Computer Security Conference*, pages 11–21, 1994.
- [11] Terran Lane and Carla E. Brodley. Temporal sequence learning and data reduction for anomaly detection. *Proceedings of the Fifth ACM Conference on Computer and Communications Security*, pages 150–158, 1998.
- [12] Ludovic Mé, Zakia Marrakchi, Cédric Michel, Hervé Debar, and Frédéric Cuppens. La détection d'intrusions : les outils doivent coopérer. In *Revue de l'Electricité et de l'Electronique*, pages No 5 50–55, 2001.
- [13] J. R. Quinlan. Comparing connectionist and symbolic learning methods. *MIT Press*, 1 : Constraints and prospects(15) :445–456, 1994.
- [14] Jake Ryan, Meng-Jang Lin, and Risto Miikkulainen. Intrusion detection with neural networks. advances in neural information processing systems. *The MIT Press*, 1998.
- [15] R. Shapire. The strength of weak learnability. *Machine Learning*, 5 :197–227, 1990.

- [16] Taeshik Shon, Jung-Taek Seo, and Jongsub Moon. Svm approach with a genetic algorithm for network intrusion detection. In *ISCIS*, pages 224–233, 2005.
- [17] Salvatore J. Stolfo, Wei Fan, Wenke Lee, Andreas Prodromidis, and Philip K. Chan. Cost-based modeling and evaluation for data mining with application to fraud and intrusion detection, 1999.
- [18] Yu.Y and Huang Hao. An ensemble approach to intrusion detection based on improved multi-objective genetic algorithm. *Journal of Software*, 18(6) :1369–1378, June 2007.

Summary

Currently, research in data mining focused on the field of intrusions detection in the computing systems. In effect, the detection of the anomalies in the data-processing networks is regarded as one problem of data classification where the use of the data mining techniques. Several approaches of machine learning are used on complex masses of data and dynamic in order to build an effective intrusion detection system (IDS). However, these approaches are confronted with problems of precision and especially the increase of false alarms "the false-positives" when they use complex and unbalanced data with a high speed traffic networks. In order to improve detection of these attacks and consequently to decrease the error rate in classification, we propose in this paper a new approach based on the aggregation of classifiers having for goal to improve the performance of a classifier by a method of vote. In fact, our approach is hybrid, since that it classifies at the same time normal connections and attacks, using a version of boosting adapted to the networks data "screens". This new approach that we proposed reduce mainly the false-positives.

Nouvelle représentation concise exacte des motifs corrélés fréquents basée sur une exploration simultanée des espaces de recherche conjonctif et disjonctif

Nassima Ben Younes, Tarek Hamrouni, Sadok Ben Yahia

Département des Sciences de l'Informatique, Faculté des Sciences de Tunis, Tunisie
{tarek.hamrouni,sadok.benyahia}@fst.rnu.tn

Résumé. La fouille de données est un processus d'extraction de connaissances valides et exploitables à partir de grands volumes de données. Les connaissances peuvent être sous forme de règles d'association qui représentent des liens entre attributs. Dans la pratique, ces derniers sont très nombreux, chose qui fait que seules les données avec une fréquence d'apparition importante sont généralement traitées. Toutefois, les motifs extraits dans pareilles situations - dits fréquents - n'offrent pas assez d'informations sur les corrélations qui régissent les items, *c.-à.-d.* le degré de dépendance des items entre eux. Dans ce papier, nous cherchons à extraire une représentation concise des motifs corrélés fréquents associés à la mesure de corrélation *bond*, tout en définissant l'opérateur de fermeture correspondant. Cet ensemble réduit offre non seulement la possibilité de retrouver tous les motifs corrélés fréquents sans perte d'information mais aussi la possibilité de dériver les supports conjonctif, disjonctif et négatif de chaque motif d'une manière efficace.

1 Introduction et motivations

L'extraction des motifs fréquents à partir d'un contexte d'extraction (appelé, en classification, *tableau d'incidence des données*) est une étape importante dans l'extraction des connaissances valides en fouille de données. Depuis son introduction, ce problème a sollicité l'intérêt de plusieurs chercheurs vu que les motifs fréquents constituent une source d'informations permettant de comprendre les relations entre les items. Vu que le nombre des motifs fréquents extraits à partir d'une base de transactions qui présente des données fortement corrélées est très élevé, plusieurs *représentations concises exactes des motifs fréquents* sont apparues. Les représentations proposées dans la littérature sont adéquates à différentes mesures. Outre le support conjonctif, des représentations récentes tiennent également compte d'une autre mesure, nommée *support disjonctif*, afin de réduire le nombre de motifs les constituant. Néanmoins, en considérant une valeur faible de seuil minimal de support *minsupp*, la taille de ces représentations demeure volumineuse et plusieurs motifs fréquents liant des items faiblement corrélés sont extraits. Ainsi, une valeur de *minsupp* élevée peut résoudre ce problème, mais beaucoup de motifs intéressants seront élagués. Pour réduire la taille des représentations et améliorer

la qualité des motifs extraits, plusieurs mesures de corrélation ont été proposées dans la littérature. En adoptant ces mesures, nous pouvons extraire des motifs corrélés fréquents qui peuvent être appliqués dans différents domaines d'application. Parmi les applications les plus courantes, nous prenons comme exemple d'illustration de la corrélation, une application médicale où nous avons trois symptômes X , Y et Z ainsi que n patients dont seul un petit nombre, ρ , présente au moins l'un de ces symptômes. La maladie qui se présente avec les symptômes X , Y et Z ne sera pas prise en compte car ayant un support largement inférieur à *minsupp*. Toutefois, il pourrait être intéressant dans ce cadre applicatif de la retenir car il pourrait s'agir d'une maladie rare puisque sa corrélation sera très élevée. Le support conjonctif entre X , Y et Z sera égal à ρ alors que son support conjonctif rapporté par son support disjonctif, qui n'est autre que sa mesure *bond*, sera très proche de 1 et dépasse le seuil minimal de *bond* noté *minbond*.

