

1.2 Le programme 'Référentiel Client'

Ce programme est stratégique pour Hewlett Packard. Il est lui aussi rattaché à l'équipe 'Customer Knowledge Management and Data Stewardship'.

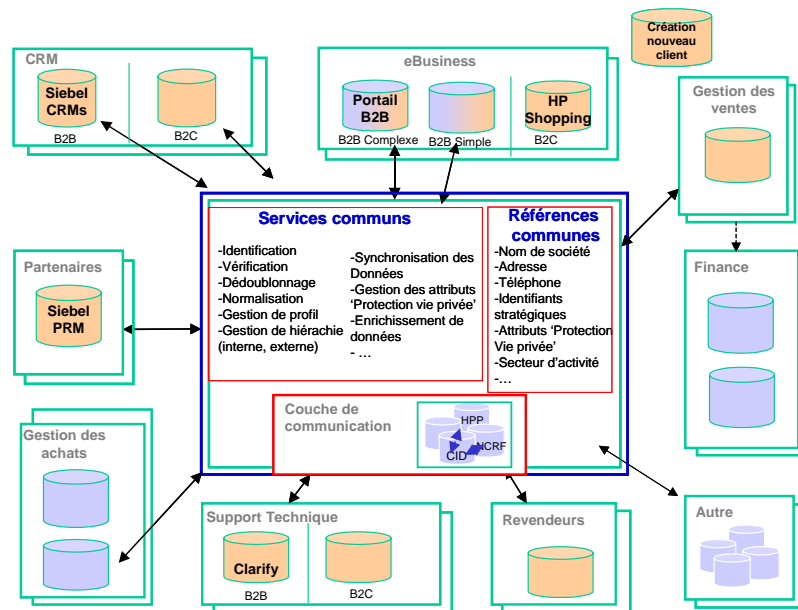


Tableau 1 : Architecture 'Référentiel Client'

'Référentiel Client' vise à dresser un inventaire de toutes les sources d'informations client au niveau Hewlett Packard monde, de leurs spécificités fonctionnelles et de leur architecture. Cet inventaire a pour but de définir en nombre limité des systèmes maîtres représentant chacun un bloc fonctionnel (Gestion des achats, Support Technique par exemple,... présentés sur le tableau 1). Ces systèmes maîtres sont, dans leur bloc fonctionnel, la source d'information et la référence pour les autres systèmes en aval. Le programme 'Référentiel Client', comme indiqué au centre du tableau 1, apporte :

- des services : débouclonnage, normalisation,...
 - une référence en terme de dictionnaire et de contenu : nom de société, adresse,...
- afin de permettre une vue centralisée et homogène des données clients chez Hewlett Packard.

1.3 L'objectif de l'audit

L'objectif est d'obtenir un baromètre qualité de l'ensemble des systèmes choisis comme maîtres dans chaque bloc fonctionnel. Comme nous l'avons dit plus haut, ces systèmes sont la référence en matière d'informations clients pour bon nombre d'autres systèmes en aval.

Il est donc nécessaire d'avoir une connaissance de la qualité de départ des références, de produire des métriques communes permettant de piloter la qualité dans le temps et d'amener les systèmes à nos objectifs par des actions correctives puis préventives ciblées.

2 Implémentation de l'audit

2.1 Le périmètre de l'audit

Parmi les différentes dimensions de la qualité de données présentées dans les ouvrages tels que Berti-Equille (1999) Huang et al. (1998), Loshin (2001) et Redman (2001), seules 2 dimensions sont analysées dans cet audit :

- La qualité intrinsèque de la donnée
- La qualité relative aux données homologues

Les autres dimensions, telles que la qualité de la représentation de la donnée par le système ou la qualité relative à l'utilisateur, seront abordées dans des étapes ultérieures.

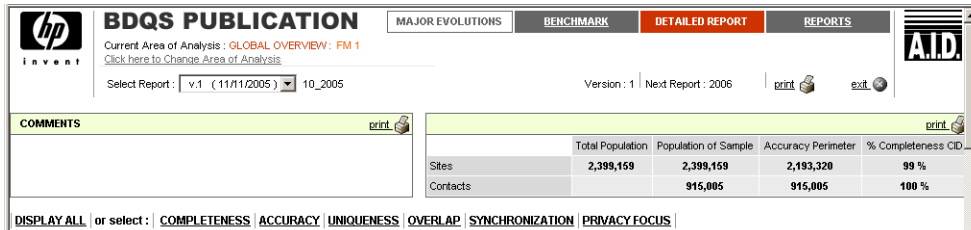
Déclinaison de ces 2 dimensions :

- La qualité intrinsèque des données de chaque système : pour les attributs inclus dans le dictionnaire de référence tels que le nom de société, l'adresse, le téléphone,..., des métriques ont été définies pour mesurer :
 - Le taux de remplissage
 - L'exactitude
- L'unicité de la représentation du client : la société 'A.I.D. 4 rue Henri Le Sidaner à Versailles' est-elle présente une seule fois dans le système analysé ? L'unicité a été mesurée en terme de sociétés, mais également en terme de personnes.
- Le taux de recouvrement des différents systèmes : nombre de clients présents dans le système des commandes et ayant appelé le support technique par exemple. Parmi ces clients communs, combien sont reconnus en tant que tels par les identifiants communs actuellement en place ?
- La synchronisation des différents systèmes : sur les clients reconnus par des identifiants communs entre 2 systèmes, quel taux d'erreur observe-t-on ? Par rapport aux métriques 'taux de recouvrement' présentées ci-dessus qui mesurent les clients communs non liés, il s'agit là de mesurer l'exactitude des liaisons : S'agit-il de la même entité ?
- La synchronisation des attributs de 'Protection de la vie privée' : l'objectif est de valider/quantifier les divergences éventuelles entre les systèmes. Par exemple, si Raoul Dupont a déclaré qu'il ne veut pas recevoir d'emails dans le système A, cette donnée est-elle bien répercutée dans le système B où Raoul Dupont est également enregistré ?

Le tableau 2 ci-dessous présente un exemple de publication résultant de l'audit avec les différents axes de mesure. Chaque système analysé est décliné par continent/pays et est comparé :

- aux autres systèmes (benchmark)
- sera comparé dans les prochaines mesures dans le temps avec lui-même (Major Evolutions).

Audit des données du 'Référentiel Client' Hewlett Packard



The screenshot shows the HP BDQS PUBLICATION interface. At the top, there are navigation tabs: MAJOR EVOLUTIONS, BENCHMARK, DETAILED REPORT (highlighted), and REPORTS. Below the HP logo, it says 'Current Area of Analysis: GLOBAL OVERVIEW: FM 1' and 'Click here to Change Area of Analysis'. There is a 'Select Report' dropdown menu showing 'v.1 (11/11/2005)' and '10_2005'. To the right, it says 'Version: 1' and 'Next Report: 2006'. Below this, there is a 'COMMENTS' section with a 'print' icon. The main content area contains a table with the following data:

	Total Population	Population of Sample	Accuracy Perimeter	% Completeness CID
Sites	2,399,159	2,399,159	2,193,320	99 %
Contacts		915,005	915,005	100 %

At the bottom, there is a 'DISPLAY ALL' button and a list of filters: COMPLETENESS, ACCURACY, UNIQUENESS, OVERLAP, SYNCHRONIZATION, and PRIVACY FOCUS.

Tableau 2. Publication des résultats

2.2 Exemples de métriques mises en place

2.2.1 Pollution

Pourquoi cette métrique ?

Les taux de remplissage, dans un système d'information, sont souvent pollués par :

- Des caractères parasites, ses pseudo-remplissages 999,AAAAAA,..., des mots 'injurieux' ou pseudo-remplissage : UNKNOWN, NONE ,...

A l'origine, on peut trouver des zones obligatoires dans un logiciel qui incitent/obligent l'utilisateur à polluer par du remplissage néfaste. Il est également possible d'avoir le fruit de données provenant de questionnaires WEB, souvent plus frustrés en contrôle en entrée. La base est alors polluée, avec des conséquences néfastes dans l'image de la société lorsqu'elle communique à un client/prospect.

Méthode de calcul

Des dictionnaires de mots, caractères, répétitions sont décrits par variable /pays avec un algorithme très simple de détection dans les enregistrements.

2.2.2 Contrôle des emails

Pourquoi cette métrique ?

L'objectif est de détecter, par anticipation, avant l'opération d'e-mailing où ils vont être utilisés, le nombre d'emails faux. Contrairement aux courriers ou téléphones, le fait d'envoyer un email faux a de faibles conséquences financières, en revanche l'emailing ne permettra, en aucun cas, de trouver le nouvel email de la personne : pas de facteur 'intelligent' ou de service de transfert de courrier, pas de disque avec nouveau téléphone. Donc, anticiper a réellement un sens pour gagner en temps et déclencher des actions pour qualifier en emails.

Méthode de calcul

En terme d'emails, à notre connaissance, il n'existe pas de moyen fiable actuellement pour valider l'email, si ce n'est de l'envoyer. En effet, les fonctions de demande d'existence disponibles dans les moteurs SMTP sont en général annulées par les responsables réseau. Les

algorithmes mis en place n'ont pas comme objectif de valider un email, *mais plutôt de détecter les emails faux : le périmètre restant sera supposé 'vraisemblable'*. Indicateurs mis en place :

- Détection de syntaxe incorrecte : pas de @, présence de caractères parasites,...
- Contrôle du nom de domaine : la validité du nom de domaine est faite on-line .
- Contrôle de la cohérence nom/prénom : spécifiquement en BtoB, il existe très souvent une 'charte' d'attribution des emails dans l'entreprise : début du prénom, nom,... Un certain nombre de formats ont été référencés afin de produire un indice de cohérence.

2.2.3 Contrôle de l'adresse

Pourquoi cette métrique ?

Avec le téléphone et l'email, l'adresse fait partie des éléments qui permettent de communiquer par un canal avec les clients/prospects. Avoir une adresse exacte signifie, en général :

- Une adresse qui existe : la ville, le code postal, la rue, le numéro dans la rue
- Une adresse libellée correctement : pas de fautes d'orthographe
- Une adresse aux normes postales : selon les pays, La Poste locale, afin d'améliorer la distribution du courrier en efficacité (lecture optique par exemple), a mis en place une norme de définition de l'adresse : positionnement du code postal, ville dans les lignes adresses, rue, bâtiment,.. ainsi que respect des référentiels rue/ville/code postal mis à disposition.

Les intérêts d'avoir une adresse exacte sont nombreux : moins de courriers inaboutis (coût important), tarifs postaux préférentiels, meilleure efficacité du dédoublement.

Méthode de calcul

Le contrôle, la conformité des adresses peuvent être contrôlés par des logiciels de normalisation d'adresse internationaux, tels que First Logic™. Dans le cas de notre projet, c'est ce dernier que nous avons utilisé. L'outil logiciel comporte plusieurs phases :

- recherche dans l'adresse des éléments clés : code postal, ville, rue, numéro dans la rue, boîte postale,...
- validation par rapport aux référentiels géographiques locaux
- si la validation n'aboutit pas directement, un certain nombre d'approximations, transformations sont opérées pour corriger l'adresse et la retrouver dans les référentiels.

L'indice produit aura la valeur :

- Adresse correcte , pas de modification nécessaire
- Adresse correcte, mais après modification nécessaire
- Adresse incorrecte

2.2.4 Contrôle du téléphone

Pourquoi cette métrique ?

La présence de faux téléphones est coûteuse, pénalisante pour les centres d'appels lorsqu'ils doivent procéder à des appels sortants. En appel entrant, si le coût des 'faux' est moins élevé, il existe également : de nombreux standards téléphoniques sont équipés de la reconnaissance des numéros de téléphone entrants, afin que l'opérateur ait directement accès à la fiche du client, du prospect. Les faux téléphones diminuent l'efficacité du système. Enfin, dans les systèmes d'information qui n'utilisent pas les téléphones, l'erreur, si elle ne génère pas de coûts directs, influe sur la qualité générale de la base, en particulier sur le taux de doubles : le téléphone est un élément de comparaison important pour retrouver les doubles sites dans une base.

Méthode de calcul

La validation d'un téléphone peut être faite selon plusieurs méthodes :

- Test du téléphone directement, par un opérateur : efficace, mais très coûteux.
- Recherche dans un annuaire
- Contrôle syntaxique simple : c'est cette dernière méthode que nous avons utilisée.

Un numéro de téléphone est composé de :

Code international d'appel (00 en général) - Code téléphonique du pays (33 pour la France)
- Code troncature (en général 0 : à concaténer au numéro uniquement pour les appels à l'intérieur du pays) - Code zone (correspond en général à une localisation géographique dans le pays : ensemble de départements en France) - Numéro de téléphone

Par exemple 00 – 44 – 0 – 20 - 72987174

L'algorithme mis en place se base sur des tables de référence compilées qui incluent en général :

- D'une manière quasi-exhaustive les longueurs possibles pour le code zone, le numéro de téléphone.
- De manière beaucoup plus inégales la liste des codes zones avec leur description en terme de villes par exemple.

