QDC 2007

Actes du 3e Atelier

Qualité des Données et des Connaissances

En conjonction avec EGC 2007

23 Janvier 2007
Namur, Belgique

Organisé par
Stéphane Lallich, Philippe Lenca et Fabrice Guillet

Troisième Atelier
Qualité des Données et des Connaissances

23 Janvier 2007, Namur, Belgique

Préface

Après le succès des précédents ateliers Qualité des Données et des Connaissances, conjointement avec la Conférence EGC, Extraction et Gestion des Connaissances - à Paris en 2005 et à Lille en 2006 - la troisième édition de cet atelier, QDC 07, est organisée cette année à Namur, en conjonction avec EGC 07.

Comme le montrent les différentes communication retenues, cet atelier se concentre sur les différentes étapes du processus de fouille des données : les méthodes d'analyse et de nettoyage des données, les approches algorithmiques utilisées pour extraire les connaissances à partir de ces données, le choix des mesures qui permettent d'en évaluer la qualité ainsi que la validation et l'exploitation des connaissances extraites.

La qualité des données sur lesquelles va porter le processus de fouille des données est un facteur décisif pour la performance de la fouille et qualité des connaissances extraites. Le travail de Delphine Clément et Brigitte Laboisse qui présentent un référentiel d'indicateurs de mesure de la qualité des données relatives à la gestion de la relation client s'inscrit dans ce thème.

Deux communications proposent des procédures algorithmiques originales. Pour faire face à l'explosion combinatoire des motifs fréquents, notamment en text-mining, Martine Cadot et Alain Lelu associent des tests statistiques réalisés par simulation sur les 2-motifs et une procédure itérative de construction de sur-motifs non redondants. Dans le domaine de la classsification associative, Abdelhamid Zemirline, Laurent Lecornu et Basel Solaiman proposent

une nouvelle méthode d'extraction de règles d'association de classe multi-labels.

Deux communications portent sur la mesure de la qualité des connaissances extraites. Régis Gras, Jérôme David, Fabrice Guillet et Henri Briand étudient la stabilité de différentes mesures de qualité, principalement l'intensité d'implication, en fonction des variations des différents effectifs qui interviennent dans le calcul de ces mesures. Stéphane Lallich, Philippe Lenca et Benoît Vaillant proposent une méthode de décentrage s'appliquant aussi bien à l'entropie de Shannon qu'à toute entropie généralisée, qui conduit à des entropies prenant leur valeur maximale pour une distribution de référence fixée par l'utilisateur.

Le décalage est souvent assez grand entre les expérimentations sur des jeux d'essais et la mise en œuvre de travaux académiques sur des données réelles en collaboration avec un expert métier. Le travail de Benoit Vaillant, Stéphanie Menou, Sorin Moga, Philippe Lenca et Stéphane Lallich qui porte sur l'analyse de données relatives à l'activité d'un serveur vocal devant guider les utilisateurs vers un service d'assistance spécifique relève de cette confrontation. Comme le montrent Jean-Hugues Chauchat et Annie Morin, la validation statistique des connaissances extraites exige que soit connu le mode de constitution de la base de données utilisée et l'interprétation des connaissances extraites est un art difficile.

Nous remercions chaleureusement les auteurs et les membres du comité de programme de l'atelier pour leur contribution au succès de l'atelier QDC 2007.

Stéphane Lallich, Philippe Lenca et Fabrice Guillet

Organisateurs de QDC 2007

Comités

Comité d'Organisation

- Stéphane Lallich
- Philippe Lenca
- Fabrice Guillet

Comité de Programme

- Alexandre Aussem, France
- Jérôme Azé, France
- Laure Berti Equille, France
- Julien Blanchard, France
- Henri Briand, France
- Béatrice Duval, France
- Régis Gras, France
- Fabrice Guillet, France
- Yves Kodratoff, France
- Stéphane Lallich, France
- Ludovic Lebart, France
- Philippe Lenca, France
- Israel-César Lerman, France
- Patrick Meyer, Luxembourg
- Sorin Moga, France
- Annie Morin, France
- Amedeo Napoli, France

- Ricco Rakotomalala, France
- Gilbert Ritschard, Suisse
- Ansaf Salleb-Aouissi, USA
- Einoshin Suzuki, Japon
- Benoit Vaillant, France
- Christel Vrain, France

Table des matières

Création d'un référentiel d'indicateurs de mesure de la qualité des données CRM
Delphine Clément, Brigitte Laboisse **5**

Simuler et épurer pour extraire les motifs sûrs et non redondants
Martine Cadot, Alain Lelu **15**

Nouvelle méthode d'extraction de règles de classification multi-labels
Abdelhamid Zemirline, Laurent Lecornu, Basel Solaiman **25**

Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association
Régis Gras, Jérôme David, Fabrice Guillet, Henri Briand **35**

Construction d'une entropie décentrée pour l'apprentissage supervisé
Stéphane Lallich, Philippe Lenca, Benoit Vaillant **45**

Qualité des règles d'association : étude de données d'entreprise
Benoît Vaillant, Stéphanie Menou, Sorin Moga, Philippe Lenca, Stéphane Lallich **55**

Quelques erreurs et abus courants dans l'interprétation des résultats de la fouille des données
Jean-Hugues Chauchat, Annie Morin **65**

Index des auteurs **71**

Création d'un référentiel d'indicateurs de mesure de la qualité des données CRM

Delphine Clément
Analyste Qualité des Données
Hewlett-Packard
14 rue du Général Caunègre – 40000 MONT DE MARSAN
delphine_clement@hp.com

Brigitte Laboisse
Directeur technique
A.I.D.
4 rue Henri Le Sidaner – 78000 VERSAILLES
blaboisse@aid.fr
http://www.aid.fr

Résumé : Le sujet de cet article est de présenter le Référentiel d'indicateurs de mesure de la qualité des données Customer Relationship Management (CRM) publié avec l'Association Française de Normalisation (AFNOR). Il s'agit avant tout d'expliquer pourquoi ce référentiel a été créé, les grands chapitres, pour quels utilisateurs, et de fournir des éléments méthodologiques autour de l'identification du client/prospect. La deuxième partie de ce papier est la présentation d'un cas pratique avec l'implémentation de ce référentiel chez Hewlett-Packard, dans un contexte international.

1 Introduction

Analyse Informatique des Données (A.I.D.) est à l'initiative et a collaboré à l'élaboration d'un référentiel d'indicateurs de la mesure de la qualité des données CRM avec l'AFNOR. La première partie de cet article décrit le groupe de travail, la méthode pour constituer ce référentiel et présente les raisons de choix des différents indicateurs. Un focus particulier est fait sur les méthodes d'identification des individus : il est rappelé les méthodes proposées par les différents outils commerciaux, celle retenue dans le référentiel et il est proposé une méthode de comparaison pour utiliser les résultats de la méthode référentiel. A notre connaissance, peu de travaux ont été publiés sur les méthodes de comparaison de dédoublonnage et ce papier propose quelques évolutions par rapport aux travaux publiés.

La seconde partie de cet article est un exemple de mise en place de ce référentiel dans un contexte international chez Hewlett Packard. Les principaux freins rencontrés sont décrits ainsi que les solutions envisagées.

Enfin, en conclusion, les évolutions envisagées pour ce référentiel sont présentées.

2 Contexte

2.1 Le métier d'A.I.D.

A.I.D. est une société de services française spécialisée dans les Bases de Données marketing, la qualité de données et l'enrichissement statistique. A.I.D. travaille à l'international de part son appartenance au groupe de communication OMNICOM et les outils, référentiels développés depuis plusieurs années, et ce au niveau mondial.

Le métier d'A.I.D. est donc autour de la donnée, spécifiquement sur les données Marketing, CRM, 'Identification du client/prospect'.

2.2 L'absence de norme, de référentiels

Dans le domaine d'intervention d'A.I.D., il n'existe pas, à notre connaissance, de normes, de référentiels. La Poste française a mis en place Visa EV@®[1] et Labeladresse®[2] qui permettent de valoriser les bases d'adresses tenues correctement :

- Avec des adresses conformes à la norme Afnor XPZ 10-011 ou Z 10-011
- Avec des adresses en cours de validité

Cette première initiative, à notre connaissance, n'a pas été poursuivie sur d'autres canaux (téléphone, email), et quand on parle de taux de faux téléphones dans un CRM par exemple, les chiffres qu'on pourra obtenir seront en général :

- L'utilisateur marketing fournira le dernier taux de la campagne remonté par le centre d'appels
- Le centre d'appels inclura les sociétés ou personnes qui ne répondent pas au téléphone
- Le responsable Système d'Information, en général, n'aura pas de réponse.

La conséquence de cette absence de norme, de référence est bien connue : l'utilisateur se sent démuni, sans éléments de comparaison, et va, sur un sentiment, une impression, fournir un jugement sur la qualité des données :

- « Il y a trop de valeurs manquantes »
- « Il y a trop de valeurs erronées »
- « C'est inutilisable »
- « En effet, j'ai comme l'impression ces derniers temps d'une dégradation de la qualité de notre base de données, ce que je ne peux accepter. Je compte donc sur vous pour faire le nécessaire »

[1] Visa EV@® est une marque déposée de La Poste.
[2] Labeladresse® est une marque déposée de La Poste.

3 Création d'un référentiel d'indicateurs de la mesure qualité des données d'un CRM

3.1 Le livrable

A.I.D. a entamé en 2005, en participation avec l'AFNOR, la démarche de mise en place d'un référentiel d'indicateurs qui permettent de mesurer la qualité de données d'un CRM. Le résultat est un document : AC X50-111 disponible à l'achat sur le site de l'AFNOR http://www.boutique.afnor.org/NEL1AccueilNormeEnLigne.aspx
Ce document n'est pas une norme, mais plutôt un 'guide de bonnes pratiques'. Il a été rédigé par les auteurs de cet article sous la direction d'un consultant AFNOR. Un groupe de professionnels : statisticiens (Gilbert Saporta, Conservatoire National des Arts et Métiers (CNAM)), responsables marketing (Air Liquide par exemple), responsables CRM (Auchan par exemple), responsables qualité de données (Hewlett Packard par exemple) a permis de valider la pertinence, l'approche opérationnelle des indicateurs retenus afin de mesurer la qualité des données d'un CRM, sur la partie Identification des individus.

3.2 Identification des individus

L'identification des individus est un vaste sujet, largement abordé dans la littérature Qualité de Données : Madnick et al. (2002) présentent par exemple une méthode d'identification des entreprises prenant en compte la notion de contexte.

3.2.1 Qu'est-ce qu'un individu ?

On distingue, en général :
Dans une base de données Business to Consumer (B2C), 2 types d'individus :
- La personne physique : Raoul Dupont, numéro de sécurité sociale 1XXXXXXXXXXXX
- Le foyer : [1] regroupement de personnes vivant à la même adresse. Cela peut inclure les couples mariés, les concubins, leurs ascendants ou descendants vivant à la même adresse.

Dans une base de données Business to Business (B2B), 2 types d'individus :
- La personne physique : Raoul Dupont, numéro de sécurité sociale 1XXXXXXXXXXXX
- Le site : [1] Etablissement d'une entreprise : nom, adresse d'une entreprise. La notion de site est en général assimilée à une relation bijective (1,1) avec un numéro issu du Système d'Identification du Répertoire des Etablissements (SIRET) (identifiant unique fourni par l'Institut National de la Statistique et des Etudes Economiques (INSEE)), mais on peut noter des exceptions : adresse non déclarée à l'INSEE : par exemple, dans le cas d'un entrepôt ou annexe. On aura alors une notion de site sans SIRET.

3.2.2 Que signifie identifier un individu?

- Qu'il est représenté d'une manière unique dans le système d'information, de manière non ambiguë
- Qu'on peut s'adresser à lui : soit par courrier, soit par téléphone, fax ou courriel
- Qu'on connaît son 'état' : décédé, en cessation d'activité,…

Il s'agit de retrouver dans un ensemble d'enregistrements ceux qui correspondent au même individu. Par exemple :

Exemple 1 :

DUPONT RAOUL	A.I.D.	4 rue le Sidaner	78000	Versailles	0139239345
DUPONT RAOUL	A.I.D.	rue Henri le Sidane	79700	Versailles Cedex	

Exemple 2 :

DUPONT RAOUL	A.I.D.	4 rue Henri le Sidaner	78000	Versailles	0139239345
DUPONT RAOUL		6 rue des Acacias	78150	Le Chesnay	0139239345

La détection de doubles est différente sur une personne, un foyer ou un site.

Détection de doubles sur une personne :

Les procédés utilisés consistent en général à 'nettoyer' les données, les confronter à un référentiel afin de pouvoir les comparer ensuite à l'identique entre elles à l'aide de 'match-keys' ou clés de rapprochement. Par exemple, on va comparer les adresses aux référentiels de La Poste (HEXAVIA, HEXAPOSTE) et corriger l'information erronée.

Dans notre exemple 1, le résultat sera :

4 rue le Sidaner	78000	Versailles	➔	4 rue Henri le Sidaner	78000 Versailles
rue Henri le Sidane	79700	Versailles	➔	rue Henri le Sidaner	78000 Versailles

Cette correction par un référentiel est en général utilisée pour l'adresse, le prénom. Sur le téléphone ou l'email, il s'agit plus d'un nettoyage de la syntaxe de part le manque de référentiels.

La comparaison à l'identique ensuite sur des clés de rapprochement (par exemple les 4ers caractères du nom, la 1ère lettre du prénom, le nom de la rue, et la ville) est motivée par un besoin de rapidité : les volumes concernés sont en général importants.

Détection de doubles sur un site :

Les méthodes utilisées sur les personnes ne sont pas performantes sur les sites :

- Un nom de site pourra apparaître sous forme de sigle : A.I.D. ou Analyse informatique de Données
- Ce nom pourra inclure des mots creux ou synonymes : Société Anonyme (SA), Mairie ou Commune,..
- On trouvera souvent une adresse physique (la rue) ou une adresse postale (boîte postale)

La comparaison à l'identique sur des clés de rapprochement ne fonctionne pas. Les logiciels du marché (Principaux éditeurs : Fuzzy Logic, Business Objects, Harte&Hanks, ..) consistent à comparer tous les enregistrements sur un sous ensemble de la population (ville ou département par exemple) et à fournir des règles telles que :

Si index1(nom1,nom2) > 90 et index2(adr1,adr2) > 70 alors double

Index1 et Index2 sont des index de proximité entre les 2 noms nom1 et nom2. Les index implémentés dans les logiciels peuvent se baser sur la phonétique, sur le nombre de modifications entre les 2 chaînes par exemple (Berti-Equille (2005) fournit une liste d'algorithmes).

Comparaison de méthodes :

Les méthodes d'identification présentées ci-dessus pourront donner des résultats différents, de part les algorithmes de proximité utilisés, mais également des règles de comparaison (variables retenues, combinaisons, seuils). Comparer les résultats de différents outils, ou du même outil sur des règles différentes est un travail laborieux, souvent fait en manuel sur quelques enregistrements. Des méthodes automatiques Hashemi et Talburt, (2006), Talburt et Hess (2004), Wang (2000) émergent : le principe consiste à comparer 2

partitions A et B et à fournir un index de proximité : nombre d'enregistrements classés ensemble par les 2 méthodes.

Soit l'exemple ci-dessous : 2 dédoublonnages ont été menés, qui ont permis de classifier la population en A et B. Dans notre contexte, par exemple, le dédoublonnage A est le résultat de la méthode préconisée dans le référentiel et le dédoublonnage B correspond à un test fait avec les outils usuels du CRM pour corriger les doubles détectés par A. On observe ainsi que la partition B3 a mis ensemble 3 individus, alors que ces derniers se répartissent pour 2 individus dans la partition A3 et pour 1 dans la partition A6. Les pourcentages horizontaux et verticaux fournissent le taux de convergence entre les 2 partitions : 100% de A3 est inclus dans B3, 66,67% de B3 est inclus dans A3. On peut calculer un index de proximité entre les partitions A et B :

Index de proximité =

$$\frac{\text{Nombre d'individus dans les cellules avec un \% vertical et horizontal >= seuil}}{\text{Nombre d'individus}}$$

Soit un seuil à 90% : 2 partitions seront considérées comme convergentes si A $\subset$ B à 90% et réciproquement B $\subset$ A à 90%.