Dans ce papier, nous nous intéressons en détail à la mesure *bond*, adaptée au contexte des motifs fréquents et règles d'association dans (Omicinski, 2003). Cette dernière est très intéressante puisqu'elle lie le support conjonctif d'un motif indiquant la fréquence de co-occurrence de ses items à son support disjonctif indiquant la complémentarité d'occurrence des items. Toutefois, peu d'études ont été dédiées à cette mesure. Une des principales raisons de cette négligence est que l'extraction des motifs vérifiant cette mesure est plus difficile que celle des motifs vérifiant *all-confidence* (Lee et al., 2003). Nous proposons alors une nouvelle représentation concise des motifs corrélés fréquents basée sur la mesure *bond*. Cette représentation concise offre la possibilité de dériver exactement les supports conjonctif, disjonctif et négatif de tout motif corrélé fréquent. À notre connaissance, cette représentation est l'unique proposée dans la littérature associée à la mesure *bond*.

Le reste de l'article est organisé comme suit. Dans la section 2, nous présentons les notions de base utilisées. Dans la section 3, nous présentons brièvement les travaux de la littérature évoquant les mesures de corrélation et les représentations concises existantes. Notre proposition est détaillée dans la section 4. Dans la section 5, nous menons une étude expérimentale qui prouve l'utilité de la représentation proposée, puis nous terminons par une conclusion.

2 Fondements mathématiques

Dans cette section, nous présentons l'ensemble des notions de base qui seront utiles dans le reste de cet article.

Définition 1 - Contexte d'extraction - Un contexte d'extraction est un triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ dans lequel \mathcal{O} et \mathcal{I} sont, respectivement, des ensembles finis d'objets et d'items, et $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ est une relation binaire entre les objets et les items. Un couple $(o, i) \in \mathcal{R}$ dénote le fait que l'objet $o \in \mathcal{O}$ contient l'item $i \in \mathcal{I}$.

Un exemple de contexte d'extraction est celui donné par le tableau 1. Dans ce contexte, nous avons $\mathcal{O} = \{1, 2, 3, 4, 5, 6\}$ et $\mathcal{I} = \{A, B, C, D, E, F\}$; les objets peuvent représenter un ensemble de patients et les items représentent des symptômes que peuvent présenter ces malades. Afin d'évaluer l'intérêt d'un motif, plusieurs mesures sont utilisées dont les plus connues sont présentées à travers la définition suivante.

	A	B	C	D	E	F
1	×	×				
2	×		×	×		
3			×	×	×	
4				×	×	×
5	×	×	×	×	×	
6	×	×	×			

TAB. 1 – Exemple de contexte d'extraction.

Définition 2 (Hamrouni *et al.*, 2009) - **Supports d'un motif** - Soient $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction et un motif $I \subseteq \mathcal{I}$. Nous distinguons trois types de supports correspondants à I :

$$\begin{aligned} \text{Supp}(\wedge I) &= |\{o \in \mathcal{O} \mid (\forall i \in I, (o, i) \in \mathcal{R})\}| \\ \text{Supp}(\vee I) &= |\{o \in \mathcal{O} \mid (\exists i \in I, (o, i) \in \mathcal{R})\}| \\ \text{Supp}(\neg I) &= |\{o \in \mathcal{O} \mid (\forall i \in I, (o, i) \notin \mathcal{R})\}| \end{aligned}$$

$\text{Supp}(\wedge I)$ (resp. $\text{Supp}(\vee I)$ et $\text{Supp}(\neg I)$) est appelé support conjonctif (resp. disjonctif et négatif) de I .

Exemple 1 Considérons le contexte illustré par le tableau 1. Nous avons $\text{Supp}(\wedge (BC)) = |\{5, 6\}| = 2$, $\text{Supp}(\vee (BC)) = |\{1, 2, 3, 5, 6\}| = 5$ et $\text{Supp}(\neg(BC)) = |\{4\}| = 1$ ⁽¹⁾.

Nous définissons maintenant les différents opérateurs de fermeture associés à ces mesures.

Définition 3 (Pasquier *et al.*, 2005) - **Fermeture conjonctive d'un motif** - La fermeture conjonctive d'un motif $I \subseteq \mathcal{I}$, notée $f_c(I)$, est égale à $\max_{\subseteq} \{I' \subseteq \mathcal{I} \mid (I \subseteq I') \text{ et } (\text{Supp}(\wedge I') = \text{Supp}(\wedge I))\}$.

Exemple 2 Soit le contexte illustré par le tableau 1. $\{A, B\}$ est l'ensemble maximal d'items communs aux objets $\{1, 5, 6\}$, $f_c(AB) = AB$ alors AB est un motif fermé conjonctif.

Définition 4 (Bastide *et al.*, 2000) - **Générateur minimal** - Un motif $g \subseteq \mathcal{I}$ est dit générateur minimal d'un motif fermé f , si et seulement si $f_c(g) = f$ et il n'existe aucun motif $g_1 \subset g$ tel que $f_c(g_1) = f$.

Exemple 3 Soit le contexte illustré par le tableau 1. $\{B\}$ est l'ensemble minimal d'items communs aux objets $\{1, 5, 6\}$, et $f_c(B) = AB$ alors B est un générateur minimal de AB .

En comparant les espaces conjonctif et disjonctif, nous remarquons que leurs structures sont très similaires. Ainsi, définissons les notions de base relatives à l'espace disjonctif.

Définition 5 (Hamrouni *et al.*, 2009) - **Fermeture disjonctive d'un motif** - La fermeture disjonctive d'un motif $I \subseteq \mathcal{I}$, notée $f_d(I)$, est égale à $\max_{\subseteq} \{I' \subseteq \mathcal{I} \mid (I \subseteq I') \text{ et } (\text{Supp}(\vee I') = \text{Supp}(\vee I))\}$.

Définition 6 (Casali *et al.*, 2005) - **Motif essentiel** - Soient $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction et un motif $I \subseteq \mathcal{I}$. I est un motif essentiel si et seulement si $\text{Supp}(\vee I) > \max\{\text{Supp}(\vee I \setminus \{i\}) \mid i \in I\}$.

¹Nous employons une forme sans séparateur pour les ensembles : par exemple, BC représente l'ensemble $\{B, C\}$.