2.2.5 Unicité

Pourquoi cette métrique ?

L'objectif est de mesurer le nombre de sociétés, de personnes présents plus d'une fois dans le système d'information sans être identifiés en tant que même entité.

Méthode de calcul

Un logiciel de dédoublement du marché (Proxim™, développé par A.I.D. dans notre cas) est mis en œuvre. Le principe est le suivant : dans une zone de comparaison (Ville, E-mail, Téléphone,...), les enregistrements sont comparés sur des variables 'pertinentes' telles que :

- Pour les sociétés : nom de l'entreprise, adresse, téléphone, numéro de TVA, ...
- Pour les personnes : nom, prénom, email, adresse, téléphone,...

La comparaison est faite à partir d'algorithmes tels que :

- Distance d'édition : Wood (1992)
- Présence de lettres communes : soit 2 chaînes de caractères S1 et S2, nc le nombre de lettres communes entre les 2 chaînes, l'indice de proximité $Ind = nc / \text{nombre maximum de caractères (S1,S2)}$. Un indice dérivé Ind' consiste à diviser par le minimum.
- De nombreux autres algorithmes publics existent, Berti-Equille (2005). Les logiciels de dédoublement tels que Proxim™ intègrent également en général des algorithmes propriétaires.

Zone de comparaison : les comparaisons des enregistrements sont restreintes à un sous périmètre pour des raisons de performance. En effet, le nombre de comparaisons $n(n-1)/2$ peut prendre rapidement un peu de temps sur des volumes importants ($n = \text{nombre d'enregistrements de la population à dédoublement}$, peut valoir plusieurs millions d'enregistrements). La combinaison de conditions en ET, OU entre les différentes variables, leurs indices de proximité, et également en utilisant les attributs qualité Wang (2000) des variables va permettre de décider si 2 enregistrements sont doublons ou pas.

2.3 Freins rencontrés lors de la mise en place des métriques

Mettre en place les métriques définies plus haut :

- Dans un contexte international
- Avec plusieurs systèmes d'information à mesurer

n'est pas neutre. Une illustration en est la notion de référentiels de contrôle : postaux, téléphones, prénoms, dictionnaires de mots de pollution, ... La qualité de ces référentiels, leur exhaustivité sont très inégaux d'un pays à l'autre.

Nous rencontrons en général le même biais :

- Les enregistrements considérés comme corrects le sont avec un taux d'erreur proche de 0.
- Les enregistrements considérés comme erronés le sont à tort, dans une proportion très variable selon le pays, la qualité des référentiels, l'adéquation de l'algorithme.

Pour pallier ce défaut, nous utilisons la méthode suivante :

Un échantillon aléatoire est tiré par pays dans les enregistrements notés erronés. Cet échantillon est validé manuellement par des personnes locales capables :

- de téléphoner pour valider
- avoir une meilleure connaissance des prénoms
- de trouver l'information dans des référentiels locaux moins connus

Le système apprend, s'enrichit. Les résultats publiés en terme d'indicateurs sont extrapolés selon le contrôle manuel de l'échantillon.

3 Résultats

3.1 Tableau de bord

Afin de porter les résultats qualité à la connaissance de nos décideurs, nous avons positionné les systèmes et leurs résultats par métrique dans un tableau de bord, divisé en trois régions (EMEA, Amériques, AP). Les systèmes étaient ainsi comparés les uns par rapport

Audit des données du 'Référentiel Client' Hewlett Packard

aux autres mais également par rapport à un objectif de résultat. Sur chaque métrique et par région, le système pouvait être vert car égal ou supérieur à l'objectif de résultat, jaune car en deçà de l'objectif mais supérieur à 50% (85% dans le cas du taux de doublons) ou rouge car en deçà de 50% (85% dans le cas du taux de doublons). La présentation d'un tel tableau de bord, assorti de commentaires, a permis de créer une véritable prise de conscience de l'intérêt et de l'importance de la qualité des données parmi les décideurs. Nous avons donc recueilli leur soutien pour la poursuite du projet et la mesure de l'évolution.

3.2 Mise en oeuvre opérationnelle

Nous avons organisé une revue individuelle plus approfondie avec les responsables de chacun des domaines fonctionnels. Nous leur avons présenté au niveau Région/Pays le détail des métriques. Nous avons par ailleurs gagné leur confiance et leur attention en montrant des échantillons de données erronées. Nous avons essayé de comprendre ensemble quels étaient les flux ou traitements qui pouvaient expliquer tel ou tel autre résultat. Nous avons partagé avec eux nos recommandations. Ainsi, au fil de la discussion, chaque responsable de domaine a défini ses priorités d'action immédiate : remplissage et exactitude d'identifiants stratégiques, protection de la vie privée, taux de doublons, etc... Sans exception, tous les utilisateurs des systèmes ont salué l'existence et l'intérêt d'un tel projet.

Conclusion

Ce premier audit était un challenge :

- De multiples sources de données avec nécessité de prendre connaissance des règles de gestion spécifiques à chaque système .
- Plus de 100 pays à couvrir, avec un nombre important de jeux de caractères différents.
- Des délais très courts d'implémentation.
- Une attente forte de la part des responsables de domaines d'obtenir des résultats directement opérationnels.

Challenge réussi, la prochaine étape prévue est de comparer les métriques dans le temps : prochain audit en février 2006. Il est également apparu une évolution nécessaire dans les analyses : définir des segments de clients. La limite des métriques produites actuellement est qu'elles restent assez techniques tant qu'elles ne sont pas déclinées par segment client. Ce dernier correspond en général à une valeur marketing différente, à des niveaux et des moyens très hétérogènes dans la qualité de données, à des niveaux d'exigence (objectifs) spécifiques. Leur introduction permettra d'adapter les recommandations/résultats à la réalité métier de Hewlett Packard.

Mener un audit a été possible de part l'organisation de Hewlett Packard, avec une cellule qualité de données existante et rattachée à un niveau hiérarchique élevé. Cela correspond à une réelle préoccupation de la qualité de données client dans cette société, ce qui est encore, à notre connaissance, un phénomène précurseur.

Lexique

Customer Data Integrity : Intégrité des données Client
Customer Knowledge Management and Data Stewardship : Gestion de la connaissance client et services sur les données
Internet and Marketing Services : Services Internet et Marketing
B2B : Business to Business : Vente aux entreprises
B2C : Business to Consumer : Vente au grand public
Benchmark : Comparaison
Major evolutions : Evolutions majeures

Références

- Berti-Equille, L. (1999), *La qualité des données et leur recommandation : modèle conceptuel, formalisation et application à la veille technologique* : Thèse de doctorat, Université de Toulon et du Var.
- Berti-Equille, L. (2005), *Journées CRM & Qualité des Données au CNAM – Qualité des données multi-sources : un aperçu des techniques issues du monde académique*.
- Huang, K. T., Lee, Yang W., Wang R. (1998), *Quality Information and Knowledge*. Prentice Hall
- Loshin, D. (2001), *Enterprise Knowledge Management : the Data Quality Approach*. Morgan Kaufmann Publishers
- Redman, T. C. (2001), *Data Quality : The Field Guide*. Digital Press
- Wang, R. (2000), *Data Quality*. Kluwer Academic Publishers
- Wood, D. (1992), *Data structures, algorithms and Performance*. Addison-Wesley

Summary

Measuring, improving the customer data quality level in an international environment with multiple information systems is the daily challenge of the Customer data quality team at Hewlett Packard. This article describes the metrics defined and the reasons of their choices, the implementation for an audit of six worldwide customer/prospect information systems. The brakes, technical (lack of control referentials in some countries), operational (to obtain the buy-in of the results by the data owners) are also described.

Normalisée d'une mesure probabiliste de qualité des règles d'association : étude de cas

Daniel Feno ^{*,**}, Jean Diatta^{*}

André Totohasina^{**}

^{*}Université de la Réunion

15-Avenue René Cassin-B.P. 7151- 97715 Saint-Denis Messag cedex 9 France

{drfeno,jean.diatta}@univ-reunion.fr,

^{**}Université d'Antsiranana-BP O - 201-Antsiranana-Madagascar

totohasina@yahoo.fr

Résumé. Cet article concerne les mesures probabilistes de qualité des règles d'association. Nous donnons une condition nécessaire et suffisante pour qu'une mesure probabiliste de qualité soit normalisable. Par ailleurs, nous considérons une trentaine de mesures probabilistes de qualité proposées dans la littérature et montrons que la plupart d'entre celles qui sont normalisables ont la même normalisée.

1 Introduction

De nombreuses de mesures de qualité des règles d'association ont été proposées dans la littérature pour tenir compte de diverses caractéristiques (Piatetsky-Shapiro, 1991; Major et Mangano, 1993; Freitas, 1999; Blanchard et al., 2005). Ainsi, pour aider l'utilisateur final dans son choix, des études expérimentales ont été menées pour l'évaluation de mesures existantes (Tan et al., 2002; Lenca et al., 2004; Lallich et Teytaud, 2004). Dans cet article, nous proposons une approche analytique de comparaison de mesures probabilistes de qualité des règles (MPQ). Cette approche est fondée sur une notion de normalité introduite dans (Totohasina, 2003). Nous caractérisons les MPQ normalisables. Par ailleurs, nous montrons que la plupart des MPQ normalisables ont la même normalisée, à savoir la mesure M_{GK} introduite dans (Guillaume, 2000) et dont les propriétés mathématiques ont été étudiées dans (Totohasina, 2003; Totohasina et al., 2005). Le reste de l'article est organisé de la façon suivante. Dans la section 2, nous introduisons les MPQ et en présentons quelques exemples. La normalisation des MPQ est présentée dans la section 3. Dans la section 4, nous donnons les normalisées de plus d'une vingtaine de MPQ avant de terminer par une brève conclusion.

2 Mesures probabilistes de qualité

Dans cet article, nous nous plaçons dans le cadre d'un contexte binaire $(\mathcal{E}, \mathcal{V})$ où \mathcal{E} est un ensemble fini d'entités et $\mathcal{V} = \{a_1, a_2, \dots, a_m\}$ un ensemble fini de variables booléennes

Normalisée d'une mesure de qualité des règles d'association

	a_1	a_2	a_3	a_4	a_5
e_1	1	0	1	1	0
e_2	0	1	1	0	1
e_3	1	1	1	0	1
e_4	0	1	0	1	1
e_5	1	1	1	0	1

TAB. 1 – Exemple d'un contexte binaire

définies sur \mathcal{E} . Les sous ensembles de \mathcal{V} seront appelés *motifs*. Le tableau 1 présente un contexte binaire ayant 5 variables définies sur un ensemble de 5 entités.

Une *règle d'association* de $(\mathcal{E}, \mathcal{V})$ est un couple (A, B) de motifs, noté $A \rightarrow B$, où B est non vide. Les motifs A et B sont appelés respectivement la “*prémisse*” et la “*conclusion*” de la règle $A \rightarrow B$.

Etant donnés deux motifs A et B :

- A' désignera l'événement de l'espace probabilisable $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$, vérifiant le motif A , i.e., $A' = \{e \in \mathcal{E} : \forall a \in A, a(e) = 1\}$;
- \overline{A} désignera la négation de A , i.e., $\overline{A}(e) = 1$ si et seulement si il existe $a \in A$ tel que $a(e) = 0$ ($(\overline{A})'$ est le complémentaire de A').

Dans tout ce qui suit, n désignera le cardinal de \mathcal{E} et p une probabilité sur $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$, définie, pour tout événement E de \mathcal{E} , par $p(E) = \frac{|E|}{n}$, où $|E|$ désigne la cardinalité de E .

Une *mesure probabiliste de qualité* des règles d'association (MPQ) est une fonction μ de l'ensemble $\mathcal{P}(\mathcal{V}) \times \mathcal{P}(\mathcal{V})$ à valeurs dans \mathbb{R} telle que $\mu(A \rightarrow B)$ s'exprime en fonction des probabilités $p(A'), p(B'), p(A' \cap B')$.

Exemple 1 – Le support (Agrawal et al., 1993) défini par $Support(A \rightarrow B) = p(A' \cap B')$ est une MPQ symétrique (i.e. $Support(A \rightarrow B) = Support(B \rightarrow A)$).

– La confiance (Agrawal et al., 1993) définie par $Confiance(A \rightarrow B) = p(B' / A')$ est une MPQ non symétrique.