Dans notre exemple = (1+1+1+1+1)/10 = 5/10 = 0.5

Par rapport aux travaux publiés cités ci-dessus, l'index que nous proposons prend en compte l'effectif dans les cases (pas uniquement 0/1 si la case est remplie) et filtre les intersections selon un pourcentage paramétrable. L'intérêt est d' écarter les différences peu significatives et de prendre en compte les volumes importants de convergence.

Fréquences

	B1	B2	B3	B4	B5	B6	B7	Total
A1	1							1
A2		1						1
A3			2					2
A4	1							1
A5				1				1
A6			1					1
A7					1			1
A8						1		1
A9							1	1
Total	2	1	3	1	1	1	1	

% horizontaux

	B1	B2	B3	B4	B5	B6	B7	Total
A1	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
A2	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
A3	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	100,00%
A4	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
A5	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	100,00%
A6	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	100,00%
A7	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	100,00%
A8	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	100,00%
A9	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	100,00%

% verticaux

	B1	B2	B3	B4	B5	B6	B7
A1	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
A2	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%
A3	0,00%	0,00%	66,67%	0,00%	0,00%	0,00%	0,00%
A4	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
A5	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%	0,00%
A6	0,00%	0,00%	33,33%	0,00%	0,00%	0,00%	0,00%
A7	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	0,00%
A8	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%
A9	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

TAB 1 – *Tableau de comparaison de 2 dédoublonnages A et B : fréquences, % horizontaux, % verticaux*

3.3 Mon CRM identifie-t-il correctement les individus ?

La démarche de mesure proposée dans le document AFNOR est volontairement pragmatique et simple. Il ne s'agit pas de trouver l'ensemble des doubles mais bien de mesurer et ce de la même manière, quel que soit le CRM. La méthode de mesure proposée, inspirée des méthodes décrites plus haut, est en 2 étapes :

- Travail manuel sur un échantillon : on recommande d'extraire une zone géographique ou une lettre de l'alphabet selon la complétude de l'information.
- Un calcul complémentaire par clés de dédoublonnage sur l'ensemble de la population est recommandé afin d'essayer de corriger le biais de l'échantillon précédent : le taux de doubles peut varier sensiblement d'une zone géographique à l'autre.

Des recommandations sur le choix de l'échantillon, les clés de dédoublonnage sont fournies ainsi que sur le calcul du taux de doubles.

Il sera possible, en fonction des résultats, de donner ces éléments de contrôle aux responsables d'outils de dédoublonnage afin qu'ils comparent (voir méthode de comparaison ci-dessus) et améliorent leur dédoublonnage en se comparant aux résultats manuels obtenus.

3.4 Puis je contacter les individus de mon CRM ?

Puis-je m'adresser aux individus par un des canaux usuels : téléphone, courrier, email ? C'est-à-dire, ai-je au minimum les coordonnées pour un canal, et cette donnée est-elle exacte ? La mesure de l'exactitude est proposée selon 2 méthodes :

- Les indicateurs a priori : on mesure, en général sur des échantillons, la qualité de l'adresse, du téléphone, de l'email. On parle ici d'indicateur a priori dans le sens où l'information est contrôlée avant son utilisation ou de manière indépendante.
- Les indicateurs a posteriori : ils résultent de l'usage observé opérationnel des données : taux de N'habite Pas à l'Adresse Indiquée (NPAI), taux de faux téléphones, taux de faux emails,…

Avantages et inconvénients de chaque type d'indicateurs :

Indicateurs a priori

Avantages : Mise en œuvre indépendante de l'usage des données, actions préventives possibles avant l'utilisation des données.

Inconvénients : Difficile d'avoir une vue complète de la qualité des données : la validation d'un téléphone passe à ce jour par un appel pour une meilleure efficacité. Les référentiels ne sont pas assez exhaustifs pour être efficaces. Méthode de mesure souvent mal reconnue par les utilisateurs

Indicateurs a posteriori

Avantages : Vue réelle de la qualité des données à l'utilisation. Méthode de mesure bien acceptée par les utilisateurs

Inconvénients : Vue partielle sur la qualité de données : périmètre utilisé peut être faible. Difficulté importante à avoir les retours d'une manière exhaustive et fiable (voir ci-dessous)

Le référentiel décrit l'ensemble de ces indicateurs, avec une recommandation sur la mise en place. Les freins, problèmes rencontrés, les biais sont décrits avec des solutions palliatives autant que faire ce peut.

4 Mise en œuvre opérationnelle : cas pratique

Dans le cadre de la gestion de la relation clients sur le périmètre des petites et moyennes entreprises en Europe, Moyen-Orient et Afrique (EMEA), Hewlett-Packard pilote, depuis un an, la mesure de quatre indicateurs a posteriori que sont le taux de NPAI, le taux de « bounced » emails ou emails non aboutis, le taux de faux téléphones et le taux de contacts obsolètes. Cette mesure est produite à la fin de chaque trimestre, avec un décalage d'un à deux mois par rapport à la fin du trimestre, afin d'assurer la récupération d'un maximum de NPAI.

4.1 Méthodes de collecte des informations non abouties et difficultés rencontrées

NPAI

Chaque pays de la région a sa propre méthode de traitement des NPAI. Nous pouvons regrouper les méthodes en 3 catégories principales :

1. Les pays qui traitent les NPAI localement et produisent eux-mêmes leur taux de NPAI. Ces pays se contentent de nous fournir un taux déclaratif ; nous n'avons malheureusement pas accès à l'enregistrement site et contact en retour, ce qui nous empêche d'analyser les caractéristiques de cet enregistrement, d'investiguer sur les causes du retour et sur les possibles corrélations avec d'autres facteurs. Cependant, à défaut, la mesure de l'évolution dans le temps de ce taux est toujours intéressante.
2. Les pays qui sous-traitent les NPAI à une agence locale. Pour ces pays, il nous est possible de récupérer les enregistrements en retour ainsi que le code campagne auquel ils se rattachent. Ceci étant, chaque pays ayant sa propre agence, les fichiers récupérés sont tous de différents formats, le niveau d'information retourné varie également d'un pays sur l'autre ; pour certains pays, nous arrivons à obtenir le motif de non-distribution indiqué sur l'enveloppe par exemple, alors que pour d'autres, nous n'avons que l'identifiant de l'enregistrement, à nous de faire ensuite un croisement avec le système d'information pour récupérer les données. De plus, même si nous attendons de un à deux mois après la fin du trimestre avant de produire la mesure, nous ne sommes jamais certains de récupérer l'exhaustivité des NPAI du trimestre.
3. Les pays qui passent en gestion centralisée de leur NPAI. La Suisse, les Pays-Bas et la Suède ont expérimenté cette méthode. Cependant, l'équipe centralisée étant en Inde, il est extrêmement onéreux d'y envoyer des caisses entières de catalogues en NPAI ; les frais engagés, ainsi que les problèmes de douane rendent cette solution non viable. Dès lors, une telle méthode doit être combinée avec une gestion locale (au sein du CRM ou au sein d'une agence) pour pallier ce problème; ce qui ne facilite pas, pour les analystes qualité, la collecte des enregistrements en retour et l'analyse.

Emails non aboutis (Hard Bounce)

Les campagnes d'emailing de Hewlett-Packard sont gérées centralement par une agence externe. De ce fait, il nous est assez facile de récupérer le fichier de campagne que cette agence publie, chaque mois, pour tous les pays, sur le site intranet de Hewlett-Packard. Ce fichier de campagne comprend l'identifiant du correspondant contacté, son nom, son prénom, sa fonction, la société pour laquelle il travaille ainsi que son opt-in, c'est-à-dire son accord à recevoir des communications marketing de la part de Hewlett-Packard sur sa boîte email. Ce fichier comprend aussi le code campagne, et surtout le statut de l'envoi. C'est en analysant ce statut que nous séparons les emails non aboutis en « soft bounces » et « hard bounces » (voir lexique ci-dessous).

Cependant, il est à noter que le statut « hard bounce » n'est pas nécessairement fiable. Par exemple, nous avons pu observer, sur une période d'étude de 3 mois, que des contacts non aboutis « hard bounce » peuvent très bien être aboutis le mois suivant. Ce qui nous a incités à créer un compteur de « hard bounces ». Ainsi, un contact doit être comptabilisé trois fois de suite (ou plus ou moins) en statut « hard bounce » pour être vraiment considéré comme tel.

Faux téléphones

Les campagnes de télémarketing de Hewlett-Packard sont gérées différemment selon les pays et selon le type de client ou le type de campagne. Parfois, ce sont les centres d'appels internes qui sont utilisés, parfois ce sont des centres d'appels locaux ou régionaux. Cette diversité est source de problèmes dans la collecte des informations concernant les faux téléphones. En effet, chaque centre d'appel a son propre script d'appel, sa propre codification et dès lors, il est très difficile d'interpréter avec certitude la signification du statut « faux téléphone ». Certains centres d'appel vont en effet qualifier comme tel un téléphone non abouti au bout de 3 à 5 tentatives, alors que d'autres ne vont utiliser ce même statut que pour un téléphone ayant une sonnerie erronée; d'autres enfin vont faire la distinction entre « société ayant déménagé – le téléphone n'est donc plus le bon » et « téléphone à la sonnerie erronée ». Cette absence de standards quant aux différents codes ou statuts entrave la bonne analyse et la bonne interprétation. L'harmonisation des codes retour est donc clairement une piste d'amélioration pour l'avenir.

Contacts Obsolètes

Les contacts obsolètes sont dans notre cas collectés lors des campagnes de télémarketing, c'est à dire d'appels sortants. Ainsi que mentionné précédemment, dans la rubrique faux téléphones, notre principal souci est le manque de cohérence entre les codes retour et les statuts « contacts obsolètes ». D'aucuns y mettront un contact qui a quitté la société dans le cadre d'une retraite, d'un licenciement, d'un décès ou d'une démission et d'autres y mettront aussi un contact qui a quitté la fonction pour laquelle il est enregistré dans la base de données mais qui a été promu à une autre fonction au sein de la même société.

Lors de la mesure du dernier trimestre, nous avons exploré une piste qui consiste à croiser les informations avec des informations provenant de bases externes. Certains fournisseurs sont en effet en mesure de fournir, plus ou moins précisément, des informations concernant les contacts obsolètes.

Enfin, nous explorons également l'idée d'analyser une population de contacts obsolètes pour voir si certains facteurs expliqueraient l'obsolescence du contact, ce qui nous permettrait d'aboutir à un modèle de prédiction du taux de contacts obsolètes dans une population définie.

4.2 Observations de nos mesures

Avec le recul d'un an de mesure de ces 4 indicateurs a posteriori, nous constatons que notre principal problème se pose au niveau du taux de faux téléphones. Nous étudions actuellement diverses pistes pour améliorer la qualité du téléphone, car nous savons par ailleurs qu'un mauvais téléphone coûte très cher à Hewlett-Packard, non pas tant en coûts directs, mais davantage en perte d'opportunités, en perte de revenu.

Conclusion

La publication de ce référentiel avec l'AFNOR est une première marche vers une norme. Notre objectif est maintenant :

- Que ce référentiel soit connu, reconnu à la fois par les professionnels du CRM, les utilisateurs et utilisé. Parvenir à un langage commun, à des indicateurs communs correspond maintenant à un travail de communication de ce référentiel, et de co-optation.
- L'étape suivante est de publier des niveaux de bonne conduite : avec le recul d'utilisation de ces indicateurs, nous avons pu voir apparaître des tendances, des niveaux de bons résultats et à contrario des taux qui sont très mauvais. La difficulté est d'avoir suffisamment d'expérimentations, de cadres d'utilisation pour produire des niveaux qui aient un sens : nous avons pu observer que :
    - Ces taux varient beaucoup d'un pays à l'autre : par exemple, en B2B, les fichiers disponibles sur le marché sont de qualité très inégale. En Italie ou en Espagne, sur des fichiers généralistes, un taux de 25% de faux téléphones est courant. En revanche, en Allemagne, un taux de 10% est plus usuel.
    - Les taux varient également selon le segment d'entreprises : des grandes entreprises, mieux suivies, avec plus de valeur et de potentiel, auront des taux qualité bien meilleurs que des Petites et Moyennes Entreprises (PME).
- Enfin, dans le cadre de l'application chez Hewlett-Packard, une autre démarche a été initiée. Chaque faux téléphone, adresse fausse a des conséquences financières : directes par les coûts supplémentaires engendrés (courriers en trop, tarifs postaux supplémentaires,…), mais également par des opportunités ratées. Il s'agit de valoriser ce coût de la non qualité et de pouvoir déclencher des actions de correction, enrichissement selon leur rentabilité prévue.

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

Bounced Emails: on distingue les soft bounces, non aboutis temporaires (la boîte email du correspondant est pleine par exemple) des hard bounces, non aboutis permanents (l'adresse email du correspondant est fausse par exemple) ; c'est en fait la population des emails non aboutis

Emailing : campagne marketing par envoi d'email

NPAI : N'habite Pas à l'Adresse Indiquée

Références

AFNOR, A.I.D. (2006), *Mesure de la qualité des données CRM – Référentiel d'indicateurs CRM*. Accord AC X50-111

Berti-Equille, L. (2005), *Journées CRM & Qualité des Données au CNAM – Qualité des données multi-sources : un aperçu des techniques issues du monde académique.*

Hashemi Ray R., Talburt John R., (2006) *Significance Test for the Talburt-Wang Similarity Index* – The 11th ICIQ (ICIQ-06, MIT IQ) Conference Program

Madnick S., Wang R. et Zhang W. (2002), *A Framework for Corporate Householding pp. 36-46*. Proceedings of the 7th International Conference on Information Quality.

Talburt John R., Hess Kimberly, Wang Richard, Kuo Emily, (2004), *An Algebraic Approach to Quality Metrics for Customer Recognition Systems* – ICIQ-04 (MIT IQ) Conference Program.

Wang, R. (2000), *Data Quality*. Kluwer Academic Publishers

Summary

This article published by AFNOR, presents the A.I.D. dictionary of CRM data quality measurement indicators. It explains why the dictionary was created and for which users, and gives the underlying methodology to help find clients and prospects. The second part of the article is a practical case study based on Hewlett Packard in an international context.

Simuler et épurer pour extraire les motifs sûrs et non redondants

Martine Cadot*
Alain Lelu**,***

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

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

1 Introduction : problématique, principes de notre approche

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

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

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

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

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

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

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

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

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

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

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

- L'indépendance totale est définie par un effectif "fictif", obtenu à partir du produit des probabilités des marges ; ainsi dans le cas de 3 variables avec un effectif total N et des sommes marginales $n_{i..}$, $n_{.j.}$, $n_{..k}$: $\hat{n}_{ijk} = Np_{ijk}$, où $p_{ijk} = p_i p_j p_k$; $p_i = n_{i..}/N$, $p_j = n_{.j.}/N$, etc. Cette indépendance est inconditionnelle en ce sens qu'elle intègre l'indépendance de toutes les combinaisons de niveau inférieur. L'écart à cette hypothèse a été exploré par Brin et al. (1997) dans notre présent cadre d'extraction de motifs et de règles d'association : ces auteurs présentent avec prudence leur approche, qui utilise la distance à l'hypothèse d'indépendance (le classique test du Khi-deux), en indiquant clairement ses limites : *For association rules, these [validity] conditions [of the Khi2 test] will frequently be broken [...] The solution of the problem is to use an exact calculation for the probability, rather than the Khi2 approximation. The establishment of such a formula is still, unfortunately, a research problem [...].*
- L'indépendance des variables d'un motif de niveau k peut n'être que conditionnelle, c'est à dire se produire en sus de dépendances (ou non) au sein des sous-motifs de niveau inférieur. Dans le présent travail, nous nous proposons de repérer les seules situations d'écart fort et incontestable à ce type d'indépendance, laissant ouverte la caractérisation fine de la situation d'indépendance.