Exemple 4 Soit le contexte défini dans le tableau 1, nous remarquons que le motif AB n'est pas essentiel car $Supp(\vee(AB)) = Supp(\vee A) = 4$.

Le support conjonctif est une mesure efficace pour l'extraction des motifs fréquents. Toutefois, ces motifs ne donnent aucune information sur la corrélation inter-transactions des items. Ainsi, nous étudions la mesure de corrélation *bond* qui permet d'éviter les lacunes des mesures traditionnelles.

La contrainte définie dans la définition 7 est très intéressante dans l'évaluation des mesures, vu qu'elle permet, si la mesure la vérifie, d'extraire d'une manière efficace l'ensemble des motifs vérifiant la mesure.

Définition 7 (Bonchi et Lucchese, 2006) - **Contrainte anti-monotone** - Soit $I \subseteq \mathcal{I}$. Une contrainte Q est dite anti-monotone si $\forall I_1 \subseteq I : I \text{ vérifie } Q \Rightarrow I_1 \text{ vérifie } Q$.

La propriété définie dans la définition 8 est très utile, vu qu'elle permet, si une mesure la vérifie, d'extraire d'une manière efficace l'ensemble des motifs vérifiant la mesure.

Définition 8 (Xiong et al., 2006) **Propriété de cross-support** Soient $\mathcal{I} = \{i_1, i_2, \dots, i_m\}$ un ensemble d'items et t un seuil minimal fixé. Un motif $X \subseteq \mathcal{I}$ vérifie la propriété de cross-support s'il contient au moins deux items x et y tel que $\frac{Supp(\wedge x)}{Supp(\wedge y)} < t$ où $0 < t < 1$.

3 Revue critique des travaux de l'état de l'art

Plusieurs travaux dans la littérature se sont focalisés sur l'extraction des motifs fréquents. Néanmoins, le support conjonctif utilisé pour évaluer la fréquence ne suffit pas pour donner l'information concernant la complémentarité d'apparition des items d'un motif X car l'apparition de tous les items de X ensemble est une contrainte forte mais elle ne donne pas d'information précise sur la corrélation entre les items. De ce fait, quelques mesures de corrélation qui permettent d'extraire des motifs corrélés fréquents ont été proposées dans la littérature dont les plus connues sont *lift* et χ^2 dans (Brin et al., 1997), *any-confidence*, *all-confidence* et *bond* dans (Omicinski, 2003), etc. Ces mesures permettent de donner une information plus complète sur la corrélation entre les items, mais seules *all-confidence* et *bond* sont caractérisées par le fait qu'elles sont à la fois anti-monotones et indépendantes du nombre de transactions. La contrainte d'anti-monotonie est fortement souhaitée, ainsi plusieurs travaux de la littérature s'y sont intéressés. Par exemple, dans (Le Bras et al., 2010), les auteurs généralisent la propriété UEUC⁽²⁾, appliquée initialement à la mesure de confiance, en établissant une condition nécessaire et suffisante pour qu'une mesure vérifie la contrainte d'anti-monotonie. Cette propriété offre une stratégie d'élagage très efficace dans la réduction de la quantité des données traitées. En outre, depuis son introduction, le problème d'extraction des motifs corrélés fréquents a sollicité l'intérêt de plusieurs chercheurs. À cet effet, plusieurs algorithmes d'extraction (Lee et al., 2003), (Omicinski, 2003), (Xiong et al., 2006), ayant adopté différentes mesures de corrélation, ont été proposés dans la littérature. Toutefois, ces derniers génèrent toujours un nombre élevé de motifs corrélés fréquents, dont la majorité est redondante. À notre connaissance, seul le travail de (Kim et al., 2004) permet d'extraire une représentation concise des

²UEUC est l'acronyme de Universal Existential Upward Closure.

motifs corrélés fréquents basée sur la mesure *all-confidence*. De plus, les études proposées s'appuient uniquement sur l'exploration de l'espace conjonctif pour l'extraction des motifs corrélés et aucune ne s'est intéressée à l'exploration de l'espace disjonctif afin d'extraire les motifs corrélés suivant la mesure *bond*. D'autre part, une étude générale sur les représentations condensées adéquates aux fonctions conservées est effectuée (Soulet et Crémilleux, 2008). Elle s'est focalisée principalement sur les représentations condensées des motifs fréquents, mais aucune représentation condensée exacte adéquate à une mesure de corrélation n'est étudiée. Dans cet article, nous allons proposer une nouvelle représentation concise exacte associée à la mesure *bond*, *c.-à.-d.*, un ensemble dont la cardinalité est inférieure à celle de l'ensemble total des motifs corrélés fréquents par rapport à *bond*. Cette mesure est très intéressante puisqu'elle lie le support conjonctif d'un motif à son support disjonctif. Donc, la représentation que nous proposons associée à *bond* est basée sur l'exploration simultanée des deux espaces de recherche conjonctif et disjonctif. C'est un compromis entre la représentation par les motifs fermés fréquents (Pasquier et al., 2005) et la représentation concise exacte basée sur les motifs fermés disjonctifs (Hamrouni et al., 2009). Elle offre les principaux avantages de ces dernières, à savoir la dérivation directe des supports conjonctif et disjonctif d'un motif corrélé fréquent. Ainsi, contrairement aux deux autres représentations, la représentation associée à la mesure *bond* et basée sur les motifs fermés corrélés fréquents permet de trouver la corrélation entre les items sans avoir recours aux identités d'inclusion-exclusion (Galambos et Simonelli, 2000).

Vu que la proposition d'une représentation concise a pour but de réduire la taille de l'ensemble extrait de motifs tout en étant capable de régénérer d'une manière efficace l'ensemble total des motifs corrélés, nous introduisons une nouvelle représentation associée à la mesure de corrélation *bond* qui permet d'atteindre ces buts surtout pour les bases denses.

4 Nouvelle représentation concise exacte des motifs corrélés fréquents

Pour réduire le nombre des motifs fréquents à forte corrélation et en utilisant des mesures de corrélation, des représentations concises des motifs corrélés sont apparues. Cette recherche de la corrélation entre les items est due aux lacunes des motifs fréquents qui n'offrent pas d'information concernant le degré de dépendance de l'apparition d'un item dans la base de transactions vis-à-vis des autres items. Ceci nous a motivé à offrir à l'utilisateur cette information clé moyennant la recherche de la corrélation entre les items en plus de leur fréquence d'apparition simultanée.