– La mesure de conviction (Brin et al., 1997) définie par $Conviction(A \rightarrow B) = \frac{p(A') \cdot p(\overline{B}')}{p(A' \cap \overline{B}')}$ est une MPQ implicative (i.e. $Conviction(A \rightarrow B) = Conviction(\overline{B} \rightarrow \overline{A})$).

– La mesure M_{GK} (Guillaume (2000)) définie par

$$M_{GK}(A \rightarrow B) = \begin{cases} \frac{p(B' / A') - p(B')}{1 - p(B')} & \text{si } A \text{ favorise } B \text{ i.e. } p(B' / A') > p(B') \\ \frac{p(B' / A') - p(B')}{p(B')} & \text{si } A \text{ défavorise } B \text{ i.e. } p(B' / A') < p(B') \end{cases}$$

est une MPQ non symétrique implicative dans le cas où A favorise B .

– La surprise (Aze et Kodratoff, 2002) définie par

$$Surprise(A \rightarrow B) = \frac{p(A' \cap B') - p(A' \cap \overline{B}')}{p(B')} \text{ est une MPQ ni symétrique ni implicative.}$$

Règle	Support [0; 1]	Confiance [0; 1]	M_{GK} [-1; 1]	Conviction [0; +∞[Surprise [-1; 1]
$A \rightarrow B$	0	0	-1	$\frac{4}{5}$	-1
$B \rightarrow C$	1	1	1	$+\infty$	1
$C \rightarrow D$	$\frac{1}{2}$	1	1	$\frac{6}{5}$	$\frac{1}{2}$
$A \rightarrow C$	$\frac{2}{3}$	$\frac{2}{3}$	$-\frac{1}{6}$	$\frac{5}{6}$	$\frac{1}{4}$

TAB. 2 – Valeurs prises par quelques MPQ

3 Normalisation d'une mesure de qualité

3.1 Motivation et définition

Considérons le contexte binaire du tableau 1 et prenons les motifs $A = a_1a_3$, $B = a_2a_4$, $C = a_5$, $D = a_1a_2$. Le tableau 2 présente les valeurs prises par quelques MPQ sur certaines règles de ce contexte.

On constate que les valeurs prises par ces mesures sont distribuées entre -1 et $+\infty$. De plus, certaines mesures ne prennent que des valeurs positives indépendamment du fait que la prémisse favorise ou défavorise la conclusion. L'objectif de la normalisation est alors, d'une part, de réduire et centrer les valeurs d'une MPQ sur l'intervalle $[-1, 1]$, et, d'autre part, de refléter les situations de référence telles que l'incompatibilité (*i.e.* $A' \cap B' = \emptyset$), la répulsion (*i.e.* $p(B'/A') < p(B')$), l'indépendance (*i.e.* $p(A' \cap B') = p(A').p(B')$), l'attraction (*i.e.* $p(B'/A') > p(B')$) et l'implication (*i.e.* $A' \subseteq B'$). D'où la définition suivante.

Une MPQ μ sera dite *normalisée* si elle vérifie les cinq conditions ci-dessous :

- (1) $0 < \mu(A \rightarrow B) \leq 1$ si A favorise B (*i.e.* $p(B'/A') > p(B')$);
- (2) $-1 \leq \mu(A \rightarrow B) < 0$ si A défavorise B (*i.e.* $p(B'/A') < p(B')$);
- (3) $\mu(A \rightarrow B) = 1$ si A implique B (*i.e.* $p(B'/A') = 1$);
- (4) $\mu(A \rightarrow B) = -1$ si A et B sont incompatibles (*i.e.* $p(A' \cap B') = 0$);
- (5) $\mu(A \rightarrow B) = 0$ si A et B sont indépendants (*i.e.* $p(A' \cap B') = p(A').p(B')$).

Exemple 2 Les MPQ suivantes sont normalisées.

- La mesure M_{GK} (Guillaume, 2000)
- La mesure de Zhang (Zhang, 2000)

3.2 Caractérisation des MPQ normalisables

Dans ce paragraphe, nous ne nous intéressons qu'aux MPQ continues par rapport aux arguments $p(A')$, $p(B')$, $p(A' \cap B')$. Nous donnons une condition nécessaire et suffisante pour qu'une telle MPQ soit normalisable. Considérons une MPQ μ et désignons par μ_n la mesure normalisée associée si elle existe. La normalisation de μ consisterait à la centrer et la réduire pour obtenir μ_n . Soit x_f (resp. y_f) le coefficient de réduction (resp. de centrage) de μ , dans le cas où A favorise B . De façon similaire, soit x_d (resp. y_d) le coefficient de réduction (resp. de centrage) dans le cas où A défavorise B . On a donc :

$$\mu_n(A \rightarrow B) = \begin{cases} x_f \cdot \mu(A \rightarrow B) + y_f & \text{si } A \text{ favorise } B \\ x_d \cdot \mu(A \rightarrow B) + y_d & \text{si } A \text{ défavorise } B \end{cases} \quad (1)$$

Normalisée d'une mesure de qualité des règles d'association

Mesure : $\mu(A \rightarrow B)$	Expression probabiliste	Référence
Multiplicateur de cote	$\frac{p(A' \cap B') \cdot p(B')}{p(A' \cap \overline{B'}) \cdot p(B')}$	(Lallich et Teytaud, 2004)
Sebag	$\frac{p(B'/A')}{p(\overline{B}/A')}$	(Sebag et Shoenuer, 1988)
Conviction	$\frac{p(A') \cdot p(\overline{B}')}{p(A' \cap \overline{B}')$	(Brin et al., 1997)
Odd Ratio	$\frac{p(A' \cap B') \cdot p(\overline{A} \cap \overline{B}')}{p(\overline{A} \cap B') \cdot p(A' \cap \overline{B}')$	cf. (Huynh et al., 2005)
Klogsen	$\sqrt{p(A' \cap B')(p(B'/A') - p(B'))}$	cf. (Huynh et al., 2005)

TAB. 3 – Liste de MPQ non normalisables

Du fait de la continuité de μ , ces quatre coefficients se déterminent par passage aux limites aux situations de référence que sont l'incompatibilité, l'indépendance et l'implication logique). Posons $\mu_{imp}(A \rightarrow B)$ la limite de $\mu(A \rightarrow B)$ à l'implication, $\mu_{ind}(A \rightarrow B)$ la limite à l'indépendance et $\mu_{inc}(A \rightarrow B)$ la limite à l'incompatibilité. Ainsi, l'équation (1) est équivalente à l'équation matricielle (2) ci-dessous.

$$\begin{pmatrix} \mu_{imp}(A \rightarrow B) & 1 & 0 & 0 \\ \mu_{ind}(A \rightarrow B) & 1 & 0 & 0 \\ 0 & 0 & \mu_{ind}(A \rightarrow B) & 1 \\ 0 & 0 & \mu_{inc}(A \rightarrow B) & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \\ x_d \\ y_d \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} \quad (2)$$

D'où la caractérisation suivante de MPQ normalisables, résultant de l'existence de solution de l'équation (2).

Théorème 1 Une MPQ μ est normalisable si et seulement si, pour toute règle $A \rightarrow B$, les conditions suivantes sont vérifiées :

- (i) les limites $\mu_{imp}(A \rightarrow B)$, $\mu_{ind}(A \rightarrow B)$ et $\mu_{inc}(A \rightarrow B)$ sont finies ,
- (ii) $\mu_{imp}(A \rightarrow B) \neq \mu_{ind}(A \rightarrow B)$ et $\mu_{ind}(A \rightarrow B) \neq \mu_{inc}(A \rightarrow B)$.

Le tableau 3 présente une liste de 5 mesures non normalisables.

4 Normalisées de mesures existantes

Dans cette section, nous donnons les normalisées associées à un certain nombre de MPQ normalisables.

Voici une liste de 19 MPQ de normalisée M_{GK} : Support, Confiance (Agrawal et al., 1993), ϕ -coefficient, Ind-Imp-Lerman (Lerman et al., 1981), Cosinus, Facteur de certitude, Laplace, Kappa (cf. (Huynh et al., 2005)), Rappel, Nouveauté, Satisfaction, Fiabilité négative, Spécificité (Lavrac et al., 1999), Surprise (Aze et Kodratoff, 2002), Piatetsky-Shapiro (Piatetsky-Shapiro, 1991), Confiance-centrée (Lallich et Teytaud, 2004), Pearl (Pearl, 1988), Lift (Brin et al., 1997), Dépendance (cf. (Huynh et al., 2005)).

Le tableau 4 donne une liste de 6 MPQ normalisables dont la normalisée associée n'est pas M_{GK} .

Mesure $A \rightarrow B$	Expression	Référence
Jaccard	$\frac{p(A' \cap B')}{p(A') + p(B') - p(A' \cap B')}$	cf. (Huynh et al., 2005)
Q-Yule	$\frac{p(A' \cap B') \cdot p(\bar{A}' \cap \bar{B}') - p(A' \cap \bar{B}') \cdot p(\bar{A}' \cap B')}{p(A' \cap B') \cdot p(\bar{A}' \cap \bar{B}') + p(A' \cap \bar{B}') \cdot p(\bar{A}' \cap B')}$	cf. (Huynh et al., 2005)
Y-Yule	$\frac{\sqrt{p(A' \cap B') \cdot p(\bar{A}' \cap \bar{B}')} - \sqrt{p(A' \cap \bar{B}') \cdot p(\bar{A}' \cap B')}}{\sqrt{p(A' \cap B') \cdot p(\bar{A}' \cap \bar{B}')} + \sqrt{p(A' \cap \bar{B}') \cdot p(\bar{A}' \cap B')}}}$	cf. (Huynh et al., 2005)
J-mesure	$p(A' \cap B') \cdot \log\left(\frac{p(A' \cap B')}{p(A')p(B')}\right) + p(A' \cap \bar{B}') \cdot \log\left(\frac{p(A' \cap \bar{B}')}{p(A')p(\bar{B}')}\right)$	(Goodman et Smyth, 1998)
Zhang	$\frac{p(A' \cap B') - p(A') \cdot p(B')}{\max\{p(A' \cap B') \cdot p(\bar{B}'); p(B') \cdot p(A' \cap \bar{B}')\}}$	(Zhang, 2000)

TAB. 4 – Liste de MPQ dont la normalisée n'est pas M_{GK} .

5 Conclusion

Nous avons étudié une trentaine de MPQ sous l'angle de la notion de normalisation qui permet de refléter les situations de référence telles que l'incompatibilité, la répulsion, l'indépendance, l'attraction et l'implication. Nous avons caractérisé les MPQ normalisables et montré que la plupart des MPQ proposées dans la littérature ont la même normalisée, à savoir la mesure normalisée M_{GK} introduite dans (Guillaume, 2000). Ce résultat nous permet d'identifier trois classes de MPQ : (a) celles qui sont non normalisables, (b) les normalisables de normalisée M_{GK} et (c) les autres.

Remerciements

Daniel Feno remercie l'A.U.F. pour son soutien financier.

Références

- Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proc. of the ACM SIGMOD International Conference on Management of Data*, Volume 22, Washington, pp. 207–216. ACM press.
- Aze, J. et Y. Kodratoff (2002). Evaluation de résistance au bruit de quelques mesures d'extraction de règles d'association. *Extraction de Connaissances et Apprentissage*. 14, 143–154.
- Blanchard, J., F. Guillet, H. Briand, et R. Gras (2005). Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *Proc. of Applied stochastic Models and Data Analysis*, ENST Bretagne, France, pp. 334–344.
- Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : Generalizing association rules to correlation. In *Proc. of the ACM SIGMOD Conference*, pp. 265–276.
- Freitas, A. (1999). On rule interestingness measures. *Knowledge-Based System* 12, 309–315.

- Goodman, R. et P. Smyth (1998). Information theoretic rule induction. In *Proc. of the ECAI-98*, pp. 357–362.
- Guillaume, S. (2000). *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales*. Ph. D. thesis, Université de Nantes, France.
- Huynh, X., F. Guillet, et H. Briand (2005). Une plateforme exploratoire pour la qualité des règles d'association : Apport pour l'analyse implicite. In *Proc. of Troisièmes Rencontres Internationales A.S.I.*, pp. 339–349.
- Lallich, S. et O. Teytaud (2004). Evaluation et validation de mesures d'intérêt des règles d'association. *RNTI-E-1*, 193–217.
- Lavrac, N., P. Flach, et B. Zupan (1999). Rule evaluation measures : A unifying view. In G. Mineau et B. Ganter (Eds.), *Ninth international workshop on Inductive Logic Programming*, Volume 1634, pp. 174–185.
- Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *RNTI-E-1*, 219–246.
- Lerman, I., R. Gras, et H. Rostam (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires. *Math Sc. Hum* 74, 5–35.
- Major, J. A. et J. Mangano (1993). Selecting among rules induced from a heuristic database. In *KDD Workshop papers, Menlo Park California*, pp. 28–41.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann.
- Piatetsky-Shapiro, G. (1991). Knowledge discovery in real databases. *IA Magazine* 11, 68–70.
- Sebag, M. et M. Shoenauer (1988). Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proc. of the European Knowledge Acquisition Workshop Conference*, pp. 28–1–28–20.
- Tan, P., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measures for association patterns. In *Proc. of the 8th KDD Conference*, pp. 32–41.
- Totohasina, A. (2003). Normalisation de mesures probabilistes de la qualité des règles. In *Proc. SFDS'03, XXXV ième Journées de Statistiques*, Volume 2, pp. 985–988.
- Totohasina, A., H. Ralambondrainy, et J. Diatta (2005). Une vision unificatrice des mesures probabilistes de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicite. In *Proc. of TAIMA'05*, pp. 375–380.
- Zhang, T. (2000). Association rules. In *PAKDD 2000*, LNAI 1805, pp. 245–256. Springer-Verlag.