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

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

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

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

3 Algorithme et indicateurs MIDOVA

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

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

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

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

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

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

3.1 Exemple de recherche de l'intervalle de variation

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

sujets	A	B	C	eff
s1, s2, s3, s4, s5, s6, s7, s8, s9, s10, s11,s12, s13, s14, s15	x	x	x	15
s16, s17, s18, s19, s20	x	x		5
s21, s22, s23, s24, s25, s26, s27, s28, s29, s30, s31, s32	x		x	12
s33, s34, s35, s36, s37, s38, s39		x	x	7
s40 s41, s42	x			3
s43		x		1
s44, s45, s46, s47, s48, s49			x	6
s50, s51, s52, s53, s54, s55, s56, s57, s58, s59, s60				11

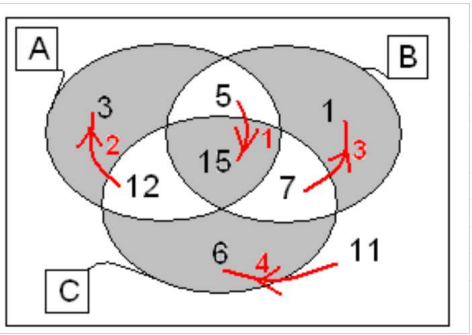

FIG. 1 – *Répartition de 60 sujets selon 3 propriétés.*

lignes, et le diagramme de Venn 8 zones. La zone où les trois propriétés sont simultanément vérifiées est grisée, ainsi que les autres zones dont le nombre de propriétés est impair, donc ici vérifiant une seule propriété. Les quatre zones restantes (blanches) sont celles où un nombre pair de propriétés (0 ou 2) sont vérifiées. Pour chercher l'intervalle de variation du support de ABC, à support constant de ses sous-motifs, on essaie d'abord d'augmenter ce support. En suivant la flèche 1, on déplace un sujet (par exemple $s16$) en lui ajoutant la propriété C qu'il n'avait pas. Il passe ainsi d'une zone non grisée à une zone grisée. Lors de ce déplacement, le support de AB ne change pas. Par contre ceux de AC et de BC augmentent d'une unité chacun. On compense cette augmentation en suivant les flèches 2 et 3, qui déplacent par exemple les sujets $s21$ et $s33$ en leur retirant la propriété C. Ce déplacement a pour conséquence une diminution du support de C d'une unité, qu'on compense en déplaçant par exemple le sujet $s50$ selon la flèche 4. Ainsi, si on désire augmenter le support de ABC, qui est dans une zone grisée, sans modifier les supports de ses sous-motifs, il faut augmenter d'autant les effectifs des 3 autres zones grisées et diminuer d'autant chaque effectif d'une zone non grisée. Comme le plus petit effectif des zones non grisées est 5, le support de ABC ne peut pas augmenter de plus de 5. Et pour faire diminuer le support de ABC, on procède de façon inverse, ce qui fait qu'il ne peut pas diminuer de plus de 1, minimum des effectifs des zones grisées. Le support de M varie ainsi entre 14 et 20, sa valeur centrale étant 17. Il ne reste plus qu'à calculer g en remplaçant le support s par 15, la longueur L par 3, et le centre c par 17, ce qui donne $g = 2^{L-1}(s - c) = 4(15 - 17)$, soit -8, ce qui veut dire qu'il faut déplacer 8 sujets pour faire passer le support de ABC de 17 à 15 sans changer les supports de ses sous-motifs. En résumé :

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

3.2 L'algorithme MIDOVA

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

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

4 Application

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

4.1 Le test de significativité des 2-motifs

Nous nous sommes d'abord intéressés aux associations de deux mots. Sur la base de 888 mots, il y en a 888*887/2 soit 393 828. La seule information pertinente que nous avons retenue sur ces associations de deux mots est le nombre de documents dans lesquels ils apparaissent simultanément. Nous avons décidé de ne garder que les associations de deux mots (2-motifs)

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

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

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

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

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

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

5 Conclusion : efficacité de la chaîne de traitement proposée

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

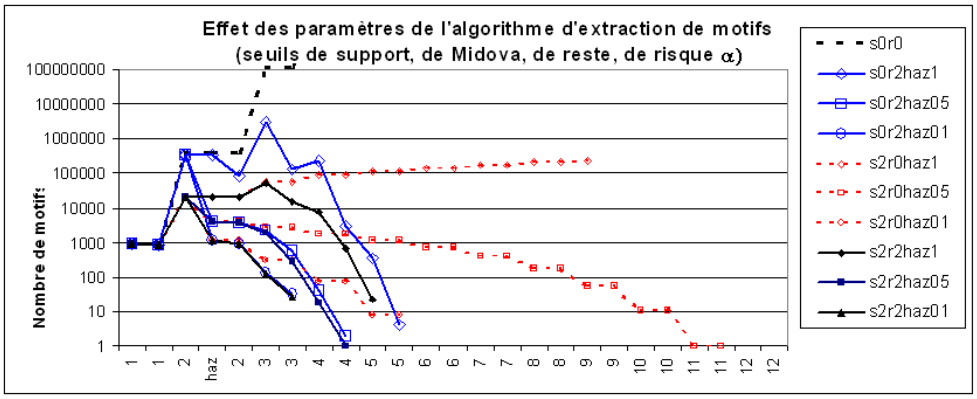

FIG. 2 – *Comparaison des efficacités respectives de l'utilisation combinée de Midova et du test d'échanges en cascade avec la méthode classique de type A Priori, avec seuil de support, où on trouve en abscisse les longueurs des motifs et en ordonnée leur nombre. Chaque longueur est présente deux fois,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

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

Références

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo (1996). *Advances in Knowledge Discovery and Data Mining*, Chapter Fast Discovery of Association Rules, pp. 307–328. Menlo Park, California : AAAI Press /MIT Press.

Bastide, Y. (2000). *Data mining : algorithmes par niveau. techniques d'implantation et applications*. Ph. D. thesis, Université Blaise Pascal, Clermont-Ferrand.

Bavaud, F. (1998). *Modèles et données : une introduction à la Statistique uni-, bi- et trivariée*. Paris ; Montréal (Qc) : L'Harmattan.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : Generalizing association rules to correlations. In J. Peckham (Ed.), *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15*, Tucson, Arizona, USA, pp. 265–276. ACM Press.

Cadot, M. (2005). A simulation technique for extracting robust association rules. In *CSDA 2005*, Chypre.

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Ph. D. thesis, Université de Franche-Comté.

Cadot, M. et A. Napoli (2003). Une optimisation de l'extraction d'un jeu de règles s'appuyant sur les caractéristiques statistiques des données. Volume 16, pp. 631–656.

Cuxac, P., M. Cadot, et C. François (2005). Analyse comparative de classifications : apport des règles d'association floues. In *EGC 2005*, pp. 519–530.

Gras, R. (1979). *Contribution à létude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. Ph. D. thesis, Université de Rennes I.

Guillet, F. (2004). Mesure de qualité des connaissances en ecd.

Jakulin (2003). Attribute interactions in machine learning. Master's thesis, University of Ljubljana, Slovenija.

Jensen, D. D. et P. R. Cohen (2000). Multiple comparisons in induction algorithms. *Machine Learning 38*(3), 309–338.

Lelu A., Tisseau-Pirot A.-G., A. A. (1997). Cartographie de corpus textuels évolutifs, un outil pour l'analyse et la navigation. *Hypertextes et Hypermédias 1*.

Lenca, P., P. Meyer, P. Picouet, et B. Vaillant (2003). Aide multicritère à la décision pour évaluer les indices de qualité des connaissances. In *EGC 2003*, pp. 271–282.

Manly, B. (1997). *Randomization, Bootstrap and Monte Carlo methods in Biology.* Texts in Statistical Science. Boca Raton, Florida, USA : Chapman & Hall/CRC.

Press, J. S. (2004). *Statistical Data Mining and Knowledge Discovery*, Chapter The role of Bayesian and frequentist multivariate modeling in statistical Data Mining, pp. 309–338. Boca Raton, US : Chapman & Hall/CRC.

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

Summary

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

Nouvelle méthode d'extraction de règles de classification multi-labels

Abdelhamid Zemirline, Laurent Lecornu, Basel Solaiman

Departement ITI, ENST Bretagne
abdelhamid.zemirline@enst-bretagne.fr,
laurent.lecornu@enst-bretagne.fr,
basel.solaiman@enst-bretagne.fr

Résumé. Les études concernant les méthodes d'extraction de règle de classification proposent de classer une instance, le plus souvent en se basant sur une seule règle, sans prendre en considération les règles du même type ayant un label de classe différent (par ex. : $X \rightarrow a$, $X \rightarrow b$). Dans certains domaines, comme par exemple le domaine médical, il est important de connaître toutes les conséquences induites par une règle donnée. Dans cette étude, nous proposons une nouvelle méthode d'extractions de règle de classification multi-labels. Elle se caractérise par sa rapidité, ceci est dû à la procédure utilisée pour l'extraction des motifs fréquents par le biais des ensembles flous. Elle se caractérise également par l'extraction des règles en fonction de leur label, les règles sont extraites à partir des ensembles de données appartenant à une même classe et non à partir de l'ensemble global des données. Cette approche permet d'extraire des règles spécifiques à chaque classe et d'extraire de nouvelles catégories de règle d'exception. La méthode développée présente une grande précision et ses performances sont comparables à celles d'autres méthodes de classification de références.

1 Introduction

Dans le domaine médical, les systèmes d'aide au diagnostic ont pour objectif de donner un ou plusieurs diagnostics au praticien, ceci induit l'utilisation de règle multi-label, c-à-d des règles à plusieurs conséquences. Il existe peu d'études sur les méthodes qui génèrent des règles multi-label (Thabtah et al. (2004, 2006)). Par contre, il y en a un grand nombre d'études qui portent sur des méthodes de génération de règle de classification avec un seul label (Li et al. (2001); Liu et al. (1998)). Dans cette étude, nous proposons une méthode d'extraction de règles de classification multi-labels, qui s'appuie sur une nouvelle approche d'extraction. Cette approche consiste à extraire les règles à partir d'un ensemble d'instances appartenant à une même classe. Une instance est un ensemble d'attribut de type alpha-numérique, un de ces attributs indique l'appartenance de l'instance à une classe. Cette approche permet d'obtenir des règles spécifiques pour chaque classe que contient la base d'apprentissage. Contrairement aux méthodes classiques qui génèrent des règles globales, extraites à partir de l'ensemble de la base d'apprentissage. De plus, notre approche permet d'extraire des règles dites d'exception.

Par exemple dans le domaine médical, ce type de règle permet aux systèmes d'aide au diagnostic d'identifier des pathologies qui sont difficiles à diagnostiquer par le praticien en raison de leur faible fréquence d'apparition et de permettre également de distinguer une pathologie qui présente un symptôme similaire à une pathologie courante. Dans cette étude, nous commençons par définir les règles de classification. Par la suite, nous présentons notre algorithme et sa description. Avant de conclure, notre méthode est comparée à d'autres algorithmes de classification.

2 Règles d'association de classes

La méthode des règles d'association de classes (Li et al. (2001); Liu et al. (1998)) comprend l'extraction des règles et la classification à l'aide de ces règles. Cette méthode définit un classifieur du type C4.5 de Quinlan (1993). Ce type de méthodes dérive des méthodes d'extraction de règle d'association de Agrawal et Srikant (1994). Le problème de règle d'association de classes se définit de façon suivante : soit D l'ensemble de données qui est composé de N cas décrits par un ensemble d'items. Soit I l'ensemble des items qui décrivent les cas de l'ensemble D. Chaque élément de D est associé à un élément de l'ensemble C des classes. Les règles d'association de classe sont implémentées de la façon suivante $X \rightarrow c$ où $X \subseteq I$ et $c \in C$. X est un itemset (c-à-d ensemble d'items). La règle $X \rightarrow c$ est extraite de l'ensemble D avec une confiance ($Conf$). $Conf$ est le rapport entre le nombre de cas de D contenant X et appartenant à c et le nombre de cas de D contenant X. La règle $X \rightarrow c$ a un support ($Supp$) dans D. $Supp$ est le rapport entre le nombre de cas de D contenant X et appartenant à c et le nombre de l'ensemble des cas de D. L'objectif de cette approche est de générer des règles d'association de classes qui ont un support et une confiance supérieurs à un certain seuil ($minSupp$, $minConf$) et de mettre en place un classificateur efficace à partir de ces règles.

Dans un problème multi-labels, on retrouve la même définition que dans le problème de règles d'association de classe. Cependant, pour une instance $d \in D$, il peut être assigné un ensemble de classes $c_1, c_2, \cdots, c_k$ pour $c_i \in C$. Ceci est représenté de la façon suivante : $(d, (c_1, c_2, \cdots, c_k))$ où $(c_1, c_2, \cdots, c_k)$ est une liste ordonnée de classes pour l'instance d. L'ordonnancement de la liste se fait à l'aide de la base de connaissance déduite du classificateur.

3 Algorithme MCCAR

Dans cette partie, nous présentons un nouvel algorithme d'extraction de règles d'association de classes que nous avons appelé l'algorithme MCCAR (Multi-label Classification based on Class Association Rule). Les particularités de notre algorithme sont les suivantes :

- Extraction de règles à partir d'un sous-ensemble d'instances appartenant à une même classe et non pas sur l'ensemble des instances,
- classification des instances en se basant sur plusieurs règles,
- extraction de règles d'exception par deux procédures (la procédure "multi-seuil" et la procédure "générer-supprimer") §3.2,
- génération de règles multi-labels.

La majorité des méthodes d'extraction de règles d'association de classes se décomposent en trois phases. La première phase consiste à générer les règles, la seconde à les évaluer et la dernière phase à utiliser les règles générées sur les instances de la base de test. Dans notre approche, nous avons quatre phases. La première phase regroupe les instances dans des sous-ensembles en fonction de leur classe et définit le support minimum pour chaque sous-ensemble. Cela permet par la suite de découvrir les items fréquents et de générer des règles dites communes et d'exception. Dans notre approche, les règles sont extraites à partir des ensembles de données appartenant à une même classe et non à partir de l'ensemble global des données. Ceci nous amène à redéfinir le calcul du support, $Supp$: le rapport entre le nombre de cas de D contenant X et appartenant à c et le nombre de l'ensemble des cas de D contenant c. La procédure permettant de générer les seuils est expliquée dans le §3.1. La génération des règles d'exception est abordée dans le §3.2. La seconde phase consiste à supprimer les instances de l'ensemble de la base d'apprentissage, qui sont couvertes par les règles générées précédemment et à générer les règles d'exception à l'aide de la procédure "générer-supprimer" (generate-and-remove). La procédure "générer-supprimer" est décrite dans le §3.2. Dans la troisième phase, les règles générées dans les phases précédentes sont évaluées afin de permettre leur exploitation par la procédure de classification. La procédure de cette phase est décrite dans §3.3. Dans la dernière phase, toutes les règles générées précédemment sont regroupées afin de produire des règles multi-labels (décrite dans §3.4). Par la suite, une procédure de classification des instances de la base de test est réalisée afin d'estimer les performances de l'algorithme. Pour réaliser cette procédure, un classificateur est mis en place. Ce classificateur manipule les règles afin de prédire la classe de l'instance à prédire (§3.5).

3.1 Sélection des itemsets fréquents et génération de règles

Plusieurs approches de sélection des itemsets fréquents et de génération de règles ont été proposées par Agrawal et Srikant (1994); Liu et al. (1998); Li et al. (2001); Yin et Han (2003); Thabtah et al. (2004); Zaki et al. (1997); Han et al. (2000); Quinlan et Cameron-Jones (1993).