4.1 Opérateur de fermeture associée à la mesure *bond*

Nous étudions dans cette sous-section les propriétés structurelles de la mesure *bond* qui offrent différents avantages dont le principal est la réduction du nombre des motifs corrélés fréquents. Dans les travaux de la littérature, d'autres mesures similaires à la mesure *bond* ont été proposées dans différents contextes d'application telles que les mesures *Coherence* dans (Lee et al., 2003), *coefficient de Tanimoto* dans (Tanimoto, 1958) et *Jaccard* dans (Jaccard, 1908). Par ailleurs, le dénominateur de la formule associée à cette mesure a toujours été considéré comme étant le cardinal de *l'univers* de X , *c.-à.-d.* l'ensemble des transactions où un item

de X apparaît. À cet égard, aucun de ces travaux n'a fait le lien entre cet univers et le support disjonctif de X . En effet, le cardinal de l'univers de X n'est autre que son support disjonctif. Nous proposons donc une nouvelle définition de la mesure *bond* exprimée comme suit :

Définition 9 - Redéfinition de la mesure bond - La mesure *bond* d'un motif X est définie par la règle suivante :

$$bond(X) = \frac{Supp(\wedge X)}{Supp(\vee X)}$$

La mesure *bond* prend ses valeurs sur l'intervalle $[0, 1]$. La valeur de la mesure *bond* d'un motif ne peut être nulle que si son support conjonctif est nul. La mesure *bond* est égale à 1 si les supports conjonctif et disjonctif d'un motif sont égaux. Cette condition est vérifiée par tous les items, ainsi que tout motif dont les items associés sont mutuellement dépendants. Cette mesure admet les propriétés intéressantes présentées par les propositions suivantes dont les preuves des trois premières sont omises faute d'espace.

Proposition 1 La mesure *bond* est une mesure indépendante du nombre de transactions, elle est donc dite descriptive.

Proposition 2 La mesure *bond* est une mesure symétrique tel que $bond(XY) = bond(YX)$.

Proposition 3 La mesure *bond* induit une contrainte anti-monotone.

Exemple 5 Soient un motif ABC et *minbond* un seuil minimal de *bond* tel que $bond(ABC) \geq minbond$. $\forall X \subseteq ABC$, $bond(X) \geq bond(ABC)$ d'où $bond(X) \geq minbond$.

Proposition 4 Tout motif $X \subseteq \mathcal{I}$ qui vérifie la propriété de *cross-support* par rapport à un seuil minimal t vérifie la propriété suivante : $bond(X) < t$. Ainsi, il suffit de trouver deux items x et $y \in \mathcal{I}$ vérifiant la propriété de *cross-support* (c.-à-d. $\frac{Supp(\wedge x)}{Supp(\wedge y)} < t$) pour déduire que tout motif contenant les items x et y n'est pas corrélé du moment où sa mesure *bond* sera strictement inférieure au seuil minimal t .

Preuve. Soient $X \subseteq \mathcal{I}$ et un seuil minimal t avec $0 < t < 1$. X vérifie la propriété de *cross-support* par rapport à t , alors $\exists x$ et $y \in X$ tel que $\frac{Supp(\wedge x)}{Supp(\wedge y)} < t$. Démontrons que $bond(X) < t$: $\frac{Supp(\wedge X)}{Supp(\vee X)} \leq \frac{Supp(\wedge (xy))}{Supp(\vee (xy))} \leq \frac{Supp(\wedge (xy))}{Supp(\vee y)} \leq \frac{Supp(\wedge x)}{Supp(\vee y)} = \frac{Supp(\wedge x)}{Supp(\wedge y)} < t$. \diamond

Exemple 6 Soient le motif $X = DF$ et $t = 0,3$. Nous avons $\frac{Supp(\wedge F)}{Supp(\wedge D)} = 0,25 < t$. Nous déduisons ainsi que tout motif contenant les items D et F n'est pas corrélé du moment où sa mesure *bond* sera strictement inférieure au seuil minimal t . Ainsi, $bond(DF) < t$ et $bond(DF) \geq bond(DF \cup \{i\})$, $\forall i \in \{A, B, C, E\}$, d'où $bond(DF \cup \{i\}) < t$.

La mesure de corrélation *bond* est le rapport entre les supports conjonctif et disjonctif d'un motif. Ainsi, si deux motifs X et Y tels que $X \subseteq Y$ (ou $Y \subseteq X$) ont la même valeur de la mesure de corrélation *bond*, ils ont donc les mêmes supports conjonctifs et disjonctifs. En effet, pour que $\frac{a}{b} = \frac{c}{d}$ il faut que $a = c$ et $b = d$ ou $a > c$ et $b > d$ ou $a < c$ et $b < d$. Par ailleurs, en ajoutant un item i à un motif X , les supports conjonctifs et disjonctifs varient d'une manière inversement proportionnelle telle que $\forall i \in \mathcal{I}$, $Supp(\wedge X) \geq Supp(\wedge (X \cup \{i\}))$ et $Supp(\vee X) \leq Supp(\vee (X \cup \{i\}))$. Ainsi, le seul cas pour que $\frac{Supp(\wedge X)}{Supp(\vee X)} = \frac{Supp(\wedge (X \cup \{i\})}{Supp(\vee (X \cup \{i\}))}$ est le cas où $Supp(\wedge X) = Supp(\wedge (X \cup \{i\}))$ et $Supp(\vee X) = Supp(\vee (X \cup \{i\}))$. Nous avons ainsi la proposition suivante :

Proposition 5 Soit $I \subseteq \mathcal{I}$. $\forall i \in \mathcal{I}$, $\text{bond}(I \cup \{i\}) = \text{bond}(I)$ si et seulement si $\text{Supp}(\wedge (I \cup \{i\})) = \text{Supp}(\wedge I)$ et $\text{Supp}(\vee (I \cup \{i\})) = \text{Supp}(\vee I)$.