Summary

This paper is concerned with probabilistic quality measures. We give a necessary and sufficient condition for a probabilistic quality measure to admit an associated normalized one. Moreover, we consider about thirty probabilistic quality measures proposed in the literature and show that most of them have the same associated normalized one.

Aide à l'interprétation des règles d'association composées.

Martine Cadot* , Pascal Cuxac** , Claire François **

* UHP/LORIA, Département Informatique, BP239, 54506 Vandoeuvre-lès-Nancy cedex
martine.cadot@loria.fr

<http://www.loria.fr/~cadot/>

**INIST-CNRS, 2 allée du Parc de Brabois, 54154 Vandoeuvre-lès-Nancy cedex
pascal.cuxac@inist.fr ; claire.francois@inist.fr

Résumé. L'extraction des règles d'association (RA) est une méthode qui est apparue pour les données type « tickets de caisse ». La création de nombreux indices de qualité a permis sa généralisation à d'autres types de données (Guillet 2004). Nous nous intéressons ici au problème de l'expert qui se trouve confronté à un nombre important de règles pas toujours faciles à interpréter. Les règles formées seulement de deux propriétés, une en partie gauche et une en partie droite s'interprètent aisément une fois l'indice de qualité choisi. Dans le cas de règles composées, c'est-à-dire comportant plus de deux propriétés, ces indices ne suffisent pas à aider l'expert à interpréter le lien entre ces propriétés. Nous proposons un modèle qui permet d'évaluer le gain d'information apporté par les règles de type $AB \rightarrow C$ et de sélectionner pour l'expert celles qui ajoutent du sens aux règles simples $A \rightarrow C$ et $B \rightarrow C$. L'application de cette méthodologie dans le cadre d'une analyse d'un corpus de textes par classification montre l'aide apportée à l'expert pour l'interprétation de cette classification. Pour faciliter l'exposé, le gain d'information que nous définissons a été appliqué à des règles formées de 3 propriétés, mais il est défini pour un nombre quelconque de propriétés.

1 Introduction

Les RA ont été créées pour extraire de la connaissance à partir de données. A et B étant deux propriétés binaires, la règle d'association est un lien entre A et B noté " $A \rightarrow B$ " ou "si A alors B". Sa définition varie selon les trois principaux courants initiés par les auteurs suivants : Gras (1979) définit des règles d'implication statistique pour aider les didacticiens à trouver des relations entre les acquisitions de notions élémentaires chez les élèves d'une classe, Guigues et Duquenne (1986) se sont plutôt intéressés à une représentation ordonnée de concepts avec les implications informatives, Agrawal et Srikant (1996) ont privilégié l'extraction optimisée de règles d'association dans de grandes bases de données.

Par la suite, ces formes ont connu des extensions dans plusieurs directions. La binarité des propriétés n'est plus obligatoire, on peut maintenant faire des RA avec des propriétés numériques (Guillaume 2000, Cadot et al. 2004b). Pour éviter l'explosion du temps d'extraction des règles, due à celle de la capacité de stockage des données, des algorithmes plus performants ont été proposés (Pasquier 2000). La sémantique des règles a été affinée grâce à de

Aide à l'interprétation des règles d'association composées.

nombreux indices de qualité (Guillet 2004), ce qui aide l'utilisateur à choisir les règles les plus adaptées à ses besoins. La navigation ainsi que l'interrogation par un langage adapté ont été mises au point (Botta et al. 2002) pour faciliter l'exploration de cet ensemble de règles.

Toutefois, à notre connaissance, peu de chercheurs en fouille de données se sont intéressés au problème d'interprétation que pose la prémisse composée d'une règle. Les contributions dans ce sens se sont concentrées sur la résolution du problème par élagage des règles qui apportent de l'incohérence au jeu de règles (Cadot et al. 2004a, Zhu H. 1998). Dans cet article, nous nous concentrons essentiellement sur l'ensemble des trois règles $A \rightarrow C$, $B \rightarrow C$ et $AB \rightarrow C$, et nous convenons de garder les deux règles à prémisse simple si elles ont été sélectionnées grâce à un choix d'indices approprié, et de ne garder la règle à prémisse composée que si elle renforce les deux autres, ce renforcement étant mesuré par un indice de gain.

Le but de cet indice de gain est d'aider l'expert à analyser les résultats issus du processus de traitement des données.

2 Le gain d'information

On ne s'intéresse dans cet article qu'aux RA extraites de tableaux numériques de type « SujetsXPropriétés » comportant les valeurs des sujets aux propriétés. Les définitions sont données pour des propriétés binaires et sont ensuite étendues à des propriétés numériques. La qualité d'une règle $A \rightarrow B$ est mesurée par de nombreux indices dont les plus courants sont le *support*, qui est le nombre d'objets vérifiant les propriétés de A et de B, c'est-à-dire de leur conjonction, appelée *motif* AB, et la *confiance*, qui est le quotient de ce support et du nombre d'objets vérifiant les propriétés de A, c'est-à-dire du support de A.

L'indice de qualité que nous proposons est construit seulement sur les règles à prémisse composée. Il mesure l'apport d'une telle règle par rapport aux règles simples qui la composent. Les règles simples ne permettent pas de calculer le support de la règle composée, mais seulement l'intervalle de ses variations. Le gain d'information mesure l'importance de l'écart entre le support observé et le milieu de l'intervalle de variation qui représente une situation d'équilibre. Notre méthode est à rapprocher de l'IPEE de Blanchard et al (2005) qui mesure l'écart à l'équilibre entre la partie gauche et la partie droite de la règle. Mais notre équilibre concerne toutes les propriétés constituant la règle. Freitas (1999) a également défini un gain d'information qu'il utilise dans son indice « Attribute surprisingness » portant sur des règles de discrimination (où le membre droit de la règle est une propriété déterminée à l'avance). Son objectif est de mesurer l'apport individuel de chaque propriété à la discrimination, alors que nous mesurons le gain d'information apporté par une propriété supplémentaire.

2.1 Définition du gain

Le principe. Pour mesurer le gain d'information d'une règle, nous nous appuyons sur les variations possibles du support du motif M obtenu en réunissant les propriétés des parties gauche et droite de cette règle. On impose à ces variations de se faire en laissant les supports des sous-motifs de M inchangés. Ainsi ce gain mesure ce que l'association de toutes les propriétés le composant apporte de plus que l'ensemble des diverses associations d'une partie de ces propriétés.

Recherche de l'intervalle de variation. Pour obtenir que le support de M augmente d'une unité, on choisit un sujet qui vérifie toutes les propriétés sauf une, et on lui rajoute cette pro-

priété. Cela a pour conséquence que le support de chaque sous-motif de M contenant cette propriété est également augmenté de une unité. On compense cette augmentation en faisant de nouveaux changements élémentaires, qui vont également devoir être compensés. Si ce processus peut se réaliser, il s'arrête nécessairement au bout de $2^{(L-1)}$ changements, L étant la longueur du motif M , c'est-à-dire son nombre de propriétés. Et le support de M peut augmenter d'autant d'unités que de répétitions possibles de ce processus. Pour le faire diminuer, on procède pareillement en faisant les changements inverses de sujets. On crée ainsi l'intervalle de variation du support de M .

Choix de la valeur du gain. L'intervalle obtenu a un centre à partir duquel le support des motifs peut augmenter ou diminuer. Nous décidons que le gain d'information correspondant aux motifs de support central est nul, et cela reste valable si l'intervalle est réduit à une valeur. Puis plus le support du motif s'éloigne de ce centre en se rapprochant des bornes de l'intervalle, plus la valeur absolue du gain augmente. Selon que le support est à droite du centre ou à gauche, le gain est positif ou négatif. Nous mesurons la valeur de ce gain en nombre de sujets dont on doit changer la valeur d'une propriété pour obtenir ce motif en partant d'un motif de support central.

La formule. Pour mesurer le gain d'information g d'une règle, nous calculons le support s du motif M sur lequel elle s'appuie, la longueur L de ce motif (le nombre de propriétés le constituant), et le centre c de l'intervalle décrit par le support de M . Pour valeur de gain, nous choisissons la fonction $g=2^{(L-1)}*(s-c)$.

2.2 Propriétés du gain d'un motif

La condition préalable au calcul du gain de motifs est qu'ils soient construits avec des propriétés dont on connaît les valeurs pour chacun des N sujets d'un ensemble donné. Bien qu'on utilise par la suite le gain pour des règles à prémisse composée, donc construites sur des motifs de longueur au moins égale à 3, le principe de calcul du gain s'étend sans problème à des motifs de longueur 2 et la formule à des motifs de longueur 1 (notons toutefois que dans ce dernier cas $g=s-N/2$, ce qui fait qu'on peut obtenir un gain avec des moitiés de sujets).

- prop 1 : Le gain d'un motif M ne peut pas dépasser $N/2$ en valeur absolue.
- prop 2 : le gain d'un motif de longueur L est un nombre entier de fois $2^{(L-2)}$
- prop 3 : Si a est l'amplitude de l'intervalle de variation du gain g d'un motif M de longueur L , l'intervalle de variation du gain de ses sur-motifs de longueur $L+1$ a une amplitude inférieure ou égale à $a-2|g|$. La valeur de a pour les motifs de longueur 1 est N , l'intervalle étant $[-N/2 ; N/2]$.

Ces trois propriétés permettent de limiter le coût machine de la recherche du gain d'un motif. Avec les propriétés 1 et 2, dès que sa longueur est telle que $2^{(L-2)}$ dépasse $N/2$ (soit $L > 1 + \log(N)/\log(2)$), le gain est nul. Avec la propriété 3, chaque fois que le gain d'un motif de longueur L est différent de 0, cela réduit l'intervalle de variation des motifs qui le contiennent. Ainsi, au fur et à mesure que la longueur du motif augmente par ajout de propriétés, ses possibilités de variation diminuent ou restent constantes, ce qui limite sa valeur possible de gain. Cet effet est accentué par la propriété 2. Cela est en adéquation avec le fait que dans le cas le plus courant, une fois que l'information essentielle est apportée par quelques propriétés, au fur et à mesure qu'on ajoute de nouvelles propriétés, l'information supplémentaire qui en résulte est de plus en plus petite.

Aide à l'interprétation des règles d'association composées.

2.3 Relation du gain avec les indices de qualité des règles

Le gain de la règle est celui du motif sur lequel elle est construite. Il fait ainsi partie des indices de qualité d'une règle au même titre que le support, la confiance et tous ceux qu'on définit habituellement (Guillet 2004). Toutefois, il ne mesure pas comme les autres indices la qualité intrinsèque d'une règle, mais la valeur additionnelle d'une règle avec prémisse composée par rapport à celles avec prémisses plus simples. Dans l'application que nous en faisons, nous extrayons d'abord les motifs de longueur quelconque ayant un support suffisant, puis les règles ayant une confiance suffisante construites sur des motifs de longueur 2. L'utilisation d'un seuil de support pour l'extraction des motifs se justifie par le besoin de généraliser les règles obtenues. Le choix de la confiance pour les règles sur des motifs de longueur 2 a été justifié a posteriori par l'interprétation satisfaisante des règles obtenues. Le gain permet de sélectionner les règles de longueur 3 qui renforcent les précédentes.