Dans ces travaux, nous proposons une nouvelle approche afin de découvrir les itemsets fréquents et de générer des règles. Cette approche génère des règles dites locales, c-à-d que ces règles sont générées à partir d'un sous-ensemble d'instances qui appartiennent à une même classe. Cela permet d'avoir des règles spécifiques pour chaque classe, contrairement aux autres méthodes qui génèrent leurs règles à partir de l'ensemble des instances. Dans notre méthode, nous nous inspirons de la méthode d'intersection proposée par Zaki et al. (1997) afin de parcourir une seule fois l'ensemble des instances pour calculer les supports des règles générées.

Notre approche divise la base d'apprentissage T en plusieurs sous-ensembles qui regroupent les instances appartenant à une seule et même classe. Chacun d'entre eux est parcouru une seule fois (grâce à la méthode d'intersection) afin de calculer la fréquence d'apparition des items. Pour chaque classe, il est défini une fonction d'appartenance en fonction des fréquences des items.

Ces fonctions d'appartenance classent les items en deux sous-ensembles : les items qui appartiennent au sous-ensemble *rare* et les autres au sous-ensemble *fréquent*. Nous faisons appel à la théorie des ensembles flous de Zadeh (1975) afin de définir les fonctions d'appartenance aux sous-ensembles *rare* et *fréquent*. La plupart des méthodes de règles d'association ont leur support seuil qui est fixé au préalable, ceci peut présenter certains désavantages tels que :

– Aucun item n'a son support qui dépasse le support minimum.

– Certains items ne sont pas pris en compte malgré que leur support sont voisins au support minimal.

Les termes "rares" et "fréquents" définissent une variable linguistique, celle-ci est caractérisée par le quintuplet $(x, T(x), U, G, M)$ (Zadeh (1975)) où :

– x est le nom de la variable linguistique, ici c'est la fréquence ;
– $T(x)$ est l'ensemble des termes associés à la valeur linguistique, ici cet ensemble est : *{ Rare, Fréquent}* ;
– U est l'espace de définition, notre domaine de définition $U = [f_{min}, f_{max}]$ (f_{min} correspond à la fréquence minimale d'apparition d'un item qui est normalisée par la fréquence maximale (max_{freq}) d'apparition d'un item pour la même classe, f_{max} correspond à la fréquence maximale (max_{freq}) normalisée, elle équivaut à 1) ;
– G est une règle syntaxique pour la génération des noms de valeur de x ;
– M est une règle sémantique pour associer à chaque valeur sa signification.

Les termes de $T(x)$ sont caractérisés par des ensembles flous définis par les fonctions d'appartenance suivantes :

– F est l'ensemble des centroïdes des ensembles flous obtenus par l'algorithme de fuzzy c-means (FCM) Bezdek (1981), tel que F se présente de la façon suivante $\{f_{rare}, f_{fréquent}\}$.
– $\mu_{i,rare}$ correspond à la fonction d'appartenance au terme linguistique *rare* pour la classe i et prend comme argument la fréquence normalisée f d'un item donné.

$$\mu_{i,rare}(f) = \begin{cases} 1 & \text{si } f \leq f_{rare} \\ 1 - (f - f_{rare})/(f_{fréquent - f_{rare}}) & \text{si } f_{rare} < f \leq f_{fréquent} \\ 0 & \text{sinon.} \end{cases}$$

– $\mu_{i,fréquent}$ correspond à la fonction d'appartenance au terme linguistique *fréquent* pour la classe i et prend comme argument la fréquence normalisée f d'un item donné.

$$\mu_{i,fréquent}(f) = \begin{cases} 0 & \text{si } f \leq f_{rare} \\ 1 - (f - f_{fréquent})/(f_{fréquent - f_{rare}}) & \text{si } f_{rare} < f \leq f_{fréquent} \\ 1 & \text{sinon.} \end{cases}$$

Pour notre méthode, les items sont considérés comme *fréquent* si le degré d'appartenance à l'ensemble *fréquent* est supérieur au degré d'appartenance au sous-ensemble *rare*. Deux seuils subjectifs sont déduits : premier support sup_1 et le second support sup_2. La valeur de sup_1 correspond à l'abscisse de l'intersection des fonctions μ_{rare} et $\mu_{fréquent}$ pour une classe donnée, multipliée par max_{freq}. La valeur de sup_1 correspond à la valeur du centroïde f_{rare} multiplié par max_{freq}. Grâce à cette méthode, des seuils supports subjectifs spécifiques à une classe donnée ont été définis. Tous les items qui ont leur degré d'appartenance à la valeur linguistique *fréquent* égal à 1 sont considérés comme des items fréquents. Les items qui ont leur degré d'appartenance à la valeur linguistique *fréquent* inférieur à 1 mais supérieur au degré d'appartenance à la valeur linguistique *rare* sont considérés comme items d'exception. La combinaison des items fréquents et d'exception va générer des règles dites communes ou d'exception selon les supports de ces règles, ceci sera précisé par la suite.

3.2 Extraction des règles dites d'exception

Tout d'abord, les règles d'exception sont définies comme étant des règles qui contredisent la croyance commune. Elles sont souvent inconnues ou négligées. Cependant, ce type de règles

permet de couvrir des cas rares. Une règle commune présente un phénomène commun avec un support et une confiance élevés. La règle d'exception contredit certaines règles dites communes et a un faible support. Cependant, elle possède une valeur de confiance aussi élevée que pour les règles dites communes.

Il existe deux types de méthodes qui permettent la découverte des règles d'exception à partir d'ensemble de règles d'association (Hussain et al. (2000); Liu et al. (1999b,a)) : la méthode à multi-supports et la méthode générer-supprimer. La méthode à multi-supports consiste à introduire deux supports minimaux. Les règles ayant leurs items entre ces deux seuils sont les règles d'exception. La méthode générer-supprimer consiste à générer un ensemble de règles R à partir d'une base d'apprentissage T à l'aide d'un algorithme d'induction, puis à supprimer les objets T' qui sont couverts par les règles R de T. Un nouvel ensemble de règles R' est généré à partir de $T - T'$ à l'aide du même algorithme d'induction. Ce processus est répété jusqu'à qu'il n'y ait plus d'élément dans la base d'apprentissage.

Les règles de classification s'écrivent de la façon suivante : $X \rightarrow c$; ceci implique une légère différence dans la définition des règles d'exception, d'où nous proposons une nouvelle définition de règles d'exception. Rappelons rapidement que l'objectif des règles de classification est de définir un classificateur afin de prédire l'appartenance à une classe de certains nouveaux objets. Les règles d'exception (pour les règles de classification) sont des règles qui indiquent l'appartenance d'un item à une classe en contradiction avec une règle qui a un poids plus fort. Elle couvre des instances qui sont inhabituelles ou ambiguës. Dans notre étude, nous définissons trois structures de règles d'exception :

Unique règle d'exception : l'itemset X de ce type de règle n'est présent que dans cette règle. La confiance pour cette règle est égale à 1. Les items qui composent X ne sont pas fréquents et la règle est générée par la méthode générer-supprimer.

$X \rightarrow c$ - règle d'exception
(support faible, confiance = 1)
$X \rightarrow \neg c$ - $\emptyset$

Pseudo règle d'exception : l'itemset X de ce type de règle est composé d'items non fréquents pour la classe c et cet itemset X apparait dans d'autres règles appartenant à des classes différentes de c.

$X \rightarrow c$ - règle d'exception
(support faible, confiance faible/élevée)
$X \rightarrow \neg c$ - règle commune
(support élevé, confiance élevée)

Semi règle d'exception : l'itemset X de ce type de règle est composé en partie d'items qui ne sont pas fréquents pour la classe c. L'itemset X peut composer d'autres règles d'exception pour des classes différentes de c.

$X \rightarrow c$ - règle d'exception
(support faible, confiance élevé)
$X \rightarrow \neg c$ - règle d'exception
(support faible, confiance faible)

Cette définition des règles d'exception permet à l'utilisateur d'évaluer les règles de faible poids d'une façon plus pragmatique et de les intégrer dans le processus de décision de la même façon que les règles dites communes. Dans notre méthode, nous combinons la méthode

générer-supprimer et la méthode à double seuils pour générer les règles d'exception. La méthode de double seuils est adaptée à notre approche car elle peut utiliser de manière aisée les fonctions d'appartenance définies dans la phase une. Cette adaptation consiste à définir des seuils de supports locaux, c-à-d chaque classe aura ses propres double supports. Dans notre algorithme, la méthode à double supports minimaux est appliquée dans la phase une et quant à la méthode générer-supprimer, elle est appelée dans la phase deux. Cette combinaison des deux méthodes permet d'extraire des règles d'exception de façon pertinente.

3.3 Evaluation des règles

Dans cette partie, nous allons parler d'évaluation des règles générées à partir d'un sous-ensemble d'instances spécifiques à une classe donnée. Cette évaluation permet par la suite de classer les règles par degré d'importance. Notre démarche consiste à évaluer les règles en fonction de leur importance dans la classe à laquelle elles appartiennent. Ceci est effectué par l'attribution de degrés d'appartenance des règles à la classe à laquelle elles appartiennent. La définition des fonctions d'appartenance s'inspire de la méthode de l'histogramme de fréquence normalisée proposée par Chauvin (1995). À partir des supports des règles d'une classe donnée, nous définissons un histogramme de fréquence normalisé par rapport à la fréquence la plus élevée. La raison qui nous motive à définir des fonctions d'appartenance pour évaluer les règles entre elles au lieu de prendre en considération la confiance de règle est la suivante : notre algorithme propose une classification à base de multi-règles, qui nécessite la combinaison de plusieurs règles afin de classer un objet. Or, la combinaison des degrés d'appartenance est plus significative que la combinaison des confiances de règles appartenant à une même classe.

3.4 Ordonnancement des règles multi-labels

Une règle multi-labels est le regroupement de règles ayant leur itemset similaire appartenant à des classes différentes d'où l'appellation règle multi-labels. Les différentes classes qui composent ce type de règles sont ordonnancées. $A \prec B$ signifie que la classe A précède la classe B. La règle multi-labels $X \rightarrow < A, B >$ sous entend que $A \prec B$. Cela signifie que le degré d'appartenance de l'itemset X à la classe A est plus important que celui à la classe B. Si l'itemset X a le même degré d'appartenance pour la classe A et B, c'est le nombre d'instances couvert par les règles $X \rightarrow A$ et $X \rightarrow B$ qui va déterminer l'ordonnancement entre ces deux classes.

3.5 Classification

La classification pour les méthodes de règles d'association de classes consiste à utiliser les règles générées afin de déterminer la classe d'une nouvelle instance. L'approche la plus utilisée est celle qui classe les règles selon leur confiance (Liu et al. (1998); Thabtah et al. (2004)). Cette approche attribue à l'instance la classe de la règle ayant la plus forte confiance. Si aucune règle n'est trouvée une classe par défaut est attribuée. Le classement des règles doit respecter certains critères, exemple de critères : soient r_a et r_b deux règles, r_a précède r_b :

1. si r_a a une confiance supérieure à r_b, sinon
2. si r_a et r_b ont une même confiance mais r_a a un support supérieur, sinon

3. si r_a et r_b ont leur confiance et leur support identiques mais r_a doit être générée en premier.

La classification des règles se présente de la façon suivante : $< r_1, r_2, \cdots, r_n, c >$ où r_i précède r_{i+1} et c est la classe par défaut.

Dans Yin et Han (2003) une approche est proposée qui utilise l'ensemble des meilleures règles qui couvrent l'instance à classer. Ils se sont basés sur :

- La récupération des règles qui couvrent l'instance à classer.
- La sélection des meilleurs règles.
- Le regroupement des règles selon leur classe et la mesure de leur combinaison.
- La sélection de la classe qui présente la meilleur mesure.

En effet, il y a été développé plusieurs méthodes de combinaison de règles appartenant à une même classe. Une d'elle consiste à calculer la moyenne des confiances des règles. Une autre consiste à utiliser la règle ayant la plus forte mesure comme représentative. Par exemple dans Li et al. (2001) la règle ayant la plus forte mesure est celle qui présente la valeur la plus élevée au test χ^2.

Pour notre méthode, nous définissons la mesure utilisée en calculant la moyenne géométrique des degrés d'appartenance des meilleures règles à une classe donnée. L'utilisation de plusieurs règles permet d'augmenter la précision de prédiction. Cependant, s'appuyer sur une seule règle pour prédire une instance, c'est négliger toutes les autres règles qui couvrent cette même instance. De plus cela ne prend pas en considération des informations qui peuvent donner un poids de certitude ou d'incertitude dans la prédiction. Dans notre cas, nous prenons les meilleures règles de chaque groupe pour la prédiction d'une instance, cela nous évite de prendre les règles qui couvrent l'instance mais dont la confiance est faible et qui ne peuvent que diminuer fortement la mesure.

4 Evaluation de l'algorithme MCCAR

Dans cette partie, nous allons évaluer les performances de notre système, mais avant de commencer la présentation de l'expérience, nous précisons que son objectif n'est pas de montrer que notre approche est plus performante que d'autres mais de monter que cette approche a sa place parmi les méthodes de classification. Cependant, la vocation de notre approche ne se limite pas seulement à définir un classificateur mais à apporter de nouvelles connaissances (règles multi-labels, règles d'exception) qui permettent à un système d'aide au diagnostic d'être plus fiable, plus informatif et transparent pour l'utilisateur.

Nous mettons en place une étude comparative de notre méthode avec d'autres méthodes de classification de référence telles que RIPPER Cohen (1995), PART Frank et Witten (1998) et CBA Liu et al. (1998). Comme pour Cohen (1995); Frank et Witten (1998); Liu et al. (1998); Thabtah et al. (2004), 26 bases de UCI Machine Learning Repository Merz et Murphy (1996) sont utilisées comme base de références afin d'évaluer les différentes méthodes. La méthode de la validation croisée à 10 parties est utilisée sur ces 26 bases de cas. L'expérience consiste à prendre chaque base de cas et à la diviser en deux parties : une partie qui sera considérée comme base d'apprentissage et une autre comme base de test. Ces bases sont appliquées aux méthodes de classification citées ci-dessus et à notre méthode, et par la suite, nous comparons les taux de bonne estimation dans la classification des cas de la base de test des différentes méthodes. Nous avons fait appel au logiciel Weka (2006) afin d'exécuter les algorithmes RIPPER et

Dataset	RIPPER	PART	CBA	MCCAR
anneal	98.1	98.32	93.43	96.5
austra	86.81	85.79	84.78	83.15
autos	73.17	75.6	64.25	86.3
breast	95.42	96.28	96.86	97.14
cleve	82.17	81.18	75.86	79.64
crx	86.08	85.5	85.07	83.5
diabetes	76.82	77.21	74.42	74.60
german	72.4	73.4	70.97	71.2
glass	72.42	73.81	65.76	64.21
heart	84.07	82.22	81.8	81.11
hepati	79.35	81.93	84.97	85.06
horse	84.78	84.51	82.86	80.95
hypo	99.08	99.24	97.43	95.25
iono	90.88	91.73	95.73	91.15
iris	94	95.33	93.25	96
labor	84.12	84.21	94.99	94.66
led7	69.7	73.56	67.01	72.93
lymph	77.7	80.4	83.4	71.78
pima	66.14	65.1	76.65	77.59
sick	93.77	93.74	96.07	94.87
sonar	80.76	75.96	76.88	76.91
tic-tac	97.8	94.25	99.6	82.24
vehicle	68.32	70.56	60.22	65.1
waveform	76	76,66	73.03	73,93
wine	95.5	94.38	96.66	96.63
zoo	86.13	92.07	92.34	91.36

TAB. 1 – *Précision de la classification de différentes méthodes*

PART. L'algorithme CBA a été exécuté à l'aide d'une application qui a été fournie par CBA (1998). Pour cette expérience, nous fixons la valeur de la confiance minimale (minConf) à 0,3 et la valeur du support minimale (minSup) à 0,03 pour CBA, RIPPER et PART.