D'après ce que nous avons présenté, il en découle la définition suivante :

Définition 10 - Fermeture f_{bond} - Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. f_c et f_d sont respectivement l'opérateur de fermeture conjonctive et celui de fermeture disjonctive. La fermeture associée à la mesure bond résulte de l'intersection entre la fermeture disjonctive et la fermeture conjonctive. Nous définissons cette fermeture comme suit :

$$\begin{aligned} f_{\text{bond}} : \mathcal{P}(\mathcal{I}) &\rightarrow \mathcal{P}(\mathcal{I}) \\ I &\mapsto f_{\text{bond}}(I) = \{i \in \mathcal{I} \mid i \in f_c(I) \cap f_d(I)\} \end{aligned}$$

Exemple 7 Soit le contexte présenté par le tableau 1.

$$\begin{array}{llll} f_d(AB) = AB & \text{et} & f_c(AB) = AB & \text{alors} & f_{\text{bond}}(AB) = AB. \\ f_d(BD) = ABCDEF & \text{et} & f_c(BD) = ABCDE & \text{alors} & f_{\text{bond}}(BD) = ABCDE. \end{array}$$

Proposition 6 L'opérateur f_{bond} est un opérateur de fermeture.

Preuve. Soient deux motifs $I, I' \subseteq \mathcal{I}$, $f_{\text{bond}}(I) = f_c(I) \cap f_d(I)$ et $f_{\text{bond}}(I') = f_c(I') \cap f_d(I')$.

(1) Extensivité (Nous cherchons à démontrer que $I \subseteq f_{\text{bond}}(I)$)

$$\begin{cases} f_c \text{ est un opérateur de fermeture} \Rightarrow I \subseteq f_c(I); \\ f_d \text{ est un opérateur de fermeture} \Rightarrow I \subseteq f_d(I); \end{cases} \Rightarrow \begin{cases} I \subseteq [f_c(I) \cap f_d(I)] \\ I \subseteq f_{\text{bond}}(I). \end{cases}$$

Ainsi, l'opérateur f_{bond} vérifie la propriété d'extensivité.

(2) Isotonie (Nous cherchons à démontrer que $I \subseteq I' \Rightarrow f_{\text{bond}}(I) \subseteq f_{\text{bond}}(I')$)

$$I' \subseteq I \Rightarrow \begin{cases} f_c(I') \subseteq f_c(I); \\ f_d(I') \subseteq f_d(I) \end{cases} \Rightarrow \begin{cases} [f_c(I') \cap f_d(I')] \subseteq [f_c(I) \cap f_d(I)] \\ f_{\text{bond}}(I') \subseteq f_{\text{bond}}(I). \end{cases}$$

Ainsi, l'opérateur f_{bond} vérifie la propriété d'isotonie.

(3) Idempotence (Nous cherchons à démontrer que $f_{\text{bond}}(f_{\text{bond}}(I)) = f_{\text{bond}}(I)$)

D'après la propriété (1), nous avons $f_{\text{bond}}(I) \subseteq f_{\text{bond}}(f_{\text{bond}}(I))$, démontrons par l'absurde que $f_{\text{bond}}(f_{\text{bond}}(I)) = f_{\text{bond}}(I)$.

Supposons que $f_{\text{bond}}(I) \subset f_{\text{bond}}(f_{\text{bond}}(I)) \Leftrightarrow \text{bond}(I) \neq \text{bond}(f_{\text{bond}}(I))$.

Ceci est impossible car $\text{bond}(f_{\text{bond}}(I)) = \text{bond}(I)$.

Ainsi, $f_{\text{bond}}(f_{\text{bond}}(I)) = f_{\text{bond}}(I)$, c.-à.-d., l'opérateur f_{bond} vérifie la propriété d'idempotence.

D'après (1), (2) et (3), l'opérateur f_{bond} est un opérateur de fermeture. \diamond

Dans ce qui suit, nous introduisons la notion de *motif fermé par f_{bond}* et de *motif minimal*.

Définition 11 - Motif fermé par f_{bond} - Soit $I \subseteq \mathcal{I}$. Le motif fermé de I par f_{bond} , noté $f_{\text{bond}}(I)$, est l'ensemble maximal d'items contenant I et ayant la même valeur de la mesure bond que le motif I .

Exemple 8 Soit le contexte présenté par le tableau 1. L'ensemble maximal d'items ayant la même valeur de la mesure bond que le motif AB est AB lui-même, alors $f_{\text{bond}}(AB) = AB$. De même, $f_{\text{bond}}(BD) = ABCDE$.

Définition 12 - Motif minimal corrélé - Soient $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction et $I \subseteq \mathcal{I}$. I est un motif minimal corrélé si et seulement si $\forall i \in I, \text{bond}(I) \geq \text{minbond}$:

$$\text{bond}(I) \neq \text{bond}(I \setminus \{i\})$$

Exemple 9 Soient le contexte illustré par le tableau 1 et $\text{minbond} = 0,3$. AB est un motif minimal corrélé car $\text{bond}(AB) \neq \text{bond}(A)$, $\text{bond}(AB) \neq \text{bond}(B)$ et $\text{bond}(AB) = 0,75 \geq 0,3$. Cependant, ABC n'est pas un motif minimal corrélé car $\text{bond}(BC) = \text{bond}(ABC) = 0,4$.

Après que nous avons défini un motif minimal, nous étudions la relation entre les motifs minimaux, les motifs essentiels et les générateurs minimaux.

Proposition 7 Un motif essentiel (resp. générateur minimal) est un motif minimal.

Preuve. Soit I un motif essentiel (resp. générateur minimal) alors $\forall i \in I, \text{Supp}(\vee I) > \text{Supp}(\vee (I \setminus \{i\}))$ et $\text{Supp}(\wedge I) \leq \text{Supp}(\wedge (I \setminus \{i\}))$ (resp. $\text{Supp}(\wedge I) < \text{Supp}(\wedge (I \setminus \{i\}))$) et $\text{Supp}(\vee I) \geq \text{Supp}(\vee (I \setminus \{i\}))$, alors $\frac{\text{Supp}(\wedge I)}{\text{Supp}(\vee I)} < \frac{\text{Supp}(\wedge (I \setminus \{i\}))}{\text{Supp}(\vee (I \setminus \{i\}))}$. Ainsi, $\text{bond}(I) \neq \text{bond}(I \setminus \{i\})$ et par conséquent I est un motif minimal. \diamond

Il est important de noter qu'un motif minimal peut n'être ni un motif essentiel ni un générateur minimal.