2.4 Le gain des règles d'association floues

Nous avons défini précédemment des supports flous et des RA floues sur des propriétés numériques (Cadot et al. 2004b). Les supports des motifs flous restent positifs, mais leurs valeurs peuvent ne pas être entières. L'intervalle de variation du support d'un motif ne peut plus se construire en déplaçant des sujets entre les 2^L parties délimitées par les valeurs aux propriétés, car leur appartenance à une partie est floue. Pas plus qu'il ne peut se construire en remplaçant la valeur à une propriété par son complément à 1, celui-ci n'ayant de sens qu'en calcul binaire. Toutefois, comme les effectifs de chacune des 2^L parties étaient calculés précédemment à partir des supports des motifs, le calcul peut toujours se faire pareillement, la différence est que les valeurs obtenues ne sont plus entières, ce sont des effectifs flous. Cela ne gêne aucunement le calcul du gain dont il importe peu qu'il soit ou non entier. Nous avons trois propriétés permettant de limiter le coût machine du calcul du gain. La deuxième propriété due au caractère entier des effectifs disparaît, mais les deux autres restent, et il en résulte une perte d'efficacité de l'algorithme du calcul du gain sur un ensemble de motifs flous. Cette perte est compensée par le choix d'un codage flou plutôt que binaire qui permet de ne pas multiplier les propriétés habituellement binarisées avec plusieurs seuils.

Nous avons ainsi défini un gain pour les règles d'association floue qui prolonge celui que nous venons de définir pour les RA classiques.

3 Application

Notre objectif est d'appliquer des RA floues sur des résultats de classifications, afin d'aider l'expert à analyser les résultats. Les règles à prémisses composées permettent de visualiser les classes qui peuvent fusionner entre deux classifications différentes ou au sein de la même classification.

3.1 Méthode de classification

Les classifications sur lesquelles nous travaillons ont été obtenues en utilisant la plateforme Stanalyst® (Polanco et al. 2001) qui permet de traiter des corpus bibliographiques et inclut la méthode des K-means Axiales comme méthode de classification (Lelu 1993).

Cette méthode est basée sur le principe de classification par centres mobiles, plus connue sous le nom de K-means (Forgy 1965), mais elle réalise une analyse factorielle sphérique sur chaque classe ; les classes sont donc matérialisées par des demi-axes représentatifs des éléments. L'utilisation de ces axes permet de quantifier l'appartenance d'un élément à une classe. De plus, au lieu d'affecter l'élément à la seule classe où sa valeur est la plus grande, on l'affecte également aux classes pour lesquelles cette valeur dépasse un certain seuil. Cet algorithme, paramétré par le nombre maximal de classes désiré et le seuil des coordonnées des éléments et descripteurs sur les axes, permet donc de construire des classes recouvrantes où les individus et descripteurs (documents et mots-clés) sont ordonnés selon un degré de ressemblance au type idéal de la classe.

Le corpus traité est constitué de 3203 notices bibliographiques extraites de la base PASCAL sur le thème de la géotechnique, publiées en 2001 et 2002 et indexées manuellement. Nous avons calculé quatre classifications avec la méthode des K-means axiales en paramétrant 20, 30, 40, 50 classes. Dans la suite de l'article elles sont nommées respectivement C20, C30, C40, C50.

3.2 Règles obtenues

Si nous calculons toutes les RA à prémisse composée nous avons 1548 règles. Le gain calculé permet de filtrer ces résultats : la variation du gain de 5 à 30 permet de passer de 90% (1395) de règles à 7% (105 règles). Pour faciliter l'analyse des résultats nous ne considérerons dans ce qui suit que les règles où les deux membres de la prémisse appartiennent à la même classification. Avec un gain supérieur à 30 on a les 12 règles résumées dans le tableau suivant (S :support, C :confiance, G : Gain) :

N°	Règle	S	C	G
R1	C20 Barrage, C20 Eau souterraine → C40 Pollution	216,3	0,76	30,56
R2	C20 Inélasticité, C20 Relation $\sigma \varepsilon$ (σ : contrainte, ε : déformation) → C50 Mécanique rupture	25,19	0,52	42,16
R3	C20 Inélasticité, C20 Relation $\sigma \varepsilon$ → C30 Résistance compression	20,7	0,43	31,66
R4	C30 Barrage, C30 Eau souterraine → C40 Pollution	16,13	0,73	30,2
R5	C30 Relation $\sigma \varepsilon$, C30 Essai sol → C20 Résistance cisaillement	21,37	0,88	37,08
R6	C30 Résistance compression, C30 Mécanique rupture → C20 Inélasticité	18,42	0,91	33,12
R7	C40 Essai sol, C40 Relation $\sigma \varepsilon$ → C20 Résistance cisaillement	19,83	0,86	33,06
R8	C50 Mécanique rupture, C50 Relation $\sigma \varepsilon$ → C20 Inélasticité	19,63	0,91	35,32
R9	C50 Conductivité hydraulique, C50 Pollution → C20 Eau souterraine	23,73	0,97	46,22
R10	C50 Conductivité hydraulique, C50 Pollution → C30 Eau souterraine	23,66	0,97	45,94
R11	C50 Pression pores, C50 Champ pétrole → C20 Inélasticité	16,83	0,91	30,04
R12	C50 Relation $\sigma \varepsilon$, C50 Essai sol → C20 Résistance cisaillement	18,17	0,91	32,54

TAB. 1 – Règles avec prémisse composée de deux classes de la même classification et avec $g > 30$ (S=support, C=confiance, G=gain).

Analysons par exemple la règle R11, constituée des règles simples suivantes :

C50 Pression Pores → C20 Inélasticité
C50 Champ pétrole → C20 Inélasticité

Aide à l'interprétation des règles d'association composées.

A première vue l'intitulé "Champ pétrole" peut paraître surprenant. L'analyse des données qui sont regroupées dans ces classes (titre des articles, résumés, indexation) permet de comprendre cette règle et de la valider. En effet la classe "Champ pétrole" est essentiellement consacrée aux roches magasins et aux distributions des contraintes dans ces roches. La classe "Inélasticité" est dominée par des aspects liés à l'élastoplasticité et à l'analyse des champs de contraintes. Cette règle est alors plus lisible puisqu'elle lie des articles parlant de la pression de pores (donc de roches poreuses plus ou moins saturées) et des articles sur la distribution des contraintes dans des roches magasin (roches poreuses plus ou moins saturées) avec des articles sur les champs de contraintes dans le domaine élastoplastique.

Cependant, certaines de ces règles, comme par exemple R1, sont difficilement interprétables. Cela est peut-être dû au fait que le gain est gonflé artificiellement par des effets d'effectifs. Ce phénomène est bien connu en statistique et nécessite qu'on sélectionne non seulement les effets les plus importants mais aussi les plus significatifs. En effet un score élevé peut être dû au hasard. La construction d'un test permettant d'établir la significativité du gain est en cours afin de les éliminer.

L'application de cette méthodologie nous a donc permis de "filtrer" les règles à prémisses composées pour ne garder que celles porteuses de valeur ajoutée.

4 Conclusion et perspectives

Le gain que nous proposons combine les avantages des indices de qualité des RA, et de l'élagage du jeu de RA. Il garde les règles simples, construites sur deux propriétés qui ont été extraites à l'aide d'un indice de qualité choisi pour sa valeur sémantique, et sont donc aisément interprétables. Les autres règles, qui ne sont gardées que si leur gain est suffisant, voire significatif renforcent l'information tirée des premières. Au final, l'ensemble des règles obtenu est de taille réduite et sans incohérence. Le gain proposé s'étend sans problème aux RA quantitatives codées de façon floue.

Nous avons vu que l'efficacité du gain doit être renforcée par un test qui en assure la significativité. D'autre part, le gain que nous avons utilisé mesure la valeur ajoutée de la règle $AB \rightarrow C$ par rapport aux règles $A \rightarrow C$ et $B \rightarrow C$ en fixant tous les sous-motifs de ABC. Il faudrait peut-être autoriser la variation du motif AB.

Références

- Agrawal, R. Srikant, H. (1994) *Fast algorithms for mining association rules in large databases*, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.
- Blanchard, J., Guillet, F., Briand, H., Gras, R. (2005) IPEE : Indice Probabiliste d'Ecart à l'Equilibre. pour l'évaluation de la qualité des règles. DKQ 2005 Paris, Atelier EGC 2005 pp. 26-34
- Botta, M., Boulicaut J.-F., Masson C., Meo R. (2002). A Comparison between Query Languages for the Extraction of Association Rules. *DaWaK 2002*, p. 1-10

- Cadot, M., di Martino, J., Napoli, A. (2004a). Réduction d'un jeu de RA par des méta-règles issues de la logique de "sens commun". *EGC'2004*. (Clermont-Ferrand, France). RNTI, 2004. p.353.
- Cadot, M., Napoli, A. (2004b) RA et codage flou des données. *SFC'04*. (Bordeaux). p.130-133.
- Forgy E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications, *Biometrics*, vol. 21, n° 3, p. 768.
- Freitas, A.A. (1999). On rule interestingness measures. *Knowledge-Based Systems* 12, p. 309-315.
- Guigues J.L., Duquenne V. (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. Hum.* n°95, pp. 5-18
- Guillaume S. (2000) *Traitement des données volumineuses, mesures et algorithmes d'extraction de RA et règles ordinales*, Thèse Nantes, 2000.
- Guillet F. (2004) Mesure de qualité des connaissances en ECD, *Cours donné lors des journées de la conférence EGC 2004*, Clermont-ferrand, 20 janvier 2004.
- Gras R., (1979) *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse, Rennes I, 1979.
- Lelu A. (1993). *Modèles neuronaux pour l'analyse de données documentaires et textuelles*. Paris, Thèse de l'Université de Paris VI, 238 pages.
- Morineau, A., Nakache, J.-P., Krzyzanowski, C. (1996) *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris 1996.
- Pasquier N. (2000), *Data Mining : Algorithmes d'Extraction et de Réduction des RA dans les Bases de Données*, Thèse, Clermont-Ferrand II, 2000
- Polanco X., François C., Royauté J., Besagni D (2001). STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology, *CSCI 2001*, Sydney, Australia, Proceedings Vol 2, pp. 871 – 873.
- Zhu H., (1998) *On-Line Analytical Mining of Association Rules*, Thèse, Simon Fraser University, 1998

Summary

When a simple rule of type $A \rightarrow C$ or $B \rightarrow C$ becomes a rule of type $AB \rightarrow C$, the two premises A and B are a new premise AB. The fusion of these premises in only one is not a problem in binary logic, which is the foundation of the association rules, but the interpretation of the rule obtained is a semantic problem. We propose a model to evaluate the profit of information brought by rule $AB \rightarrow C$ and to select for the expert those which reinforce sufficiently the semantics of simple rules $A \rightarrow C$ and $B \rightarrow C$. The application of this methodology to the analysis of a corpus by classification enables us to help the expert to interpret this classification.

Agrégation de mesures d'intérêt de règles d'association

Jean-Pierre Barthélemy*, Angélique Legrain*
Philippe Lenca*, Benoît Vaillant**,*

*GET – ENST Bretagne – Laboratoire TAMCIC, UMR 2872 CNRS
prenom.nom@enst-bretagne.fr
**IUT de Vannes, Département STID
benoit.vaillant@univ-ubs.fr

Résumé. L'un des principaux problèmes posés par l'extraction de règles d'association est l'évaluation de la qualité des règles produites par les algorithmes de type APRIORI. De nombreuses mesures ont été définies afin de pouvoir classer les règles dites *intéressantes*. Très hétérogènes, elles produisent des classements forts variés. C'est pourquoi, plutôt que de privilégier une mesure il paraît intéressant de tenir compte des différentes informations apportées par les mesures. Ainsi, nous avons adopté une nouvelle approche : l'agrégation à l'aide de relations valuées permettant de mesurer le degré d'intensité de préférence d'une règle sur une autre. Elles permettent d'une part de retranscrire la nature numérique des mesures, et d'autre part de réduire les problèmes d'incomparabilité entre les mesures.

Nous avons étudié différents opérateurs d'agrégation. Dans cet article, nous illustrons les résultats obtenus à l'aide d'un exemple jouet en utilisant le plus simple opérateur d'agrégation.

1 Considérations basiques

1.1 Règles d'association

Dans ce papier, nous nous restreignons au cas où les données sont des objets décrits par q attributs binaires : l'objet i satisfait la propriété x (codée en 1), ou non (codée en 0). On note $N = \{1, \dots, n\}$ l'ensemble des objets et $Q = \{a, b, \dots\}$ l'ensemble des propriétés.

Nous considérons des règles d'association $A \rightarrow B$ telles que définies par Agrawal et al. (1993) : si un sous-ensemble de N possède significativement les caractéristiques de l'ensemble $A \subseteq Q$, alors il possédera significativement les caractéristiques de l'ensemble $B \subseteq Q$. Une règle d'association est un 2-uplet (A, B) de sous-ensembles de Q tels que $A \cap B = \emptyset$.