Notre système présente des résultats de taux de fiabilité très intéressants. Les résultats obtenus dans le tableau 1 montrent que notre méthode présente le même niveau de précision que les méthodes de classification de référence. Les cas où notre méthode présente des résultats inférieurs aux méthodes de références, l'écart est en moyenne pas plus de 2 points, sauf pour les bases d'apprentissage *lymph* et *tic-tac*. Pour la base d'apprentissage *tic-tac*, la mauvaise estimation que notre méthode nous renvoie est due au fait que les valeurs des attributs ont une répartition équiprobable dans l'ensemble des cas des différentes classes. Ce qui implique que notre méthode génère les même règles pour les différentes classes et ceci induit des ambiguïtés dans la classification de certaines instances. Pour la base d'apprentissage *lymph*, l'écart important de mauvaise estimation s'explique en partie par la petite taille de la base d'apprentissage.

Certaines classes ne peuvent pas déduire les items qui leur sont réellement fréquents. Il y a une autre cause à cet écart important, elle est du même type que celle décrite pour la base d'apprentissage *tic-tac*.

5 Conclusion

Dans cette étude, nous avons proposé une nouvelle méthode d'extraction de règles multi-labels. Cette approche a vocation à être intégrée dans un système d'aide au diagnostic. Elle se compose de plusieurs caractéristiques qui ne sont pas présentes dans les méthodes d'extraction de règles de classification traditionnelles telles que : (1) couverture de l'ensemble des instances, (2) un seul parcours de la base d'apprentissage, (3) extraction de règles d'exception et (4) utilisation des degrés d'appartenance pour l'évaluation des règles. Ces caractéristiques associées au fait que les performances obtenues sont comparables aux autres méthodes de règles de classification de références, montrent la puissance de l'approche proposée.
Dans la méthode présentée, nous faisons appel aux fonctions d'appartenance afin d'évaluer les règles de classification. Par ce biais, nous pouvons facilement intégrer un opérateur de fusion qui peut par la suite être introduit dans des systèmes d'aide au diagnostic. Ces opérateurs de fusion auront la tâche d'intégrer dans un seul système des règles provenant de différentes sources afin de générer un classificateur ayant des performances très élevées.
L'algorithme MCCAR présente une nouvelle approche d'extraction de règles multi-labels et en particulier l'extraction de règles d'exception. Pour les règles d'exception, il serait intéressant par la suite de mettre une procédure de filtrage, afin de mettre à l'écart certaines règles qui sont considérées comme du bruit pour un domaine donné et mettre en avant celles qui peuvent avoir un grand intérêt dans la prise de décision pour ce même domaine.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499. Morgan Kaufmann.

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. *Plenum Press*.

CBA (1998). *CBA : The DM-II system*. http ://www.comp.nus.edu.sg/ dm2/.

Chauvin, S. (1995). *Thèse : Evaluation des théories de la décision appliquées à la fusion de capteurs en image satellitaire*. Ph. D. thesis, Thèse de Doctorat d'Université, Nantes.

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pp. 115–123. Morgan Kaufmann.

Frank, E. et I. H. Witten (1998). Generating accurate rule sets without global optimization. In *Proc. 15th International Conf. on Machine Learning*, pp. 144–151. Morgan Kaufmann, San Francisco, CA.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In W. Chen, J. Naughton, et P. A. Bernstein (Eds.), *2000 ACM SIGMOD Intl. Conference on Management of Data*, pp. 1–12. ACM Press.

Hussain, F., H. Liu, E. Suzuki, et H. Lu (2000). Exception rule mining with a relative interestingness measure. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 86–97.

Li, W., J. Han, et J. Pei (2001). Cmar : Accurate and efficient classification based on multiple class-association rules. In *ICDM*, pp. 369–376.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *KDD*, pp. 80–86.

Liu, B., W. Hsu, et Y. Ma (1999a). Mining association rules with multiple minimum supports. In *Knowledge Discovery and Data Mining*, pp. 337–341.

Liu, H., H. Lu, L. Feng, et F. Hussain (1999b). Efficient search of reliable exceptions. In *PAKDD*, pp. 194–203.

Merz, C. et P. Murphy (1996). *UCI repository of machine learning databases*. Department of Information and Computer Science, University of California, Irvine,.

Quinlan, J.-R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

Quinlan, J. R. et R. M. Cameron-Jones (1993). FOIL : A midterm report. In *Machine Learning : ECML-93, European Conference on Machine Learning, Proceedings*, Volume 667, pp. 3–20. Springer-Verlag.

Thabtah, F. A., P. I. Cowling, et Y. Peng (2004). Mmac : A new multi-class, multi-label associative classification approach. In *ICDM*, pp. 217–224.

Thabtah, F. A., P. I. Cowling, et Y. Peng (2006). Multiple labels associative classification. *Knowl. Inf. Syst. 9*(1), 109–129.

Weka (2006). *weka : data mining software in java*. http ://www.cs.waikato.ac.nz/ml/weka.

Yin, X. et J. Han (2003). CPAR : Classification based on predictive association rules. In *Proceedings of 2003 SIAM International Conference on Data Mining*, San Fransisco, CA.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning - ii. *Information Sciences 8*, 301–357.

Zaki, M. J., S. Parthasarathy, M. Ogihara, et W. Li (1997). New algorithms for fast discovery of association rules. Technical report, Rochester, NY, USA.

Summary

In this paper, we propose a new algorithm of multi-labels classification MCCAR (Multi-labels Classification based on Class Association Rule). Our approach is based on producing association rules and knowledge base informing about the accuracy of rules for the classification. Our algorithm finds reliable exceptions with a simple and efficient approach. A rule has a multi-label class. To rank class labels for a rule, we use the degree of membership of rule to a class. This new approach enables us to evaluate the importance of rules among themselves.

Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association

Régis Gras*, Jérôme David*
Fabrice Guillet*, Henri Briand*

*LINA - FRE CNRS 2729
Equipe COnnaissances & Décision
Ecole Polytechnique de l'université de Nantes
La chantrerie BP 50609, 44306 NANTES Cedex 3
regisgra@club-internet.fr, jerome.david,fabrice.guillet,henri.briand@univ-nantes.fr,
http://www.polytech.univ-nantes.fr/COD

Résumé. L'analyse statistique implicative permet d'extraire d'un ensemble de données des règles de type $a \rightarrow b$ au sens où b est généralement réalisé dès lors que a l'est. Une mesure de qualité l'intensité d'implication, est associée à cette quasi-implication. Il est important d'étudier la stabilité de la règle lorsque les paramètres dont elle dépend varient sensiblement au voisinage de leur observation. Plusieurs méthodes le permettent. Nous étudions ici la stabilité de la mesure de qualité en utilisant la différentielle de l'indice fondamental qui la fonde.

1 Introduction

Le problème de la sensibilité aux faibles perturbations des paramètres en jeu, donc de la stabilité des indices de mesure de qualité des règles d'association se pose dès lors que les données sont susceptibles d'être bruitées. Trois méhodes nous semblent appropriées dans le but d'examiner la sensibilité des indices qui permettent de mesurer les qualités des règles d'association, en particulier celles de la forme $a \rightarrow b$ où a et b sont des variables observées sur un ensemble de sujets :

1. la simulation consistant à partir de fichiers plus ou moins artificiels à travers lesquels sont modifiées les paramètres intervenant dans la définition des indices (Gras, 2005) ;
2. la méhode du bootstrap consistant à effectuer des changements de certaines valeurs des paramètres tout en conservant constantes certaines d'entre elles dont l'effectif de la population de sujets ;
3. une méthode mathématique consistant à étudier par l'analyse les variations des paramètres par l'examen de leurs dérivées partielles et donc du gradient de l'indice global (Gras, 2005), (Lenca et al., 2006), (Vaillant et al., 2006).

C'est cette dernière méthode que nous retenons ici. Nous porterons notre attention de façon privilégiée sur l'indice à la base de l'Analyse Statistique Implicative (A.S.I.) et comparerons

les résultats obtenus à ceux dérivant d'autres indices retenus pour des mesures de qualité de règles.

2 Rappel sur l'Analyse Statistique Implicative

Précisons. Un ensemble fini V de v variables est donné : $a, b, c, ...$ Dans la situation paradigmatique classique, il s'agit des performances (réussite-échec) à des items d'un questionnaire. A un ensemble fini E de n sujets x, on associe, par abus d'écriture, les fonctions du type : $x \rightarrow a(x)$ où $a(x) = 1$ (ou $a(x) = vrai$) si x satisfait ou possède le caractère a et 0 (ou $a(x) = faux$) sinon. En intelligence artificielle, on dira que x est un exemple ou une instance pour a si $a(x) = 1$ et un contre-exemple dans le cas contraire (Gras, 1979) et (Gras et al., 1996).

La règle $a \rightarrow b$ est logiquement vraie si pour tout x de l'échantillon, $b(x)$ n'est nul que dans le cas où $a(x)$ l'est aussi ; autrement dit si l'ensemble A des x pour lesquels $a(x) = 1$ est contenu dans l'ensemble B des x pour lesquels $b(x) = 1$. Cependant, cette inclusion stricte n'est qu'exceptionnellement observée dans les expériences réelles. Dans le cas d'un questionnaire de connaissances, on pourrait en effet observer quelques rares élèves réussissant un item a et ne réussissant pas l'item b, sans que ne soit contestée la tendance à réussir b quand on a réussi a. Relativement aux cardinaux de E (soit n), mais aussi de A (soit n_a) et B (soit n_b), c'est donc le "poids" des contre-exemples (soit $n_{a\overline{b}}$) qu'il faudra prendre en compte pour accepter statistiquement de conserver ou non la quasi-implication ou la quasi-règle $a \rightarrow b$. Ainsi, c'est à partir de la dialectique exemples-contre-exemples que la règle apparaît comme le dépassement de la contradiction.

Pour formaliser cette quasi-règle, nous considérons, comme le fait I.C. Lerman (Lerman, 1981) pour la similarité, deux parties quelconques X et Y de E, choisies aléatoirement et indépendamment (absence de lien a priori entre ces deux parties) et de mêmes cardinaux respectifs que A et B. Soit $\overline{Y}$ et $\overline{B}$ les complémentaires respectifs de Y et de B dans E de même cardinal $n_{\overline{b}} = n - nb$.

Nous dirons alors :

Définition 1 : $a \rightarrow b$ est admissible au niveau de confiance $1 - \alpha$ si et seulement si $Pr[card(X \cap \overline{Y}) \leq card(A \cap \overline{B})]$

Il est établi (Lerman et al., 1981) que, pour un certain processus de tirage, la variable aléatoire $card(X \cap \overline{Y})$ suit la loi de Poisson de paramètre $\frac{n_a \cdot n_{\overline{b}}}{n}$. Mais d'autres processus légitimes de tirage conduisent à une loi binomiale, voire une loi hypergéométrique. Dans le cas où $n_{\overline{b}} \neq 0$, nous réduisons et centrons cette variable de Poisson en la variable :

$Q(a, \overline{b}) = \frac{card(X \cap \overline{Y}) - \frac{n_a \cdot n_{\overline{b}}}{n}}{\sqrt{\frac{n_a \cdot n_{\overline{b}}}{n}}}$

Dans la réalisation expérimentale, la valeur observée de $Q(a, \overline{b})$ est $q(a, \overline{b})$. Elle estime un écart entre la contingence ($card(A \cap \overline{B})$) et la valeur qu'elle aurait prise s'il y avait eu indépendance entre a et b.

Définition 2 : $q(a, \overline{b}) = \frac{n_{a \wedge \overline{b}} - \frac{n_a \cdot n_{\overline{b}}}{n}}{\sqrt{\frac{n_a \cdot n_{\overline{b}}}{n}}}$ est appelé indice d'implication, nombre retenu comme indicateur de la non-implication de a sur b. Dans les cas légitimant convenablement l'approximation (par exemple, $\frac{n_a \cdot n_{\overline{b}}}{n} > 4$), la variable suit approximativement la loi normale centrée

réduite. L'intensité d'implication, qualité de l'admissibilité de $a \rightarrow b$, pour $na \leq nb$ et $nb \neq n$, est alors définie à partir de l'indice $q(a, \overline{b})$ par :

Définition 3 : Dans le cas où $n_b \neq n$, l'intensité d'implication qui mesure la qualité inductive de a sur b est : $\varphi(a, b) = 1 - Pr\left[Q(a, \overline{b}) - q(a, \overline{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\overline{b})}^{\infty} e^{-\frac{t^2}{2}} dt$ Ainsi dépend de 4 variables quasi-indépendantes, mais sous contraintes :

- n, la taille de l'échantillon
- n_a (resp. n_b)les exemples de a (resp. de b)
- $n_{a\wedge\overline{b}}$ les contre-exemples de l'implication
- $n_{a\wedge\overline{b}} \leq n_a \leq n_b \leq n$ les contraintes.

La formule définissant la qualité de la règle implicative, soit $\varphi(a, b)$, montre que cette qualité croît lorsque $q(a, \overline{b})$ décroît. Nous préciserons cette dépendance plus loin.

3 Indice en Analyse Statistique Implicative

3.1 Examen des variations des cardinaux

Etudier la stabilité de q, revient à examiner ses petites variations au voisinage des 4 valeurs entières observées (n, n_a, n_b et $n_{a\wedge\overline{b}}$). Pour ce faire, il est possible d'effectuer différentes simulations en croisant ces 4 variables entières dont q dépend (Fleury, 1996) ou (Gras et al., 2004). Mais, considérons ces variables comme nombres réels et q fonction continûment différentiable par rapport à ces variables contraintes à respecter les inégalités : n_a, n_b et $n_{a\wedge\overline{b}} \leq inf[n_a, n_b]$ et $sup[n_a, n_b] \leq n$. Il suffit alors d'examiner la différentielle de q par rapport à ces variables et d'en conserver la restriction aux valeurs entières des paramètres de la relation $a \rightarrow b$.

Différentielle de q

$$dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial n_a} dn_a + \frac{\partial q}{\partial n_b} dn_b + \frac{\partial q}{\partial n_{a\wedge\overline{b}}} dn_{a\wedge\overline{b}}$$

Si l'on veut étudier comment varie q en fonction de $n_{\overline{b}}$, il suffit de remplacer n_b par $n - n_b$ et donc changer le signe de la dérivée de n_b dans la dérivée partielle. En fait, l'intérêt de cette différentielle réside dans l'estimation de l'accroissement (positif ou négatif) de q soit Δq par rapport aux variations respectives Δn, Δn_a, $\Delta n_{\overline{b}}$ ou Δn_b et $\Delta n_{a\wedge\overline{b}}$.

Cas où seuls varient n_b et $n_{a\wedge\overline{b}}$ (dérivées partielles de n et n_a nulles)

$$\frac{\partial q}{\partial n_b} = \frac{1}{2} n_{a\wedge\overline{b}} \left(\frac{n_a}{n}\right)^{-1/2} (n - n_b)^{-3/2} + \frac{1}{2} \left(\frac{n_a}{n}\right)^{1/2} (n - n_b)^{-1/2} > 0$$

$$\frac{\partial q}{\partial n_{a\wedge\overline{b}}} = \frac{1}{\sqrt{\frac{n_{a\wedge\overline{b}}}{n}}} > 0$$

Ainsi, si les accroissements Δn_b et$\Delta n_{a\wedge\overline{b}}$ sont positifs, l'accroissement de $q(a,\overline{b})$ est également positif. Ceci s'interprète ainsi : si le nombre d'exemples de b et celui des contre-exemples de l'implication augmentent l'intensité d'implication diminue pour n et n_a constants. Autrement dit, cette intensité est maximum aux valeurs observées (n_b et $n_{a\wedge\overline{b}}$) et minimum aux valeurs $n_b + \Delta n_b$ et $n_{a\wedge\overline{b}} + \Delta n_{a\wedge\overline{b}}$.