Exemple 10 Soit le contexte présenté dans le tableau 1. D'après l'exemple 9 (cf. page 8) AB est un motif minimal, mais nous avons démontré dans l'exemple 3 (cf. page 3) qu'il n'est pas un générateur minimal et dans l'exemple 4 (cf. page 4) qu'il n'est pas un motif essentiel.

La proposition suivante introduit la propriété d'idéal d'ordre des motifs minimaux corrélés.

Proposition 8 L'ensemble des motifs minimaux corrélés est un idéal d'ordre.

À présent, nous définissons la classe d'équivalence associée à la fermeture f_{bond} .

Définition 13 Classe d'équivalence associée à la fermeture f_{bond}

Une classe d'équivalence associée à la fermeture f_{bond} est un ensemble regroupant des éléments qui possèdent la même valeur de la mesure bond . Chaque classe est représentée par le motif fermé corrélé fréquent qui est le plus grand motif dans cet ensemble. Par contre, les plus petits motifs sont les motifs minimaux corrélés.

Un schéma illustratif d'une classe d'équivalence associée à la fermeture f_{bond} est donnée par la figure 1. La relation entre les éléments d'une telle classe et ceux des classes d'équivalence disjonctive et conjonctive est illustrée dans la même figure. En effet, la classe d'équivalence associée à la mesure bond est une présentation intermédiaire entre les deux classes d'équivalence disjonctive et conjonctive. Elle a la même structure que ces dernières : un motif maximal – un motif fermé corrélé – ainsi qu'un ensemble de motifs minimaux incomparables (selon la relation d'inclusion) – des motifs essentiels, des générateurs minimaux ou des motifs minimaux corrélés quelconques. Tout motif corrélé, compris entre un élément maximal et un ou plusieurs éléments minimaux, admet les mêmes caractéristiques – supports et fermeture – que ces derniers.

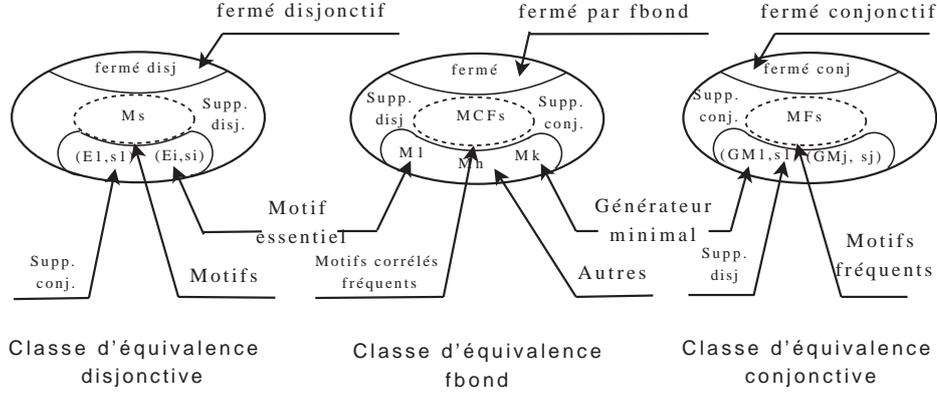


FIG. 1 – Caractérisation structurelle d'une classe d'équivalence induite par f_{bond} .

4.2 Nouvelle représentation concise associée à la mesure *bond*

En utilisant la fermeture associée à *bond*, nous pouvons construire deux représentations adéquates à cette mesure qui couvrent les mêmes motifs corrélés fréquents, sauf que la représentation basée sur les motifs fermés corrélés fréquents est plus compacte que celle basée sur les motifs minimaux corrélés fréquents. De ce fait, nous nous intéressons à la représentation la plus concise.

Dans un premier temps, définissons l'ensemble des motifs de cette représentation concise.

Définition 14 Désignons par $\mathcal{MF}\mathcal{CF}$ l'ensemble des motifs fermés corrélés fréquents défini comme suit :

$$\mathcal{MF}\mathcal{CF} = \{ X \subseteq \mathcal{I} \mid bond(X) > bond(X \cup \{i\}), \forall i \in \mathcal{I} \setminus X \}$$

Définition 15 Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. Désignons par $\mathcal{RM}\mathcal{F}\mathcal{CF}$ la représentation basée sur l'ensemble $\mathcal{MF}\mathcal{CF}$ définie comme suit :

$$\mathcal{RM}\mathcal{F}\mathcal{CF} = \{ (X, Supp(\wedge X), Supp(\vee X)) \mid X \in \mathcal{MF}\mathcal{CF} \}$$

Notons que le fait de maintenir à la fois les supports conjonctifs et disjonctifs des motifs retenus dans la représentation est nécessaire afin d'éviter le coût élevé de l'utilisation des identités d'inclusion-exclusion lors de la phase de régénération de l'ensemble total des motifs fréquents. Le théorème suivant montre que les éléments retenus dans la représentation forment une représentation exacte.

Théorème 1 La représentation $\mathcal{RM}\mathcal{F}\mathcal{CF}$ est une représentation concise exacte de l'ensemble $\mathcal{M}\mathcal{C}\mathcal{F}$ des motifs corrélés fréquents.