Les algorithmes de type APRIORI (Agrawal et al., 1993) permettant de calculer les règles d'association produisent en général un nombre considérable de règles. Il est ainsi nécessaire de filtrer les règles en fonction de leur intérêt. Nous considérons dans cette étude les mesures d'intérêt dites objectives et basées sur des opérations de comptage dans les données, *i.e.* sur les grandeurs n , n_A , n_B et n_{AB} , où n_A (resp. n_B , n_{AB}) représente le nombre d'objets vérifiant toutes les propriétés de A (resp. B , $A \cup B$), selon les notations classiques.

1.2 Mesures de qualité

Le support, $p_{AB} = n_{AB}/n$, et la confiance, $p_{AB}/p_A = n_{AB}/n_A$ sont utilisés comme premier filtre pour extraire un ensemble \mathcal{R} de règles avec les algorithmes de type APRIORI. Il est nécessaire d'utiliser dans un second temps d'autres mesures de qualité. En effet, on considère que le support et la confiance n'ont que peu de bonnes propriétés pour ordonner un ensemble de règles si on les compare à d'autres mesures (Piatetsky-Shapiro, 1991; Tan et al., 2002; Lenca et al., 2003). Nous avons sélectionné et étudié une vingtaine de mesures (Lenca et al., 2003). Dans Lenca et al. (2004), nous proposons une aide à la sélection d'un ensemble de mesures selon les données et les attentes de l'utilisateur. On met en évidence que le choix d'une mesure dépend particulièrement des attentes de ce dernier.

Les mesures constituent un paysage très hétérogène : on peut observer d'importantes variations entre les formules (les mesures ne traduisent pas les mêmes caractéristiques des règles), et de grandes différences dans les co-domaines ($[0, 1]$, $[0, +\infty[$, $]-\infty, 1]$, bornes fonction de n_A, n_B , et/ou $n_{AB} \dots$). Ainsi certaines règles peuvent être très bien classées par une mesure, et mal classées par une autre. La comparaison des préordres totaux induits par les mesures sur une base de règles permet de mettre en évidence cette observation (Vaillant et al., 2004).

1.3 L'agrégation pour faire face à l'hétérogénéité des mesures

Il se pose ainsi naturellement la question suivante : quelle est ou quelles sont les meilleures règles étant donné nos mesures de qualité et les données à traiter ? On peut tenter d'y répondre selon deux voies principales :

- (i) position dictatoriale : choisir une mesure préférée et ne pas tenir compte des autres
- (ii) position consensuelle : trouver un consensus entre les mesures.

Dans cet article, nous suivons cette seconde piste pour laquelle trois voies apparaissent naturellement :

- a) l'agrégation directe des mesures en une seule, en utilisant une sorte de moyenne généralisée. Une difficulté est alors posée par la diversité des échelles de mesures. Comment agréger des mesures dont les co-domaines sont $[0,1]$, $[0,+\infty[$, $]-\infty,1]$?
- b) l'agrégation des rangs induits par les différentes mesures en un rang unique. Les classements ne tiennent alors pas compte des différences dans les évaluations. De plus, une agrégation ordinale implique des problèmes "logiques" périlleux (se référer par exemple au théorème d'Arrow (Arrow, 1951))
- c) l'agrégation de relations valuées. C'est la voie que nous explorons dans cette étude. Les classements sont des préordres totaux que nous généralisons sous forme de relations valuées.

Pour cela, nous rappelons qu'une relation valuée (parfois appelée " relation floue ") sur un ensemble S est une transformation R de $S \times S$ dans l'intervalle unité $[0,1]$. Les relations valuées permettent d'échapper aux effets d'échelles en préservant les différences d'échelles puisqu'elles tiennent compte de telles différences. L'agrégation de relations valuées a été étudiée très précisément par Fodor et Roubens (1994). Elles peuvent être affectées de nombreuses propriétés, parmi lesquelles nous retenons quelques formes particulières de transitivité.

2 Construction d'une relation valuée sur un ensemble de mesures

L'un des avantages de la modélisation de relations valuées est la retranscription des évaluations numériques qu'elle permet. Pour définir des relations valuées il va donc falloir poser certaines conditions. Pour une mesure μ et deux règles r_i et $r_{i'}$, on dira par exemple que $R(r_i, r_{i'})$ est négligeable si $\mu(r_i)$ est faiblement supérieur à $\mu(r_{i'})$. Cette faible différence peut être définie de deux façons : soit par le quotient des évaluations, soit par leur différence.

Nous utilisons trois types de transitivité, la transitivité faible ($\{R(s, t) \geq \frac{1}{2} \text{ et } R(t, u) \geq \frac{1}{2}\} \Rightarrow R(s, u) \geq \frac{1}{2}$), la min-transitivité ($R(s, u) \geq \min\{R(s, t), R(t, u)\}$) et la max- Δ -transitivité ($R(s, u) \geq \max\{0, R(s, t) + R(t, u) - 1\}$) car elles permettent la préservation de certaines propriétés après agrégation.

Soit $\mathcal{R} = \{r_1, \dots, r_k\}$ un ensemble de règles, et μ_1, \dots, μ_m les mesures sélectionnées. Chaque mesure μ_j induit une relation valuée R_j sur \mathcal{R} . L'idée générale est que $R_j(r_i, r_{i'})$ correspond à une différence normalisée entre les valeurs prises par la mesure μ_j sur les règles $r_i, r_{i'}$ et doit permettre de modéliser un système de préférences sur l'ensemble des règles.

Nous présentons ci-dessous, une des relations valuées que nous avons étudiée et proposée par Brans et Mareschal (1994). C'est une variante de la différence linéaire permettant de lisser les transitions entre "non préférence" et "faible préférence" ainsi qu'entre la "préférence faible" et la "préférence forte" (le paramètre σ_j représente un seuil entre les "préférences faibles" et les "préférences fortes" –point d'inflexion de la courbe) :

$$R_j(r_i, r_{i'}) = \begin{cases} 1 - \exp\left(-\frac{(\mu_j(r_i) - \mu_j(r_{i'}))^2}{2\sigma_j^2}\right) & \text{si } \mu_j(r_i) - \mu_j(r_{i'}) > 0 \\ 0 & \text{sinon} \end{cases} \quad (1)$$

3 Agrégation

3.1 Généralités

Un opérateur d'agrégation est une fonction C de $\cup_{m \geq 1} [0, 1]^m$ dans $[0, 1]$ non décroissante de chaque composant, idempotent, et qui satisfait $C(0, \dots, 0) = 0$ et $C(1, \dots, 1) = 1$. Si R^* est un m -uplet (R_1, \dots, R_m) de relations valuées, un opérateur d'agrégation C va produire une relation de consensus de R^* , notée $C(R^*) : C(R^*)(r_i, r_{i'}) = C(R_1(r_i, r_{i'}), \dots, R_m(r_i, r_{i'}))$.

De nombreux opérateurs d'agrégation aux propriétés différentes ont été étudiés dans la littérature (cf. Fodor et Roubens (1994)) : moyennes généralisées, opérateurs OWA, intégrales de Choquet et de Sugeno, maximum et minimums pondérés, etc. Nous avons choisi de nous concentrer sur les plus simples d'entre-eux, les moyennes généralisées :

$$M(u_1, \dots, u_m) = f^{-1}\left(\sum_{1 \leq j \leq m} w_j f(u_j)\right)$$

où f est une fonction monotone continue, f^{-1} sa réciproque, et les w_j des poids non négatifs. Ainsi on distingue certains cas particuliers : la moyenne arithmétique pondérée (WMean, $f(u) = u$) la moyenne géométrique pondérée (WGeom, $f(u) = \log(u)$) la moyenne d'ordre α (RPM, $f(u) = u^\alpha, \alpha \in \mathbb{R}^*$), et la moyenne harmonique pondérée (WHarm, $f(u) = 1/u$). Dans cet article, nous présentons des résultats obtenus avec RPM et WMean.

3.2 Comportement vis à vis de la transitivité et de certaines propriétés

Saminger et al. (2002); Peneva et Popchev (2003) ont étudié les propriétés préservées par ces opérateurs. Seules celles qui jouent un rôle dans la modélisation des préférences nous intéressent dans le cadre de ce travail. Nous avons par conséquent éliminé différentes formes de dissimilarités. Nous nous concentrons principalement sur la transitivité, qui est l'une des propriétés qui se rapproche le plus des attentes de l'utilisateur. Elle garantit que si un ensemble de règles est mieux évalué qu'un autre, alors cet ordre sera préservé par la procédure d'agrégation.

4 Résultats expérimentaux

Le tableau 1 contient le rangement de 21 règles pour les 20 mesures que nous avons étudiées (les notations utilisées sont des abréviations du nom des mesures, SUP pour le support, IIE pour l'intensité d'implication entropique, etc., voir Lenca et al. (2004)). Les données "jouet" et le mode de calcul de ces règles sont présentés dans (Barthélemy et al., 2006), ainsi que les évaluations numériques des règles pour chaque mesure.

La relation évaluée retenue pour exprimer la préférence individuelle d'une règle sur une autre pour une mesure donnée nécessite de fixer le paramètre σ_j (cf. formule 1). Pour ce faire, nous avons choisi de tenir compte des différences d'évaluations, et utilisons la valeur pour un quantile donné. Par exemple, la valeur prise à un quantile de 0% correspond à la plus petite valeur absolue de la différence (qui est évidemment 0, puisque la différence des évaluations d'une règle sur elle-même est nulle). La valeur prise par un quantile de 100% est la plus grande valeur absolue de la différence d'évaluation entre toute paire de règles et un quantile de 50% amène à la valeur médiane de l'absolue des différences. Dans notre exemple, le quantile est fixé à 60% de toutes les différences absolues, et c'est donc une valeur élevée. Nous expliquons plus loin pourquoi nous avons choisi une telle valeur. Les poids choisis sont tous égaux à 1/20. D'autres poids ont été proposés dans (Legrain, 2004) grâce à une modélisation de préférences d'utilisateurs (Lenca et al., 2004).

Une fois ces paramètres fixés, la procédure d'agrégation produit une matrice carrée de valeurs entre 0 et 1, chaque valeur représentant la préférence agrégée d'une règle sur une autre. Pour obtenir des résultats lisibles (i.e. binaires), on utilise un seuil de coupe λ et on compare l'indice d'agrégation à λ . Les valeurs plus faibles que λ sont considérées comme égales à 0 et les valeurs supérieures égales à 1. On représente alors les préférences par des graphes (cf. figures 1 et 2, pour RPM et WMean) où les arcs entre deux sommets représentent la préférence d'une règle sur une autre, la flèche pointant sur la règle qui est préférée. On considère qu'aucune règle n'est préférée à elle-même. Il apparaît que les règles qui sont bien classées par toutes les mesures (comme r_{18}) restent bien classées, et celles qui étaient mal classées (comme r_{19}) restent mal classées. Le cas des règles plus controversées, comme r_2 nécessite d'approfondir la méthode. Notons que WHarm et WGeom produisent un effet de veto (Barthélemy et al., 2006). En effet, quand une mesure a la même valeur pour deux règles, l'utilisation du logarithme et de la fonction inverse produit des valeurs "infinies", ce qui fait que toute autre différence n'est pas prise en compte. Ceci explique qu'un grand nombre de règles resteront incomparables en terme d'agrégation de préférences avec de tels agrégateurs. Le fait que SUP divise les règles en seulement deux groupes joue aussi un rôle sur la valeur que nous avons choisie pour fixer σ_j : si nous avons choisi une valeur plus faible que 60%, alors la valeur