Remarque 1 : On pourrait de la même façon étudier les variations de l'intensité lors de variations des autres variables. Par exemple, la dérivée partielle de q par rapport à n_a est :

$$\frac{\partial q}{\partial n_a} = \frac{n_{a\wedge\overline{b}}}{2\sqrt{\frac{n_{\overline{b}}}{n}}}.\frac{1}{n_a^{3/2}} - \sqrt{\frac{n_{\overline{b}}}{n.n_a}} < 0$$

Ainsi, sur $[0, n_b]$, $q(a,\overline{b})$ est toujours décroissante par rapport à n_a et est donc minimum pour $n_a = n_b$.

3.2 Examen des variations des fréquences

On examine maintenant les variations de q en fonction des fréquences relatives des variables précédentes où le référentiel a pour cardinal n.

Ainsi, on notera $f_i = \frac{n_i}{n}$ chacune des fréquences des variables respectives n, n_a, $n_{\overline{b}}$ (que nous privilégions par rapport à n_b pour des raisons de calculs) et $n_{a\wedge\overline{b}}$ s'écrit alors :

$$q(a,\overline{b}) = \sqrt{n}.\frac{f_{a\wedge\overline{b}} - f_a.f_{\overline{b}}}{\sqrt{f_a.f_{\overline{b}}}} = \sqrt{n}.\frac{f_{a\overline{b}}}{\sqrt{f_a.f_{\overline{b}}}} - \sqrt{n}.\sqrt{f_a.f_{\overline{b}}}$$

On étudie alors la stabilité à partir des dérivées partielles de q par rapport aux 4 variables fréquentielles :

$$dq = \frac{\partial q}{\partial n}dn + \frac{\partial q}{\partial f_{a\wedge\overline{b}}}df_{a\wedge\overline{b}} + \frac{\partial q}{\partial f_a}df_a + \frac{\partial q}{\partial f_{\overline{b}}}df_{\overline{b}} = \overrightarrow{grad}.\begin{bmatrix} dn \\ df_{a\wedge\overline{b}} \\ df_a \\ df_{\overline{b}} \end{bmatrix}$$

Or

$$\frac{\partial q}{\partial n} = \frac{1}{2}.\frac{f_{a\wedge\overline{b}} - f_a f_{\overline{b}}}{\sqrt{n f_a f_{\overline{b}}}};$$

$$\frac{\partial q}{\partial f_{a\wedge\overline{b}}} = \sqrt{n}.\frac{1}{\sqrt{f_a f_{\overline{b}}}};$$

$$\frac{\partial q}{\partial f_a} = -\frac{\sqrt{n}}{2}.\frac{f_{a\wedge\overline{b}}}{\sqrt{f_{\overline{b}}}}.\frac{1}{f_a^{3/2}} - \frac{\sqrt{n}}{2}.\sqrt{\frac{f_{\overline{b}}}{f_a}};$$

$$\frac{\partial q}{\partial f_{\overline{b}}} = -\frac{\sqrt{n}}{2}.\frac{f_{a\wedge\overline{b}}}{\sqrt{f_a}}.\frac{1}{f_{\overline{b}}^{3/2}} - \frac{\sqrt{n}}{2}.\sqrt{\frac{f_a}{f_{\overline{b}}}}$$

Remarque 2 : En calculant $\frac{\partial q}{\partial f_b}$ au lieu de $\frac{\partial q}{\partial f_{\overline{b}}}$, on constate que cette dérivée partielle est positive comme dans le 2. En effet : $\frac{\partial q}{\partial f_b} = \frac{\sqrt{n}}{2}.\frac{f_{a\wedge\overline{b}}}{\sqrt{f_a}}.\frac{1}{f_{\overline{b}}^{3/2}} + \frac{\sqrt{n}}{2}.\sqrt{\frac{f_a}{f_{\overline{b}}}}$

Remarque 3 : La "vitesse" d'accroissement de q (en valeur absolue) quand s'accroît le nombre de contre-exemples $a \wedge \overline{b}$, est donc, n et n_a restant constants, inversement proportionnelle à celle de la racine carrée de $n_{\overline{b}}$. Autrement dit : si $n_{\overline{b}}$ devient 4 fois plus petite (donc n_b 4 fois plus grand), cette vitesse est accélérée et multipliée par 2 (l'intensité d'implication devient bien meilleure).

3.3 Comparaison d'une transition d'un noeud d'un graphe implicatif à un suivant

Ce problème se présente dans l'examen de la recherche de l'alignement de deux ontologies (David et al., 2006b,a). Dans le cas qui nous intéresse où n et n_a sont constants, n_b décroît d'un noeud S_1 du graphe à un noeud S_2, donc $n_{a\wedge\overline{b}}$ va croître. On veut alors comparer les deux états contingents $S1$ et $S2$ représentés par ces noeuds successifs, il suffit de calculer les dérivées partielles des variables non constantes (ici $f_{a\wedge\overline{b}}$ et $f_{\overline{b}}$) au point S_1.

$$\Delta q(S_1, S_2) = \frac{\partial q}{\partial f_{a\wedge\overline{b}}}(S_1)\Delta f_{a\wedge\overline{b}} + \frac{\partial q}{\partial f_{\overline{b}}}(S_1)\Delta f_{\overline{b}}$$

les dérivées partielles étant calculées au point S_1, où $\Delta f_{a\wedge\overline{b}} = f_{a\wedge\overline{b}_2} - f_{a\wedge\overline{b}_1} < 0; \Delta f_{\overline{b}} = f_{\overline{b}_2} - f_{\overline{b}_1} > 0$ et $q(a, \overline{b}_2) = \sqrt{n}.\frac{\lfloor f_{a\overline{b}_2} - f_a f_{\overline{b}_2} \rfloor}{\sqrt{f_a f_{\overline{b}_2}}}, q(a, \overline{b}_1) = \sqrt{n}.\frac{\lfloor f_{a\overline{b}_1} - f_a f_{\overline{b}_1} \rfloor}{\sqrt{f_a f_{\overline{b}_1}}}$

Comme $q(S_1)$ et $q(S_2)$ sont négatifs dans le cas où n_a est inférieur à n_{b_i}, il suffit alors de comparer la variation observée $q(S_2) - q(S_1)$ -que l'on souhaite négative (meilleure intensité en S_2 qu'en S_1)- à la variation attendue par le gradient calculé. C'est le signe de la différence $\Delta q(S_1, S_2) - (q(S_2) - q(S_1))$ qui nous indiquera si l'intensité d'implication s'est améliorée ou non au cours de la transition $S_1 \rightarrow S_2$. *Si cette différence est positive, l'amélioration attendue est plus intéressante pour l'intensité qu'elle aurait été si l'évolution d'un noeud S_1 au suivant S_2 avait suivi le gradient de q en S_1.*

3.4 Examen des variations de l'intensité d'implication

On détermine, en conséquence, les variations de $\varphi(a, b)$ lors de la transition T du noeud S_1 vers le noeud S_2. Pour cela, on calcule l'intensité $\varphi(T)$, approximation gaussienne de la loi centrée réduite de la v.a. $Q(a, \overline{b})$ des contre-exemples, afférente à la variation attendue du gradient de q, à savoir $q(T) = q(a, \overline{b_1}) + \Delta q(S_1, S_2)$:

$$\varphi(T) = \frac{1}{\sqrt{2\pi}} \int_{q(T)}^{\infty} e^{\frac{-t^2}{2}} dt$$

Pour estimer le gain due à la transition, il suffit alors de comparer cette intensité à celle qui a été observée

$$\varphi(a, \overline{b_2}) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\overline{b_2})}^{\infty} e^{\frac{-t^2}{2}} dt$$

Le gain sera positif ou négatif suivant que l'intensité observée sera supérieure ou égale à l'intensité attendue du gradient. Il sera exprimé en pourcentage de $\varphi(T)$ par le rapport :

$$\frac{\varphi(a, b_2) - \varphi(T)}{\varphi(T)}$$

On dira, par exemple, que le gain lié à la transition vers S_2 est de 3% de l'intensité attendue de l'observation en S_1 et du gradient de q en ce noeud.

Cette méthode est **généralisable** quelles que soient les variables en jeu, c'est-à-dire dans les cas où d'autres variables peuvent être modifiées. Il suffit d'avoir recours à la différentielle de q selon ces 4 variables actives $(n, n_a, n_b, n_{a\wedge\overline{b}})$.

Remarque 4 : Considérons l'intensité d'implication φ comme fonction de $q(a, \overline{b})$:

$$\varphi(q) = \frac{1}{\sqrt{2\pi}} \int_q^{\infty} e^{\frac{-t^2}{2}} dt$$

On peut alors examiner comment $\varphi(q)$ varie lorsque q varie au voisinage d'une valeur donnée (a, b), sachant comment q varie lui-même en fonction des 4 paramètres qui le déterminent. Par dérivation de la borne d'intégration, on obtient :

$$\frac{d\varphi}{dq} = -\frac{1}{\sqrt{2\pi}} e^{\frac{-q^2}{2}} < 0$$

Ce qui confirme bien que l'intensité croît lorsque q décroît, mais la vitesse de croissance est précisée par la formule, ce qui permet d'étudier avec plus de précision les variations de $\varphi(q)$.

4 Examen d'autres indices

Contrairement à l'indice de base q et l'intensité d'implication qui mesure la qualité à travers une probabilité (cf. définition 3), les autres indices les plus courants se veulent eux-mêmes directement des mesures de qualité . Nous examinerons leurs sensibilités respectives aux variations des paramètres retenus dans la définition de ces indices. Nous conservons les notations adoptées au paragraphe 2 et choisissons des indices qui sont rappelés dans (Lenca et al., 2004).

4.1 L'indice Lift

Il s'exprime par : $l = \frac{n.n_{a\wedge b}}{n_a.n_b}$. Cette expression peut encore s'écrire pour mettre en évidence le nombre de contre-exemples : $l = \frac{n.(n_a - n_{a\wedge\overline{b}})}{n_a.n_b}$.

Pour étudier la sensibilité de l aux variations des paramètres, nous formons :

$$\frac{\partial l}{\partial n_{a\wedge\overline{b}}} = -\frac{n}{n_a.n_b}$$

Ainsi, la variation de l'indice Lift est indépendante de celle du nombre de contre-exemples. C'est une constante qui ne dépend que des variations des occurrences de a et de b. l décroît donc lorsque le nombre de contre-exemples croît, ce qui sémantiquement, est acceptable mais la vitesse de décroissance ne dépend pas de la vitesse de croissance de $n_{a\wedge\overline{b}}$.

4.2 L'indice MC (Multiplicateur de cote)

Il s'exprime ainsi : $m = \frac{n_a - n_{a\wedge\overline{b}}}{n_b.n_{a\wedge\overline{b}}}.n_{\overline{b}}$ (Lallich et Teytaud, 2004). Remarquons qu'en étant indépendant de n, il n'a pas de sens statistique aussi intéressant.

Sa dérivé partielle par rapport au nombre de contre-exemples est :

$$\frac{\partial m}{\partial n_{a\wedge\overline{b}}} = -\frac{n_a.n_{\overline{b}}}{n_b}.\left(\frac{1}{n_{a\wedge\overline{b}}}\right)^2$$

L'indice m décroît donc lorsque $n_{a\wedge\overline{b}}$ croît et la vitesse de décroissance est même plus rapide qu'avec le Lift et qu'avec l'indice q de base dans l'intensité d'implication. Il ne résiste pas à l'instabilité du nombre de contre-exemples.

4.3 La confiance

Cet indice c est le plus connu et, historiquement, après celui de J. Lovinger, le plus utilisé grâce à la caisse de résonance dont dispose une publication anglo-saxonne (Agrawal et al., 1993). Il est à l'origine de plusieurs autres indices communément employés qui n'en sont que des variantes satisfaisant telle ou telle exigence sémantique. De plus, il est simple et s'interprète aisément et immédiatement.

$$c = \frac{n_{a\wedge b}}{n_a} = 1 - \frac{n_{a\wedge\overline{b}}}{n_a}$$

La première forme s'interprète ainsi comme une fréquence conditionnelle des exemples de b quand a est connu. La sensibilité de cet indice aux variations des occurrences des contre-exemples se lit avec la dérivée partielle :

$$\frac{\partial c}{\partial n_{a\wedge\overline{b}}} = -\frac{1}{n_a}$$

Par conséquent, la confiance croît quand $n_{a\wedge\overline{b}}$ décroît ce qui est sémantiquement acceptable, mais la vitesse de variation est constante, indépendante de la vitesse de décroissance de ce nombre, des variations de n et de n_b. Le gradient de c ne s'exprime que par rapport à $n_{a\wedge\overline{b}}$ et à n_a. Ceci peut apparaître comme une restriction du rôle des paramètres dans l'expression de la sensibilité de l'indice.

5 Conclusion

Au moyen d'une approche utilisant plutôt l'analyse mathématique (calcul différentiel) que celles traditionnellement utilisées en data mining, nous avons étudié l'effet du bruit dans les données, de perturbations dans les cardinaux des paramètres, sur la mesure de qualité des règles d'association. Nous avons ainsi obtenu des moyens directs d'approximation de la mesure. La comparaison avec d'autres types de mesure a montré l'intérêt de la prise en compte en A.S.I. de paramètres négligés bien souvent dans d'autres types de mesure de qualité.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *the 1993 ACM SIGMOD international conference on Management of data*, pp. 207–216. ACM Press.

David, J., F. Guillet, R. Gras, et H. Briand (2006a). Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts. In *proc. of the 17th European Conference on Artificial Intelligence*, pp. 357–358.

David, J., F. Guillet, V. Philippé, H. Briand, et R. Gras (2006b). Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d ?implication. In *Extraction des Connaissances : Etat et Perspectives, RNTI-E-5*, pp. 409–413. Cepaduès Editions.

Fleury, L. (1996). Extraction de connaissances dans une base de données pour la gestion de ressources humaines. Thèse d'Université, Université de Nantes.

Gras, R. (1979). Contribution à l'étude experimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didacticques mathématiques. Thèse d'Etat, Université de Rennes.

Gras, R. (2005). Panorama du développement de l'a.s.i. à travers des situations fondatrices. In *3ème Rencontre Internationale A.S.I., Supplément nř 15 de la Revue "Quaderni di Ricerca in Didattica"*, pp. 9–33. Université de Palerme.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, Ratsimba-Rajohn, et A. Totohasina (1996). *L'implication statistique, une nouvelle méthode exploratoire de données*. Collection associée à Recherche en Didactique des Mathématiques, La Pensée Sauvage, Grenoble.

Gras, R., R. Couturier, M. Bernadet, J. Blanchard, H. Briand, F. Guillet, P. Kuntz, R. Lehn, et P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. *Revue des Nouvelles Technologies de l'Information 1*(RNTI-E).

Lallich, S. et O. Teytaud (2004). Évaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information 1*(RNTI-E), 193–217.

Lenca, P., P. Meyer, B. Vaillant, P. Picouet, et S. Lallich (2004). Evaluation et analyse multicritères des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information 1*(RNTI-E), 219–246.

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

Lerman, I.-C. (1981). *Classification et analyse ordinale des données*. Dunod.

Lerman, I.-C., R. Gras, et H. Rostam (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Mathématiques et Sciences Humaines* (74), 5–35.

Vaillant, B., S. Lallich, et P. Lenca (2006). Modeling of the counter-examples and association rules interestingness measures behavior. In S. Crone, S. Lessmann, et R. Stahlbock (Eds.), *The 2006 International Conference on Data Mining*, pp. 132–137.

Summary

The implicative statistical analysis allows the extraction of "type $a \rightarrow b$" rules from a set of data in the sense that b is usually achieved as soon as a is achieved. A measurement of this rule's quality is associated with this near implication. It is the implication intensity. It is important to study the stability of the rule when the parameters it depends on vary noticeably

when close to their observations. Several methods allow that. We study here the quality's measurement stability by differential analysis.