Preuve. Nous démontrons par récurrence que pour tout motif $X \subseteq \mathcal{I}$ corrélé fréquent, la fermeture de X est dans l'ensemble résultat $\mathcal{MF}\mathcal{CF}$: $f_{bond}(X) \in \mathcal{MF}\mathcal{CF}$. L'hypothèse de récurrence est vérifiée pour les items i qui correspondent aux motifs minimaux de taille 1 insérés dans $\mathcal{M}\mathcal{M}_1$ et dont la fermeture $f_{bond}(i)$ est insérée dans $\mathcal{MF}\mathcal{CF}_1$ si $Supp(\wedge i) \geq minsupp$. Ainsi, $f_{bond}(i) \in \mathcal{MF}\mathcal{CF}$. Supposons maintenant que $\forall X \subseteq \mathcal{I}$ tel que $|X| = n$ nous avons $f_{bond}(X) \in \mathcal{MF}\mathcal{CF}$. Nous démontrons alors que $\forall X \subseteq \mathcal{I}$ tel que $|X| = n + 1$,

Nouvelle représentation concise exacte des motifs corrélés fréquents

nous avons $f_{bond}(X) \in \mathcal{MFCCF}$. Soit un motif X de taille $n + 1$. Trois cas sont alors possibles :

- (a) X est un motif fermé corrélé fréquent : Si $X \in \mathcal{MFCCF}$ alors $f_{bond}(X) \in \mathcal{MFCCF}$ (évident).
- (b) X est un motif minimal corrélé fréquent : Si $X \in \mathcal{MM}_{n+1}$, alors $f_{bond}(X) \in \mathcal{MFCCF}_{n+1}$, et par conséquent $f_{bond}(X) \in \mathcal{MFCCF}$.
- (c) X est un motif quelconque corrélé fréquent : Si $X \notin \mathcal{MFCCF}$ et $X \notin \mathcal{MM}_{n+1}$ alors $\exists Y \subset X$ tel que $|Y| = n$ et $bond(X) = bond(Y)$, ainsi, $f_{bond}(X) = f_{bond}(Y)$, or d'après l'hypothèse, nous avons $f_{bond}(Y) \in \mathcal{MFCCF}$. Nous avons ainsi $f_{bond}(X) \in \mathcal{MFCCF}$. \diamond

Il est important de noter que la représentation \mathcal{RMFCCF} est une représentation parfaite de l'ensemble \mathcal{MCF} . En effet, la taille de ce dernier dépasse toujours celle de \mathcal{RMFCCF} quelle que soit la base et les valeurs de $minsupp$ et de $minbond$. Ceci résulte de la propriété de non-injectivité de l'opérateur de fermeture. Par ailleurs, la régénération à partir de la représentation \mathcal{RMFCCF} peut se faire d'une manière très efficace. En effet, au sein d'une classe d'équivalence associée à la mesure $bond$, les motifs possèdent la même valeur de la mesure $bond$ et par conséquent le même support conjonctif, disjonctif et négatif. De ce fait, pour dériver les données correspondantes à un motif, nous n'avons pas besoin des identités d'inclusion-exclusion. En effet, il suffit de chercher le plus petit fermé corrélé fréquent qui le couvre.

5 Résultats expérimentaux

Dans cette section, nous allons mener une étude expérimentale, réalisée sur des bases "benchmark", qui tendra à prouver que la représentation concise associée à la mesure $bond$ constitue un ensemble réduit de l'ensemble des motifs corrélés fréquents. Pour notre étude expérimentale, nous avons comptabilisé les cardinalités de notre représentation \mathcal{RMFCCF} , celle de l'ensemble \mathcal{MMCF} des motifs minimaux corrélés fréquents ainsi que celle de l'ensemble total \mathcal{MCF} des motifs corrélés fréquents, extraits à partir de bases "benchmark" ⁽³⁾. Toutes les expérimentations ont été réalisées sur une machine munie d'un processeur Intel®, ayant une fréquence d'horloge de 2 GHz, 4 Go de mémoire vive (avec 2 Go d'espace d'échange ou Swap) tournant sur une plateforme Linux Ubuntu 9.04. Notre algorithme est implanté en langage C++. Les programmes ont été compilés avec le compilateur gcc 4.3.3.

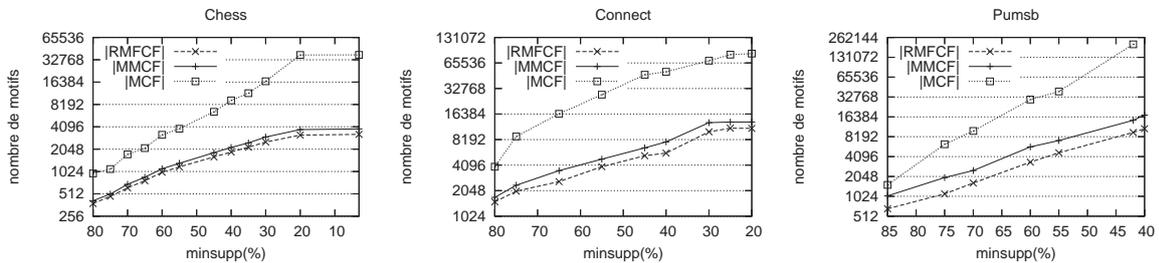
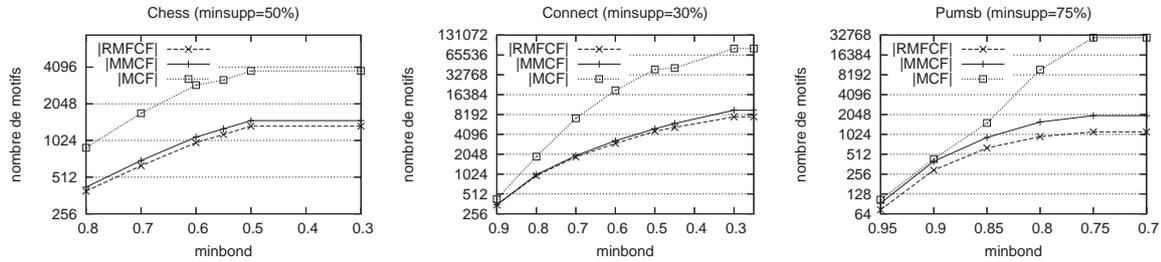


FIG. 2 – Variation des cardinalités en fonction de $minsupp$ avec $minbond = 0,25$.

³Ces bases sont disponibles à l'adresse suivante : <http://fimi.cs.helsinki.fi/data>.

FIG. 3 – Variation des cardinalités en fonction de *minbond*.