TAB. 1 – Ordonnancement induit par les mesures sur une base de règles

mesure	rank 1	rank 2	rank 3	rank 4	rank 5	rank 6	rank 7	rank 8	rank 9	rank 10	rank 11	
LIFT	r_{18}	$r_1 r_2 r_{17}$	r_{21}	$r_6 r_7$	$r_3 r_{16} r_{20}$	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$			
CONFGEN	$r_1 r_{17}$	r_{18}	r_2	r_{21}	$r_3 r_6 r_{16}$	r_7	r_{20}	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$	
CONF	$r_1 r_3 r_{16} r_{17}$	$r_6 r_8 r_{11}$	$r_{18} r_{19} r_{20} r_{21}$	$r_5 r_7 r_{12} r_{13} r_{14} r_{15}$	$r_2 r_4 r_9 r_{10}$							
SUP	$r_5 r_6 r_7 r_8 r_{11} r_{12} r_{13} r_{14} r_{15}$	$r_1 r_2 r_3 r_4 r_9 r_{10} r_{16} r_{17} r_{18} r_{19} r_{20} r_{21}$										
IIE	$r_1 r_{17}$	$r_3 r_{16}$	r_{18}	$r_6 r_{21}$	r_{20}	r_2	r_7	$r_4 r_5 r_8 r_9 r_{10} r_{11} r_{12} r_{13} r_{14} r_{15} r_{19}$				
INTIMP	$r_1 r_{17}$	r_{18}	r_2	$r_3 r_{16}$	$r_6 r_{21}$	r_7	r_{20}	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$	
IQC	r_{18}	$r_1 r_2 r_{17}$	$r_6 r_7 r_{21}$	r_{20}	$r_3 r_{16}$	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$			
CONV	$r_1 r_3 r_{16} r_{17}$	r_{18}	$r_6 r_{21}$	r_2	r_{20}	r_7	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$		
GI	r_{18}	$r_1 r_2 r_{17}$	r_{21}	$r_6 r_7$	$r_3 r_{16} r_{20}$	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$			
-INDIMP	$r_1 r_{17}$	r_{18}	r_2	$r_3 r_{16}$	$r_6 r_{21}$	r_7	r_{20}	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$	
IPD	$r_1 r_{17}$	r_{18}	r_2	$r_3 r_{16}$	$r_6 r_{21}$	r_7	r_{20}	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$	
LAP	$r_1 r_3 r_{16} r_{17}$	$r_6 r_8 r_{11}$	$r_{18} r_{19} r_{20} r_{21}$	$r_5 r_7 r_{12} r_{13} r_{14} r_{15}$	$r_2 r_4 r_9 r_{10}$							
LOE	$r_1 r_3 r_{16} r_{17}$	r_{18}	$r_6 r_{21}$	r_2	r_{20}	r_7	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$		
MC	$r_1 r_3 r_{16} r_{17}$	r_{18}	r_2	r_{21}	r_6	$r_7 r_{20}$	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$		
PS	r_{18}	$r_1 r_2 r_{17}$	$r_6 r_7 r_{21}$	$r_3 r_{16} r_{20}$	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$				
R	r_{18}	$r_1 r_2 r_{17}$	$r_6 r_7 r_{21}$	$r_3 r_{16}$	r_{20}	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$			
SEB	$r_1 r_3 r_{16} r_{17}$	$r_6 r_8 r_{11}$	$r_{18} r_{19} r_{20} r_{21}$	$r_5 r_7 r_{12} r_{13} r_{14} r_{15}$	$r_2 r_4 r_9 r_{10}$							
MOCO	r_{18}	$r_1 r_{17}$	r_6	$r_7 r_{21}$	$r_3 r_8 r_{11} r_{16}$	$r_2 r_{12} r_{13} r_{20}$	$r_5 r_8 r_{14} r_{15} r_{19}$	$r_4 r_9 r_{10}$				
TEC	$r_1 r_3 r_{16} r_{17}$	$r_6 r_8 r_{11}$	$r_{18} r_{19} r_{20} r_{21}$	$r_5 r_7 r_{12} r_{13} r_{14} r_{15}$	$r_2 r_4 r_9 r_{10}$							
ZHANG	$r_1 r_3 r_{16} r_{17}$	r_{18}	r_2	r_{21}	r_6	$r_7 r_{20}$	$r_{12} r_{13}$	$r_4 r_8 r_9 r_{10} r_{11}$	r_{19}	$r_5 r_{14} r_{15}$		

du quantile aurait été nulle. Un moyen de palier à une telle situation pourrait être d’ajouter un léger degré de bruit dans les valeurs prises par les mesures. Cette stratégie, connue comme le “jittering”, est souvent utilisée dans des outils de visualisation, afin de représenter un nombre important de points, autrement confondus. La représentation graphique a pour but d’illustrer les résultats sur un exemple jouet. Les résultats obtenus sur une base de données réelles sont présentés dans (Legrain, 2004), mais les représentations graphiques nécessitent alors des méthodes plus élaborées. Cette possibilité a notamment été explorée en fouille visuelle de règles d’associations par Lehn et al. (1999) et Blanchard et al. (2003). Notons cependant que l’exploration visuelle n’est pas requise pour sélectionner les règles, c’est un moyen confortable et centré sur l’expert.

5 Conclusion et perspectives

Dans cette étude, nous nous sommes intéressés à la construction d’opérateurs d’agrégation de relations valuées définies à partir des évaluations des mesures d’intérêt de règles d’association. Les attentes de l’utilisateur contraignent les choix possibles. De telles contraintes peuvent

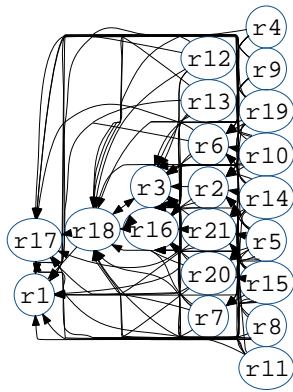


FIG. 1 – RPM ($\alpha = 2$)

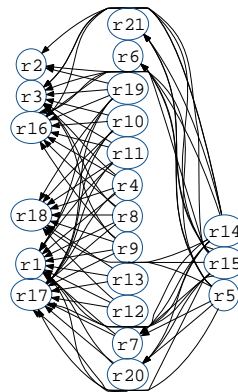


FIG. 2 – WMean

être explicitées mathématiquement. Nous nous sommes particulièrement intéressés à la propriété de transitivité. Parmi les opérateurs d'agrégation classiques, peu d'entre eux respectent de telles contraintes, excepté la moyenne généralisée. Un exemple jouet illustre notre approche. Des expériences prometteuses effectuées sur de plus grandes bases de données laissent entrevoir des résultats visuels intéressants. Bien qu'il persiste certains conflits entre le traitement de grandes bases de règles et leur représentation intelligible sous forme de graphe, nous pensons qu'une ouverture possible serait d'autoriser des zooms sur certaines régions significatives.

Une perspective d'étude intéressante est d'élargir le choix des opérateurs d'agrégation à des opérateurs de type compromis, comme l'intégrale de Choquet ou de Sugeno. Le classement final d'une règle pourrait aussi être consolidé si une coalition importante de mesures lui était favorable. Le choix du paramètre σ_j est déterminant. Nous l'avons fixé à l'aide d'une approche expérimentale basée sur une méthode de quantile. D'autres possibilités peuvent être envisagées, par exemple nous proposons de le fixer à l'aide d'un expert des données. Dans tous les cas, une étude de la robustesse de la solution proposée doit être menée.

Références

- Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Int. Conf. on Management of Data*, pp. 207–216.
- Arrow, K. J. (1951). *Social Choice and Individual Values*. Cowles Foundations and Wiley.
- Barthélemy, J. P., A. Legrain, P. Lenca, et B. Vaillant (2006). Aggregation of valued relations applied to association rule interestingness measures. In *MDAI'06 (To be published)*, LNAI. Springer-Verlag.
- Blanchard, J., F. Guillet, et H. Briand (2003). A user-driven and quality-oriented visualization for mining association rules. In *Third IEEE ICDM*, pp. 493–496.
- Brans, J. et B. Mareschal (1994). The PROMETHEE-GAIA decision support system for multi-criteria investigations. *Investigation Operativa* 4(2), 102–117.

- Fodor, J. et M. Roubens (1994). *Fuzzy preference modelling and multicriteria decision support*. Kluwer Academic Publishers.
- Legrain, A. (2004). Agrégation de mesures de qualité de règles d'association, Rapport de DEA MIASH. Master's thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne.
- Lehn, R., F. Guillet, P. Kuntz, H. Briand, et J. Philippé (1999). Felix : An interactive rule mining interface in a KDD process. In P. Lenca (Ed.), *HCP'99*, pp. 169–174.
- Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité en ECD. *RNTI (Entreposage et Fouille de données)* (1), 123–134.
- Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2004). A multicriteria decision aid for interestingness measure selection. Technical Report LUSSE-TR-2004-01-EN, ENST Bretagne.
- Lenca, P., P. Meyer, B. Vaillant, et P. Picouet (2003). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. *EGC 2003 1*(17), 271–282.
- Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2004). Évaluation et analyse multicritère des mesures de qualité des règles d'association. (RNTI-E-1), 219–246.
- Peneva, V. et I. Popchev (2003). Properties of the aggregation operators related with fuzzy relations. *Fuzzy Sets and Systems, numéro 139*, 615–633.
- Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro et W. Frawley (Eds.), *KDD*, pp. 229–248. AAAI/MIT Press.
- Saminger, S., R. Mesiar, et U. Bodenhofer (2002). Domination of aggregation operators and preservation of transitivity. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10*(Suppl.), 11–35.
- Tan, P.-N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Eighth ACM SIGKDD Int. Conf. on KDD*, pp. 32–41.
- Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. In *Discovery Science*, Volume 3245 of *LNAI*, pp. 290–297. Springer-Verlag.

Summary

One of the concerns of knowledge discovery in databases is the evaluation of association rules' interestingness. Many interestingness measures have been introduced in order to rank such rules according to their interest. These measures are heterogeneous and the ranking of rules may differ largely. This is why, rather than privileging one measure it appears interesting to take account the various information brought by the measures. Thus we explore a new approach: the aggregation of valued relations that represent the intensity of preference of a rule over another. The aim in using such tools is to take into account the numerical nature of measures and reduce incomparability issues.

We studied several aggregation operators. In this contribution we discuss results obtained on a toy example using the simplest of them.

Extraction de mesures d'intérêt représentatives pour le post-traitement des règles d'association

Xuan-Hiep Huynh*, Fabrice Guillet*, Henri Briand*

*LINA CNRS FRE 2729 - Ecole polytechnique de l'université de Nantes
La Chantrerie, BP 50609, 44306 Nantes cedex 3, France
{xuan-hiep.huynh,fabrice.guillet,henri.briand}@univ-nantes.fr

Résumé. Cet article s'intéresse au calcul d'une base réduite de mesures d'intérêt pour le post-traitement des règles d'association. Ces mesures, appelées mesures d'intérêt représentatives, sont calculées à l'aide d'une classification par méдиоіdes. L'intérêt de cette approche est de pouvoir offrir à un utilisateur un ensemble réduit de mesures les mieux adaptées à la nature de ses données, et ainsi de lui faciliter la validation des meilleures règles. Nous dressons un état de l'art sur le post-traitement et les travaux relatifs aux mesures d'intérêt. Puis notre approche est appliquée sur un jeu d'essai comportant 40 mesures d'intérêt et environ 120000 règles. Sur cette expérimentation, nous avons pu montrer l'intérêt de notre approche et proposer de réduire le nombre de mesures à seize mesures d'intérêt représentatives.

1 Introduction

Depuis les travaux précurseurs (Agrawal et al., 1993) (Agrawal et Srikant, 1994) en extraction de connaissances dans les données, la validation des règles d'association demeure un verrou scientifique pénalisant l'usage de cette technique.

En effet, bien que le modèle des règles d'association ait l'avantage de permettre une extraction non supervisée de règles exprimant des tendances implicatives dans les données, il a malheureusement l'inconvénient de produire de très grandes quantités de règles. L'étape finale de validation de ces règles étant laissée à l'utilisateur (un décideur ou un analyste), celui-ci se trouve alors confronté à une tâche généralement inextricable : comment repérer les règles intéressantes parmi plusieurs milliers, voire plusieurs millions de règles ?

Il s'avère donc indispensable d'aider l'utilisateur par la mise en oeuvre d'une phase de post-traitement des règles produites. Ce post-traitement a pour objectif de réduire le volume des règles, en présélectionnant un nombre réduit de règles potentiellement intéressantes pour l'utilisateur, au sens de ses critères de préférence et en fonction de la structure des données.

A cette fin, cinq approches complémentaires de post-traitement sont proposées dans la littérature : par représentations graphiques, par recherche interactive de résumés, par réduction des redondances, par aide à la décision multicritères, et par mesures d'intérêt. La première consiste à utiliser des représentations graphiques afin d'améliorer la lisibilité des résultats (Blanchard et al., 2003). La seconde consiste à considérer l'ensemble de règles comme une base de données sur laquelle l'utilisateur extrait des sous ensembles de règles par des requêtes intégrant des

contraintes (Blanchard et al., 2003) (Blanchard et al., 2005a) (Blanchard et al., 2005b), (Hilderman et Hamilton, 2001), (Boulicaut et al., 1998). La troisième consiste à opérer une réduction très significative (généralement 10%) en ne conservant que les règles non redondantes au sens logique (Lehn et al., 2004). La quatrième consiste à concevoir la sélection des meilleures règles comme un problème d'optimisation en aide à la décision multicritères (Lenca et al., 2004). La dernière consiste à valuer les règles par des indicateurs numériques : les mesures d'intérêt.

Notre travail s'insère dans le cadre de cette dernière approche par mesures d'intérêt. Depuis l'introduction du support et de la confiance (Agrawal et Srikant, 1994), de nombreuses mesures ont été proposées dans la littérature (Bayardo-Jr. et Agrawal, 1999) (Hilderman et Hamilton, 2001) (Tan et al., 2004). Cette abondance de mesures d'intérêt pose un second problème : comment aider l'utilisateur à choisir les mesures les mieux adaptées à ses buts et à ses données, afin de détecter in fine les meilleures règles ?