Construction d'une entropie décentrée pour l'apprentissage supervisé

Stéphane Lallich*, Philippe Lenca**, Benoît Vaillant***

*Université Lyon 2, Laboratoire ERIC
5 avenue Pierre Mendès-France
69676 Bron Cedex, France
stephane.lallich@univ-lyon2.fr
http://eric.univ-lyon2.fr/~lallich/
**GET - ENST Bretagne - Département LUSSI
CNRS UMR 2872 TAMCIC
Technopôle de Brest Iroise, CS 83818,
29238 Brest Cedex, France
philippe.lenca@enst-bretagne.fr
http://perso.enst-bretagne.fr/~lenca/
***UBS - IUT de Vannes - Département STID
Laboratoire VALORIA
8, rue Montaigne,
BP 561, 56017 Vannes, France
benoit.vaillant@univ-ubs.fr

Résumé. En apprentissage supervisé, de nombreuses mesures sont fondées sur la notion d'entropie. Une caractéristique majeure des entropies est qu'elles sont maximales lorsque la distribution des modalités de la variable de classe est uniforme, ce qui peut être un inconvénient lorsque cette distribution est très éloignée de l'uniformité. Pour traiter ce cas, nous proposons une entropie décentrée qui prend sa valeur maximale pour une distribution donnée. Cette distribution peut être la distribution *a priori* des classes ou une distribution tenant compte des coûts de mauvaise classification ou plus généralement une distribution fixée par l'utilisateur.

1 Motivations

En apprentissage supervisé à partir de variables catégorielles, par exemple en induction par arbres, de nombreux algorithmes d'apprentissage utilisent des mesures d'association prédictive fondées sur l'entropie de Shannon (1948). Considérons une variable de classe Y à q modalités et un prédicteur catégoriel X à p modalités. La fréquence relative conjointe du couple (x_i, y_j) est notée p_{ij}, $i = 1, 2 \ldots k$; $j = 1, 2, ...q$. En outre, on désigne par $h(Y)$ l'entropie de Shannon *a priori* de Y, $h(Y) = -\sum_{j=1}^{q} p_{.j} \log_2 p_{.j}$, et par $h(Y/X)$ l'espérance de l'entropie de Y conditionnellement à $X, h(Y/X) = E(h(Y/X = x_i))$. Parmi les mesures usuelles fondées sur l'entropie de Shannon, étudiées notamment par Wehenkel (1996), on citera en particulier :

- le gain d'entropie (Quinlan (1975)), qui vaut $h(Y) - h(Y/X)$;
- le coefficient u de Theil (1970), qui est le gain relatif d'entropie de Shannon, à savoir le gain normalisé par l'entropie *a priori* de Y, valant $\frac{h(Y)-h(Y/X)}{h(Y)}$;
- le gain-ratio de Quinlan (1993) qui rapporte le gain d'entropie dû à X à l'entropie de X plutôt qu'à l'entropie *a priori* de Y, afin de pénaliser les prédicteurs ayant le plus de modalités, ce qui correspond à $\frac{h(Y)-h(Y/X)}{h(X)}$;
- le coefficient de Kvalseth (1987), qui normalise le gain d'entropie par la moyenne des entropies de X et de Y, valant $\frac{2(h(Y)-h(Y/X))}{h(X)+h(Y)}$.

La particularité de ces coefficients est que l'entropie de Shannon d'une distribution est maximale lorsque la distribution est uniforme. Même si c'est le gain d'entropie par rapport à l'entropie *a priori* de Y qui figure au numérateur de chacun des coefficients précités, les entropies de Y et de $Y/X = x_i$ qui interviennent dans ce gain sont évaluées sur une échelle dont le "zéro" est la distribution uniforme des classes.

Il serait plus logique d'apprécier directement le gain d'entropie à l'aide d'une échelle dont le "zéro" serait la distribution *a priori* des classes. Cette caractéristique des coefficients fondés sur l'entropie est particulièrement contestable lorsque les classes à apprendre sont de fréquences très inégales ou lorsque les coûts de classification sont très inégaux.

Nous proposons dans ce papier une version décentrée de l'entropie qui permet d'apprécier directement à quel point le prédicteur candidat permet d'améliorer la distribution de la variable de classe. Après avoir présenté les travaux de référence par rapport au but poursuivi (section 2), nous exposons de façon détaillée le principe de décentrage de l'entropie de Shannon dans le cas d'une variable booléenne (section 3). Nous généralisons ensuite la méthode proposée au cas d'une variable ayant un nombre quelconque de modalités (section 4) et nous montrons comment étendre la démarche au cas d'une entropie généralisée (section 5) pour ensuite conclure (section 6).

2 Etat de l'art

Le principe de construction de cette entropie décentrée a été esquissé dans (Lallich et al., 2005) pour le cas où la variable de classe est booléenne. Dans ce précédent travail, nous proposons une version paramétrée de différentes mesures statistiques de l'intérêt des règles d'association du type $A \rightarrow B$, en particulier de l'intensité implication entropique (Gras et al., 2001). Pour construire une mesure statistique, Lerman et al. (1981) proposent de procéder comme suit : on commence par choisir une grandeur d'intérêt, par exemple le nombre de contre-exemples de la règle, ainsi qu'un modèle aléatoire et une hypothèse nulle H_0 qui spécifie la valeur de référence θ pour apprécier la confiance de la règle. On détermine ensuite la loi de la grandeur d'intérêt sous H_0. On construit alors une mesure statistique en centrant et réduisant la grandeur d'intérêt sous H_0 ou une mesure probabiliste en calculant le complément à 1 de la p-valeur associée à la mesure statistique. Dans la mesure où l'intensité d'implication entropique est la moyenne géométrique de l'intensité d'implication et d'un indice d'inclusion reposant sur les entropies des variables booélennes B/A et $\overline{A}/\overline{B}$, le paramétrage de l'intensité d'implication entropique exige le paramétrage de l'indice d'inclusion et par là même le décentrage de l'entropie qui doit prendre sa valeur maximale pour $p_{b/a} = \theta$ et non plus pour $p_{b/a} = 0.5$, tel que fait dans l'indice d'inclusion.

Dans une optique différente, directement reliée à la recherche d'une mesure d'association prédictive, tout particulièrement dans le cas des arbres de décision, Zighed, Ritschard et Marcellin ont proposé une entropie asymétrique et consistante. Cette mesure est asymétrique au sens où l'on peut choisir la distribution pour laquelle elle est maximale ; par consistante il faut entendre qu'elle prend en compte la taille de l'échantillon.

Dans un premier papier (Marcellin et al. (2006)), ces auteurs traitent d'abord le cas d'une variable de classe booléenne, de fréquences p pour $Y = 1$ et $1 - p$ pour $Y = 0$. Ils rappellent d'abord les propriétés classiques de l'entropie de Shannon de Y, notée $h(p) = -p \log_2(p) - (1 - p) \log_2(1 - p)$. Celle-ci est une fonction réelle non négative de p, qui, entre autres propriétés, vérifie notamment :

1. **Invariance par permutation des modalités**
   $h(p)$ ne change pas lorsque l'on permute les modalités de Y.
2. **Maximalité**
   La valeur de $h(p)$ est maximale lorsque la distribution de Y est uniforme, c'est-à-dire de fréquences égales à $1/2$ pour chacune des deux modalités de Y.
3. **Minimalité**
   La valeur de $h(p)$ est minimale lorsque la distribution est certaine, concentrée sur l'une des modalités de Y, toutes les autres modalités étant de fréquence nulle.
4. **Concavité stricte**
   L'entropie $h(p)$ est une fonction strictement concave.

Marcellin et al. (2006) conservent la propriété 4 (*concavité stricte*) mais modifient la propriété 2 de telle sorte que l'entropie soit maximale pour une distribution laissée au choix de l'utilisateur (maximale pour $p = \theta$, où θ est choisi par l'utilisateur), ce qui impose de renoncer à la propriété 1 (*invariance par permutation des modalités*). Ils proposent :

$$h_\theta(p) = \frac{p(1-p)}{(1-2\theta)p + \theta^2}$$

On observera que pour $\theta = 0.5$, cette entropie asymétrique se confond avec l'entropie quadratique de Gini. Dans un second papier, les mêmes auteurs étendent leur approche au cas d'une variable de classe à q modalités (Zighed et al. (2007)). En outre, dans la mesure où l'on ne peut qu'estimer la distribution réelle $(p_j)_{j=1,2\ldots,q}$ par la distribution empirique $(f_j)_{j=1,2\ldots,q}$, ils souhaitent que pour une même distribution de fréquences empiriques, la valeur de l'entropie soit d'autant plus faible que n est grand (propriété 5, nouvelle propriété dite de *consistance*). Ils sont ainsi amenés à modifier la propriété 3 (*Minimalité*) en une propriété $3'$ (*Minimalité asymptotique*) : l'entropie d'une variable certaine est simplement astreinte à tendre vers 0 lorsque $n \to \infty$. Pour satisfaire ces nouvelles propriétés $3'$ et 5, ils suggèrent d'estimer les fréquences théoriques p_j par leur estimateur de Laplace, $\widehat{p_j} = \frac{nf_j+1}{n+q}$. Ils proposent ainsi une entropie asymétrique consistante définie par :

$$h_\theta(p) = \sum_{j=1}^{q} \frac{\widehat{p_j}(1-\widehat{p_j})}{(1-2\theta_j)\widehat{p_j} + \theta_j^2}$$

Une des particularités du principe de décentrage que nous proposons dans ce papier, par rapport à celui proposé par Zighed et al. (2007) est qu'au lieu de donner une seule entropie décentrée, il s'adapte à n'importe quel type d'entropie, que ce soit une entropie de Shannon ou plus généralement une entropie d'ordre bêta de Daroczy (Daroczy (1970)).

3 Entropie décentrée pour les variables booléennes

3.1 Principe de construction

On considère une variable de classe Y qui comporte $q = 2$ modalités. La distribution de fréquences de Y pour les valeurs 0 et 1 est notée $(1-p, p)$. Nous voulons définir une entropie décentrée associée à $(1-p, p)$, notée $\eta_\theta(p)$, qui soit maximale lorsque $p = \theta$, où θ est fixé par l'utilisateur, et non pas lorsque $p = 0.5$ (cas d'une distribution uniforme). Pour définir cette entropie décentrée, suivant la démarche décrite dans Lallich et al. (2005), nous proposons de transformer la distribution $(1-p, p)$ en $(1-\pi, \pi)$, où :

$$\pi = \frac{p}{2\theta} \text{ si } 0 \leq p \leq \theta,\ \pi = \frac{p+1-2\theta}{2(1-\theta)} \text{ si } \theta \leq p \leq 1$$

En toute rigueur, les fréquences transformées devraient être notées $1-\pi_\theta$ et π_θ. Elles sont notées $1-\pi$ et π dans un souci de simplification. Ce sont bien des fréquences, soit $0 \leq \pi \leq 1$. L'entropie décentrée $\eta_\theta(p)$ est alors définie comme l'entropie de $(1-\pi, \pi)$:

$$\eta_\theta(p) = -\pi \log_2 \pi - (1-\pi) \log_2 (1-\pi)$$

Par rapport à la distribution $(1-p, p)$, il est clair que $\eta_\theta(p)$ n'est pas une entropie au sens strict du terme. Ses propriétés doivent être analysées en tenant compte du fait que $\eta_\theta(p)$ est l'entropie de la distribution transformée $(1-\pi, \pi)$, soit $\eta_\theta(p) = h(\pi)$. Le comportement de cette entropie est illustré par la figure 1 pour $\theta = 0.2$.

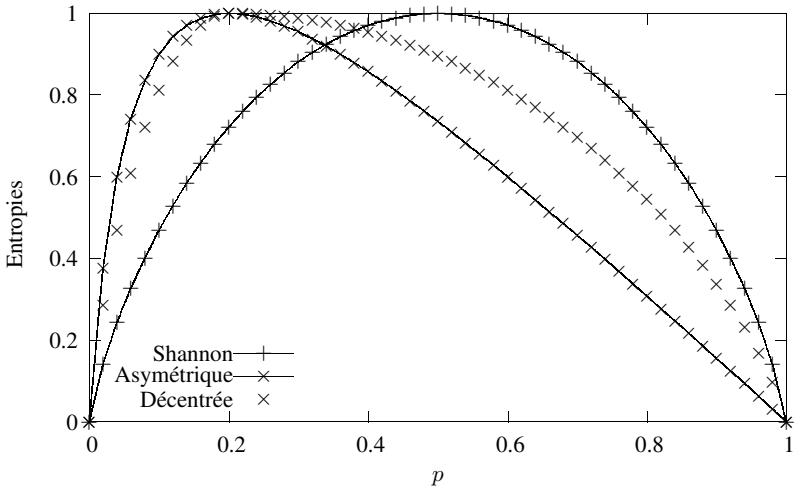

FIG. 1 – *Entropies décentrée, asymétrique et de Shannon*

3.2 Propriétés

L'entropie décentrée conserve différentes propriétés de l'entropie, parmi celles étudiées notamment par Zighed et Rakotomalala (1998) dans un contexte de Data mining. Ces propriétés sont faciles à démontrer dans la mesure où $\eta_\theta(p)$ est définie comme une entropie en fonction de π et en possède les caractéristiques.

Au préalable, pour démontrer certaines des propriétés de $\eta_\theta(p)$ en fonction de p, il faut calculer ses dérivées première et seconde par rapport à p sachant que $\eta_\theta(p) = h(\pi)$ est une fonction concave de π (entropie) où π est une fonction linéaire croissante par morceaux de p. Pour faciliter ces calculs, on considère la fonction $\eta(x) = h(f(x))$, où h est concave et f est linéaire croissante, $f(x) = ax + b$, $a > 0$, soit $f'(x) = a$ et $f''(x) = 0$. Alors les dérivées première et seconde de $\eta(x)$, par rapport à x, s'écrivent respectivement $\eta'(x) = h'(f(x))f'(x) = ah'(f(x))$ et $\eta''(x) = a^2h''(f(x))$.

1. **Invariance par permutation des catégories**
   Cette propriété des entropies est volontairement abandonnée, puisque l'on décentre l'entropie.
2. **Maximalité**
   $\eta_\theta(p)$ est maximale, de valeur 1, pour $\pi = 0.5$, soit $p = 0.5 \times 2\theta = \theta$. Sa dérivée première par rapport à θ s'écrit :
   – $\eta'_\theta(p) = \frac{1}{2\theta}h'(\pi) = \frac{1}{2\theta}\left(\log_2(1-\pi) - \log_2 \pi\right)$, pour $0 \leq p \leq \theta$,
   – $\eta'_\theta(p) = \frac{1}{2(1-\theta)}h'(\pi) = \frac{1}{2(1-\theta)}\left(\log_2(1-\pi) - \log_2 \pi\right)$, pour $\theta \leq p \leq 1$
   La dérivée est nulle pour $\pi = 0.5$, soit $p = \theta$. L'entropie décentrée $\eta_\theta(Y)$ est donc une fonction dérivable en tout point qui prend la valeur maximale 1 pour $p = \theta$
3. **Minimalité**
   $\eta_\theta(p)$ est minimale pour $\pi = 0$ et $\pi = 1$, donc pour $p = 0$ et $p = 1$.
4. **Concavité**
   D'après le calcul préalable :
   – $\eta''_\theta(p) = \frac{1}{4\theta^2}h''(\pi) = \frac{-1}{4\theta^2 Ln2}\frac{1}{\pi(1-\pi)}$, pour $0 \leq p \leq \theta$,
   – $\eta''_\theta(p) = \frac{1}{4(1-\theta)^2}h''(\pi) = \frac{-1}{4(1-\theta)^2 Ln2}\frac{1}{\pi(1-\pi)}$, pour $\theta \leq p \leq 1$
   Par suite, $\eta_\theta(p)$ est une fonction concave de p. On remarquera qu'au point $p = \theta$, la dérivée seconde à gauche est distincte de la dérivée seconde à droite.