D’après les résultats expérimentaux, nous remarquons que la taille de notre représentation est non seulement plus petite que la taille de l’ensemble total des motifs corrélés fréquents mais aussi que la taille de l’ensemble des motifs minimaux corrélés fréquents (*cf.* Figures 2 et 3). Ce résultat est obtenu grâce à l’opérateur de fermeture f_{bond} qui permet de regrouper les motifs ayant les mêmes caractéristiques. De plus, d’après les courbes, nous constatons que les cardinalités augmentent avec la diminution de *minsupp* et/ou de *minbond*. Toutefois, il est clair que pour des seuils faibles, la taille de l’ensemble MCF est très élevée comparée à celle de la représentation $RMFCF$. Par contre, pour des seuils élevés la différence entre ces cardinalités est faible. Ceci est expliqué par le fait qu’en considérant une valeur élevée de seuil minimal de support, le nombre des motifs corrélés fréquents est très proche du nombre d’items corrélés fréquents. En effet, le fait d’avoir des items différents de leurs fermés est très rare puisque, dans ce cas, les items ayant une même fermeture doivent apparaître exactement dans les mêmes transactions.

Les résultats obtenus prouvent que la représentation proposée est très compacte pour les bases denses. Pour les bases éparées, sa taille est toujours nettement plus réduite que l’ensemble total des motifs corrélés fréquents. Toutefois, le taux de réduction est relativement faible comparé au cas des bases denses. Par exemple, pour la base ACCIDENTS, le taux de réduction est seulement égal à 20% pour *minsupp* = 45% et *minbond* = 0,25.

6 Conclusion et perspectives

Dans ce papier, nous nous sommes concentrés sur la proposition d’une nouvelle représentation concise des motifs corrélés fréquents basée sur la fermeture associée à la mesure de corrélation *bond*. Cette mesure symétrique qui vérifie la propriété d’anti-monotonie permet de réduire le nombre des motifs corrélés fréquents et de minimiser le coût de calcul de la représentation $RMFCF$. Suite aux expérimentations effectuées sur des contextes “benchmark” denses, nous avons démontré que cette représentation est concise. De plus, elle offre une régénération simple vu que la détermination des différents supports se fait sans avoir recours aux identités d’inclusion-exclusion.

Les perspectives de travaux futurs concernent : (1) La conception, l’implémentation et l’évaluation d’algorithmes efficaces d’extraction de la représentation que nous avons proposée et de régénération de l’ensemble total de motifs corrélés fréquents. (2) La déduction de l’ensemble total des motifs corrélés fréquents par rapport à la mesure de corrélation *all-confidence*, à partir de notre représentation concise. (3) Une autre perspective intéressante serait d’explorer la représentation proposée afin d’extraire de nouvelles règles présentant une disjonction d’items corrélés fréquents en prémisse ou en conclusion.

Références

- Bastide, Y., N. Pasquier, R. Taouil, L. Lakhal, et G. Stumme (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the International Conference DOOD'2000*, pp. 972–986.
- Bonchi, F. et C. Lucchese (2006). On condensed representations of constrained frequent patterns. *Knowledge and Information Systems* 9(2), 180–201.
- Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD'1997*, pp. 265–276.
- Casali, A., R. Cicchetti, et L. Lakhal (2005). Essential patterns : a perfect cover of frequent patterns. In *Proceedings of the 7th International Conference on DaWaK'2005*, pp. 428–437.
- Galambos, J. et I. Simonelli (2000). *Bonferroni-type inequalities with applications*. Springer.
- Hamrouni, T., S. Ben Yahia, et E. Mephu Nguifo (2009). Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets. *Data Knowledge Engineering* 68(10), 1091–1111.
- Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles* 44, 223–270.
- Kim, W. Y., Y. K. Lee, et J. Han (2004). CCMINE : efficient mining of confidence-closed correlated patterns. In *Proceedings of the 8th International Conference on PAKDD'2004*, pp. 569–579.
- Le Bras, Y., P. Lenca, et S. Lallich (2010). Mining interesting rules without support requirement : a general universal existential upward closure property. *Annals of Information Systems* 8, 75–98.
- Lee, Y. K., W. Y. Kim, Y. D. Cai, et J. Han (2003). COMINE : efficient mining of correlated patterns. In *Proceedings of the 3rd International Conference ICDM'2003*, pp. 581–584.
- Omiecinski, E. R. (2003). Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering* 15(1), 57–69.
- Pasquier, N., Y. Bastide, R. Taouil, G. Stumme, et L. Lakhal (2005). Generating a condensed representation for association rules. *Journal of Intelligent Information Systems* 24(1), 25–60.
- Soulet, A. et B. Crémilleux (2008). Adequate condensed representations of patterns. *Data Mining and Knowledge Discovery* 17(1), 94–110.
- Tanimoto, T. T. (1958). An elementary mathematical theory of classification and prediction. *Technical Report, I.B.M. Corporation Report*.
- Xiong, H., P.-N. Tan, et V. Kumar (2006). Hyperclique pattern discovery. *Data Mining and Knowledge Discovery* 13(2), 219–242.

Summary

Data mining is an extraction process of manageably-sized knowledge from huge large sets of data. The mined knowledge can be under the form of association rules which are correlations between attributes. However, the number of attributes is very high, what induced that the focus has been mainly on data with high frequency. The mined patterns are hence called frequent patterns. Unfortunately, such class of patterns does not offer information about the correlation ratio amongst the items that constitute a given pattern. In this respect, we propose in this paper a new concise representation of frequent correlated patterns, while defining the corresponding closure operator associated to the correlation measure *bond*. This reduced set makes it possible not only to derive all the frequent correlated patterns without information loss but also to derive the conjunctive, disjunctive and negative supports of each pattern in an efficient way.

Index des auteurs

- B -

Bahri, E., 41
Ben Yahia, S., 53
Ben Younes, N., 53

- D -

Dhouha, G., 15

- G -

Guillaume, S., 15

- H -

Hamrouni, T., 53

- L -

Lallich, S., 29
Le Bras, Y., 29
Lenca, P., 29

- M -

Mephu Nguifo, E., 15
Meyer, P., 29

- N -

Nouria, H., 41

- P -

Pierkot, C., 1