Dans cet article, nous proposons une technique d'analyse de données pour calculer les meilleures mesures d'intérêt objectives sur un jeu de règles. Pour cela, nous utilisons une classification par médioïdes (PAM : partitioning around medoids (Kaufman et Rousseeuw, 1990)) avec comme métrique l'indice de corrélation linéaire, afin de partitionner quarante mesures d'intérêt en k clusters. Le sous-ensemble de mesures d'intérêts représentatives du jeu de données étudié est alors constitué par les k médioïdes obtenus.

L'article est organisé en cinq parties. Après une synthèse des travaux relatifs aux mesure d'intérêt des règles d'association, nous présentons notre approche de calcul des mesures représentatives en employant la technique d'analyse de données PAM. Puis, nous décrivons l'ensemble de données et les mesures d'intérêt étudiées. Enfin, nous expérimentons notre méthode sur la base de données mushroom (Newman et al., 1998) et discutons les résultats obtenus.

2 Travaux relatifs sur la qualité des connaissances

Dans les travaux précurseurs sur les règles d'association (Agrawal et al., 1993) (Agrawal et Srikant, 1994), deux premières mesures statistiques¹ sont introduites : le support et la confiance. Celles-ci sont bien adaptées aux contraintes algorithmiques (cf Apriori), mais ne sont pas suffisantes pour capturer l'intérêt des règles pour l'utilisateur. Afin de contourner cette limite, de nombreuses mesures d'intérêt complémentaires ont été proposées dans la littérature. Freitas (1999) distingue deux types de mesures d'intérêts : les mesures subjectives, et les mesures objectives. Les mesures *subjectives* dépendent des buts, connaissances, croyances de l'utilisateur et sont combinées à des algorithmes supervisés spécifiques afin de comparer les règles extraites avec ce que l'utilisateur connaît ou souhaite (Padmanabhan et Tuzhilin, 1998) (Liu et al., 1999). Ainsi, les mesures subjectives proposent de capturer la nouveauté (novelty) ou l'inattendu (unexpectedness) d'une règle par rapport aux connaissances/croyances de l'utilisateur. Les mesures *objectives*, quand à elles, sont des indices statistiques qui évaluent la contingence d'une règle dans les données.

Dans la littérature, de nombreuses synthèses traitent des mesures d'intérêt selon deux aspects différents : la définition de l'ensemble des principes constitutifs d'une bonne mesure

¹Lallich et al. (2005) ont proposé une autre façon de distinguer une mesure d'intérêt objective par sa "nature" *statistique* ou *descriptive*. Les mesures descriptives ne varient pas avec l'expansion de cardinalité. Bien au contraire, les mesures statistiques varient avec l'expansion de cardinalité. Ici, nous disons simplement les mesures statistiques dans tous ces deux cas.

d'intérêt, et leur comparaison par classification en fonction de critères théoriques ou expérimentaux sur des données.

Dans la perspective d'établir les principes d'une bonne mesure d'intérêt, Piatetsky-Shapiro (1991) a présenté une nouvelle mesure d'intérêt, appelé Rule-Interest, et propose trois principes fondamentaux pour une mesure sur une règle $a \rightarrow b$: (P1) valeur 0 quand a et b sont indépendants, (P2) croissant avec $a \wedge b$, (P3) décroissant avec a ou b . Hilderman et Hamilton (2001) ont proposé cinq principes : minimum value, maximum value, skewness, permutation invariance, transfer. Tan et al. (2004) ont défini cinq principes d'intérêt : symmetry under variable permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas (1999) a proposé un principe de "surprise" d'attribut. Gras et al. (2004) ont proposé un ensemble de dix critères constitutifs d'une bonne mesure d'intérêt.

Parmi ces synthèses, certaines abordent également la comparaison des mesures d'intérêt à partir de leur classification. Bayardo-Jr. et Agrawal (1999) ont conclu que les meilleures règles selon toutes les mesures d'intérêt qu'ils étudient doivent résider le long d'une frontière de support/confiance. Kononenco (1995) a utilisé onze mesures pour estimer la qualité des attributs à valeurs multiples, et montre que les valeurs des mesures : information-gain, j-mesure, gini-index, et relevance tendent à augmenter linéairement avec le nombre de valeurs d'un attribut. Zhao et Karypis ont proposé un algorithme d'optimisation de huit critères et montrent qu'une partie des critères suffit pour extraire les meilleures règles. Gavrilov et al. (1999) ont étudié la similitude des mesures afin de les classer. Hilderman et Hamilton (2001) ont proposé cinq principes pour ranger des résumés de bases de données en employant seize mesures de diversité et montrent que : (1) six mesures ont satisfait cinq des principes proposés, (2) neuf des mesures restantes ont satisfait au moins un des principes proposés. En étudiant vingt et une mesures, Tan et al. (2004) ont montré qu'aucune mesure n'est adaptée à tous les cas et que la corrélation des mesures augmente avec la diminution du support. Par une méthode d'aide à la décision multicritère intégrant huit critères, (Vaillant et al., 2004) (Lenca et al., 2004) ont extrait un pré-ordre sur vingt mesures et identifient quatre clusters de mesures. (Carvalho et al., 2005) (Carvalho et al., 2003) ont évalué onze mesures d'intérêt objectives afin de les ranger en fonction de leur intérêt effectif pour un décideur. Choi et al. (2005) ont utilisé une approche d'aide à la décision multi-critère pour trouver les meilleures règles d'association. Lallich et Teytaud (2004) ont utilisé quinze mesures et ont proposé des critères pour les évaluer. (Blanchard et al., 2005a) (Blanchard et al., 2005b) ont classé dix-huit mesures objectives en quatre groupes selon trois critères : indépendance, équilibre, et caractère descriptif ou statistique. Huynh et al. (2005b) ont proposé une approche de classification par graphes de corrélation qui permet d'identifier onze clusters sur trente-quatre mesures d'intérêt.

Enfin, deux outils d'expérimentation sont disponibles : HERBS (Vaillant et al., 2003) et ARQAT (Huynh et al., 2005a).

3 Calculs des mesures représentatives

3.1 Formalisation des données et de la distance corrélative entre des mesures d'intérêt

Soient $R(D) = \{r_1, r_2, \dots, r_p\}$ un ensemble de p règles d'association extraites d'un ensemble de données D . Chaque règle $a \rightarrow b$ est décrite par ses deux itemsets (a, b) et ses cardinalités $(n, n_a, n_b, n_{a\bar{b}})$. Soit $M = \{m_1, m_2, \dots, m_q\}$ un ensemble de q mesures d'intérêt. Chaque mesure est calculée par une fonction numérique sur les cardinalités d'une règle : $m(a \rightarrow b) = f(n, n_a, n_b, n_{a\bar{b}})$. Précisément, n est le nombre total d'enregistrements de D , n_a (resp. n_b) le nombre d'enregistrements de D satisfaisant a (resp. b), et $n_{a\bar{b}}$ le nombre d'enregistrements satisfaisant $a \wedge \bar{b}$ (les exemples négatifs).

Chaque règle étant évaluée par q mesures, nous disposons en entrée de notre problème d'une matrice numérique $((m_{ij}))$ de dimensions $q \times p$ dont chaque élément $m_{ij} = m_i(r_j)$ correspond à la valeur de la mesure m_i sur la règle r_j , avec $i = 1..q$ et $j = 1..p$.

Nous calculons ensuite $((d_{ij}))$ la matrice carrée de dimension q des distances entre mesures d'intérêt, à l'aide d'une distance corrélative :

$$d_{ij} = 1 - |\eta_{ij}|$$

où η_{ij} est le coefficient de corrélation linéaire (Saporta, 1990) entre deux mesures m_i et m_j :

$$\eta_{ij} = \frac{\sum_{k=1}^p [(m_{ik} - \bar{m}_i)(m_{jk} - \bar{m}_j)]}{\sqrt{[\sum_{k=1}^p (m_{ik} - \bar{m}_i)^2][\sum_{k=1}^p (m_{jk} - \bar{m}_j)^2]}}$$

3.2 Calcul des mesures représentatives avec PAM

Nous partons de la matrice $((d_{ij}))$ des distances entre mesures d'intérêts, afin de partitionner l'ensemble des mesures en k clusters, puis d'extraire le meilleur représentant de chaque cluster. Pour cela, nous avons choisi d'utiliser la méthode PAM (Partitioning Around Medoids (Kaufman et Rousseeuw, 1990)). Cette méthode de classification non supervisée, variante des k -moyennes (MacQueen, 1967), à l'avantage d'être plus robuste que cette dernière.

De plus, la méthode PAM que nous utilisons (PAM dans \mathbb{R}^2 , (Kaufman et Rousseeuw, 1990)), à le grand avantage de permettre la visualisation graphique de la classification obtenue selon les deux premiers axes principaux d'une ACP.

La partition en clusters est basée sur la notion de médioïde. Un médioïde m_i^* d'un cluster ψ_i est défini comme l'élément de ce cluster le plus proche de tous les autres³.

Partant d'un ensemble de k médioïdes initiaux, la méthode procède en plusieurs itérations jusqu'à obtenir la stabilisation des médioïdes. A chaque itération, chaque objet est classé dans le cluster du médioïde le plus proche, puis le médioïde de chaque cluster est recalculé.

Nous rappelons quelques notions définies dans Kaufman et Rousseeuw (1990).

- La qualité de la classification sur les k médioïdes à l'aide d'une fonction, appelée *fonction objective*, définie par : $\sum_{j=1}^q \min_{i=1, \dots, k} d(j, m_i^*)$, où q est le nombre d'objets à classer.

²<http://www.r-project.org/>

³En calculant la distance moyenne dans notre approche.

- Un cluster ψ est dit isolé si tout objets du cluster est toujours plus proche d'un second objet de son cluster que de tout autre objet des autres clusters : $\forall i, j \in \psi, \forall h \notin \psi, \max(d_{ij}) < \min(d_{ih})$
- Le diamètre d'un cluster ψ est la plus grande distance entre deux objets du même cluster : $\max(d_{ij}), i, j \in \psi$
- La séparation d'un cluster ψ est définie comme la distance minimale entre un objet d'un cluster et un objet d'un autre cluster : $\min(d_{ij}), i \in \psi, j \notin \psi$
- La distance moyenne à un médioïde j d'un cluster ψ est donnée par : $\frac{\sum_{i \in \psi} (d_{ij})}{|\psi|}$
- La distance maximale à un médioïde j d'un cluster ψ est donnée par : $\max(d_{ij}), i \in \psi$

4 Expérimentation

4.1 Description des données, des règles, et des mesures étudiées

Nous expérimentons notre approche sur la base de données catégoriques mushroom (D) issue du dépôt d'Irvine (Newman et al., 1998). Nous avons ensuite calculé l'ensemble de règles d'association R sur ces données à l'aide de l'algorithme Apriori (Agrawal et Srikant, 1994). Le tableau Tab. 1 récapitule les caractéristiques principales des données étudiées.

Ensemble de données	Article	Transactions	Longueur moyenne des transactions	Nombre de règles (seuil de support)	Jeu de règles
D_1	118	8416	22	123228 (12%)	R_1

TAB. 1 – Description des données.

Cet ensemble de 123228 règles est évalué par la matrice $((m_{ij}))$ sur quarante mesures d'intérêt. La plupart des références détaillées sur ces mesures peuvent être trouvées dans les synthèses (Hilderman et Hamilton, 2001) (Tan et al., 2004). Nous utilisons trente-quatre mesures d'intérêt dont les formules peuvent être trouvées dans Huynh et al. (2005b), les 6 mesures restantes étant définies dans le tableau Tab. 2.

4.2 Clusters de mesures d'intérêt obtenues

La classification obtenue est illustrée par trois figures. La première (Fig. 1) présente les résultats sous une forme graphique en projetant les mesures selon les deux axes principaux obtenus par une analyse en composantes principales. Chaque mesure y est représentée par un symbole, et chaque cluster par un numéro. Bien que déformée par la projection, cette visualisation s'avère très utile pour avoir une vue synthétique de la classification, mais aussi pour valider le choix du nombre k de clusters, ici fixé à seize (Ce choix de seize clusters est issu de travaux préliminaires présentés dans Huynh et al. (2005c)).

Le tableau Tab. 3 récapitule les clusters obtenus et leur médioïde. Le tableau Tab. 4 détaille les caractéristiques numériques de chaque cluster.

On peut y observer que les clusters 1 (Causal Confidence, Causal Confirmed-Confidence, Confidence, Descriptive Confirmed-Confidence, Ganascia, Laplace) et 2 (Causal Confirm,