3.3 Autres propriétés

1. Prise en compte de la taille de l'échantillon (consistance) et minimalité asymptotique.

Pour satisfaire cette propriété introduite par Zighed et Rakotomalala (1998) pour les arbres de décision, deux possibilités s'offrent à nous, que nous n'avons pas encore exploitées. En premier lieu, on peut suivre la démarche utilisée par Zighed et al. (2007) pour construire une entropie asymétrique consistante, qui consiste à estimer les fréquences théoriques à l'aide de l'estimateur de Laplace. Par ailleurs, on peut avoir recours à la méthode delta (Goodman et Kruskal (1972)) pour estimer la variance asymptotique de l'entropie et raisonner sur la borne basse de l'intervalle de confiance.

2. Condition de transfert (Pigou-Dalton).

p	$h(p)$	$\eta_\theta(p)$
0	0	0
0.1	0.469	0.811
0.2	0.722	1
0.3	0,881	0.989
0.4	0.971	0.954
0.5	1	0.896
0.6	0.971	0.811

TAB. 1 – *Quelques valeurs remarquables de $h(p)$ et $\eta_\theta(p)$.*

On sait que l'entropie de Shannon vérifie la condition de transfert (dite de Pigou-Dalton) au sens où elle augmente lorsqu'une modalité plus fréquente transfère une partie de sa masse de fréquence sur l'autre modalité sans que l'ordre soit modifié. Pour l'entropie décentrée, comme l'illustre la figure 1, cette condition de transfert reste vraie pourvu que la position de chacune des deux modalités par rapport à θ ne soit pas modifiée. Le tableau 1 illustre ce phénomène (dans le cas $\theta = 0.2$). Si l'on transfère une fréquence de 0.1, sur la modalité de plus faible fréquence, l'entropie $h(p)$ augmente tant que l'ordre des 2 modalités reste inchangé (chaque modalité doit rester du même côté de 0.5 à l'issue du transfert). Dans le cas de $\eta_\theta(p)$, il faut que le transfert conserve la position de chaque modalité par rapport à θ.

3. Propriétés perdues

Certaines propriétés des entropies semblent perdues, ou sans objet, ainsi l'insensibilité aux modalités de fréquence nulle, l'uniformité partagée, le comportement en cas de scission-fusion de modalités et la pseudo-additivité. Il faut donc souligner que les entropies décentrées ne sont pas des entropies au sens stricte du terme.

4 Entropie décentrée pour une variable à q modalités

Pour étendre la définition de l'entropie décentrée au cas d'une variable Y ayant q modalités, $q > 2$, on utilise une stratégie similaire à celle utilisée dans le cas booléen. On note $\underline{p} = (p_1, p_2, ...p_q)$ le vecteur des fréquences de Y et $\underline{\theta} = (\theta_1, \theta_2, ...\theta_q)$, le vecteur des fréquences de la distribution de référence, par exemple la distribution *a priori* de Y en apprentissage supervisé.

4.1 Recherche de la forme décentrée

L'entropie de $\underline{p}$ s'écrit $h(\underline{p}) = -\sum_{j=1}^{q} p_j \log_2 p_j$, alors que l'on recherche l'entropie décentrée sous la forme $\eta(\underline{p}) = -\sum_{j=1}^{q} \pi_j \log_2 \pi_j$, où pour être analogues à des fréquences relatives, les π_j doivent vérifier :

– $0 \leq \pi_j \leq 1$
– $\sum_{j=1}^{q} \pi_j = 1$

C'est ainsi que :

– π_j passe de 0 à $1/q$, lorsque p_j passe de 0 à θ_j
– π_j passe de $1/q$ à 1, lorsque p_j passe de θ_j à 1

Cherchons les π_j sous la forme $\pi_j = \frac{p_j - b_j}{a_j}$.
Dans le cas $0 \leq p_j \leq \theta_j$ on a :
– $p_j = 0$, $\pi_j = 0$, soit $0 = -\frac{b_j}{a_j}$ et $b_j = 0$
– $p_j = \theta_j$, $\pi_j = 1/q$, soit $1/q = \frac{\theta_j - b_j}{a_j}$ et $a_j = q\theta_j$
Il vient :

$$\pi_j = \frac{p_j}{q\theta_j}, \text{ si } 0 \leq p_j \leq \theta_j$$

Dans le cas $\theta_j \leq p_j \leq 1$:
– $p_j = \theta_j$, $\pi_j = 1/q$, soit $1/q = \frac{\theta_j - b_j}{a_j}$ et $a_j = q(\theta_j - b_j)$
– $p_j = 1$, $\pi_j = 1$, soit $1 = \frac{1 - b_j}{a_j}$ et $a_j = 1 - b_j$
Par suite, $1 - b_j = q(\theta_j - b_j)$, d'où $b_j = \frac{q\theta_j - 1}{q-1}$ et $a_j = \frac{q(1-\theta_j)}{q-1}$
On obtient alors :

$$\pi_j = \frac{p_j - b_j}{a_j} = \frac{q(p_j - \theta_j) + 1 - p_j}{q(1 - \theta_j)}$$

A titre de vérification :
– si $q = 2$, on retrouve bien les formules qui précèdent (section 3),
– par construction, π_j prend les valeurs $0, 1/q$ et 1, lorsque p_j prend les valeurs $0, \theta_j$ et 1,
– on a bien $0 \leq \pi_j \leq 1$, $j = 1, 2, \ldots, q$,
– seul problème, la condition de normalisation, qui n'est automatiquement vérifiée que pour $q = 2$.

Pour régler ce problème, il suffit de normaliser les π_j. On obtient ainsi les π_j^* définis par $\pi_j^* = \frac{\pi_j}{\sum_{j=1}^{q} \pi_j}$. Les propriétés précitées sont conservées, puisque le facteur de normalisation vaut 1 pour $p_j = \theta_j$, car alors les π_j sont égaux à $1/q$. L'entropie décentrée pour une variable à q modalités est alors définie par $\eta_{\underline{\theta}}(\underline{p}) = h(\underline{\pi}^*)$.

4.2 Représentation graphique

Pour illustrer le comportement de l'entropie décentrée, il est possible de représenter les valeurs de $\eta_{\underline{\theta}}(\underline{p})$ dans le cas de $q = 3$ catégories. Les fréquences étant liées par la condition de normalisation, on peut représenter l'entropie décentrée par une surface située dans un hyperplan de R^4.

5 Décentrage des entropies généralisées

L'entropie de Shannon n'est pas la seule fonction de diversité ou d'incertitude utilisable pour construire des coefficients d'association prédictive. Déjà, Goodman et Kruskal (1954) avaient proposé une présentation unifiée des trois coefficients usuels que sont le λ de Guttman, le u de Theil et le τ de Goodman et Kruskal, sous l'appellation de coefficient *PRE* (*Proportional Reduction in Error*). De façon plus générale (*cf.* Lallich (2002) où l'on retrouvera le détail des coefficients cités ici), nous avons construit les coefficients du type *Proportional Reduction in Diversity* (*PRD*), qui sont l'analogue du gain normalisé lorsque l'entropie de Shannon est

remplacée par une fonction d'incertitude quelconque. Pour qu'une telle construction soit justifiée, comme le note C. d'Aubigny (d'Aubigny, 1980), il suffit que la fonction d'incertitude soit concave, afin que la réduction moyenne de diversité de Y due au conditionnement suivant X soit positive, grâce à l'inégalité de Jensen. Si la fonction I choisie est l'entropie quadratique de Gini, $I(Y) = 2(1 - \sum_{j=1}^{q} p_j^2)$ (indice de diversité de Gini-Simpson) le gain relatif correspond au coefficient τ de Goodman et Kruskal, alors que si la fonction choisie est $I(Y) = q - 1$ (indice de diversité du nombre d'espèces, en écologie) le gain relatif correspond au coefficient λ de Guttman, Goodman et Kruskal.

Plus généralement encore, nous avons remarqué que les fonctions d'incertitude utilisables étaient soit des entropies généralisées d'ordre β de Daroczy (1970), soit des diversités de rangs d'ordre ρ introduites par Patil et Taillie (1982). Nous avons ainsi proposé (Lallich (2002)) une écriture unique pour la quasi-totalité des coefficients usuels sous la forme d'une réduction normalisée d'entropie généralisée ou de diversité de rangs :

$$\lambda_\alpha(Y/X) = \frac{I(Y) - I(Y/X)}{\alpha I(Y) + (1-\alpha) I(X)}$$

Dans cette formule, I renvoie aussi bien aux entropies d'ordre β qu'à leur équivalent en termes de diversité de rangs d'ordre ρ, alors que α est à la disposition de l'utilisateur pour arbitrer entre les deux normalisations usuelles. Cette expression permet de retrouver les coefficients usuels ($\alpha = 1$) fondés sur une entropie généralisée ($\beta = 0$: nombre de catégories ; $\beta = 1$: Theil ; $\beta = 2$: Gini) ou de rangs ($\rho = 0$: Guttman ; $\rho = 1$: Utton) ainsi que des analogues du gain-ratio ($\alpha = 0$) et du coefficient de Kvalseth ($\alpha = 0.5$) et d'en générer de nouveaux. La stratégie de décentrage que nous avons proposée s'applique sans difficultés au cas où la fonction d'incertitude choisie est une entropie généralisée ou une entropie de rangs.

Par exemple, la formule générale des entropies généralisées d'ordre β s'écrit $H_\beta(\underline{p}) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1 - \sum_{j=1}^{q} p_j^\beta\right)$. Pour décentrer cette entropie, il faut d'abord transformer les fréquences p_j en π_j, puis normaliser les π_j, pour obtenir les pseudos-fréquences π_j^*, suivant la procédure décrite dans la section qui précède. On obtient l'évaluation de la distribution $\underline{p}$ par l'entropie décentrée d'ordre β en formant :

$$\eta_\beta(\underline{p}) = H_\beta(\underline{\pi^*}) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1 - \sum\nolimits_{j=1}^{q} \pi_j^{*\beta}\right)$$

La version décentrée des entropies de rangs est construite suivant le même procédé. Par exemple, dans le cas $\rho = 0$, qui correspond à la logique du coefficient de Guttman, on a $H_{\rho=0}(\underline{p}) = 2\left(1 - \max\left\{p_j, j = 1, 2, ..., q\right\}\right)$, d'où :

$$\eta_{\rho=0}(\underline{p}) = H_{\rho=0}(\underline{\pi^*}) = 2\left(1 - \max\left\{\pi_j^*, j = 1, 2, ..., q\right\}\right)$$

Pour illustrer le comportement des entropies généralisées décentrées, nous avons représenté η_β, pour $\beta = 0, 0.5, 1, 2, 5$, ainsi que $\eta_{\rho=0}$ (figure 2) et l'entropie asymétrique de Zighed et al. (Zighed et al. (2007)) dans le cas où la distribution *a priori* de la variable de classe est $(0.8, 0.2)$, ce qui correspond à $\theta = 0.2$. Les différences de comportement apparaissent clairement sur cette figure qui montre bien l'intérêt et la spécificité de l'entropie asymétrique que nous proposons. C'est en fait un "kit de décentrage" que l'on peut appliquer à n'importe quelle mesure d'association prédictive reposant sur un gain d'incertitude. Le choix de la valeur de β ou ρ dépend de la réactivité que l'on attend de la mesure.

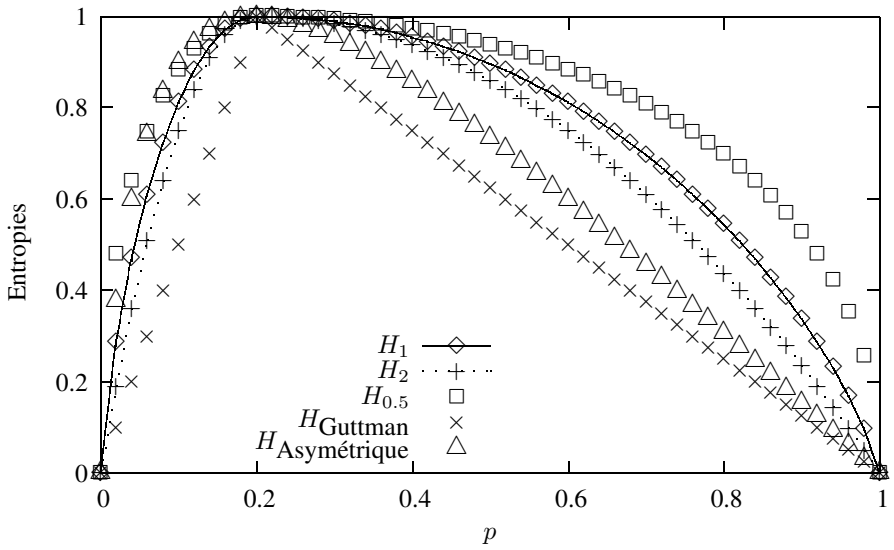

FIG. 2 – *Décentrage des entropies généralisées*

6 Conclusion et travaux futurs

Les mesures d'association prédictive usuelles peuvent être exprimées sous forme d'un gain normalisé associé à une fonction d'incertitude, entropie généralisée ou diversité de rangs. Au terme de cette première étape, nous proposons une méthode de décentrage qui permet d'associer une entropie décentrée à n'importe quelle entropie généralisée d'ordre β, ou entropie de rangs d'ordre ρ.

La phase suivante de ce travail est bien sûr la mise en œuvre des entropies décentrées sur des données réelles, notamment lorsque celles-ci présentent une variable de classe très déséquilibrée, pour examiner dans quelle mesure elles permettent d'améliorer les performances des algorithmes de classification supervisée.

Références

Daroczy, A. (1970). Generalized information functions. *Information and Control* (16), 36–51.

d'Aubigny, C. (1980). *Etude de la morphologie des indices d'association*. Ph. D. thesis, Thèse de 3e cycle, Université Joseph Fourier, Grenoble, France.

Goodman, L. A. et W. H. Kruskal (1954). Measures of association for cross classifications, i. *JASA I*(49), 732–764.

Goodman, L. A. et W. H. Kruskal (1972). Measures of association for cross classifications, iv. *JASA IV*(67), 415–421.

Gras, R., P. Kuntz, R. Couturier, et F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Conférence Extraction des connaissances et apprentissage (EGC 2001) 1*(1-2), 69–80.

Kvalseth, T. O. (1987). Entropy and correlation : some comments. *IEEE Trans. on Systems, Man and Cybernetics 17*(3), 517–519.

Lallich, S. (2002). *Mesure et validation en extraction des connaissances à partir des données*. Ph. D. thesis, Habilitation à Diriger des Recherches, Université Lyon 2, France.

Lallich, S., B. Vaillant, et P. Lenca (2005). Parametrised measures for the evaluation of association rule interestingness. In *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA 2005), Brest, France*, pp. 220–229.

Lerman, I., R. Gras, et H. Rostam (1981). Elaboration et évaluation d'un indice d'implication pour données binaires. *Mathématiques et Sciences Humaines 74*, 5–35.

Marcellin, S., D. Zighed, et G. Ritschard (2006). An asymmetric entropy measure for decision trees. In *Proceedings in Computational Statistics, Berlin : Springer, CD*, pp. 975–982.

Patil, G. et C. Taillie (1982). Diversity as a concept and its measurement. *Journal of American Statistical Association 77*(379), 548–567.

Quinlan, J. (1975). *Machine Learning*, Volume 1.

Quinlan, J. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technological Journal* (27), 379Ű423, 623Ű656.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology* (76), 103–154.

Wehenkel, L. (1996). On uncertainty measures used for decision tree induction. In *Proceedings of Info. Proc. and Manag. Of Uncertainty*, pp. 413–418.

Zighed, D., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. *Actes 7e Conférence EGC, Extraction et Gestion des Connaissances, Namur to appear*.

Zighed, D. A. et R. Rakotomalala (1998). *Graphes d'induction et apprentissage machine*. Hermès Paris.

Summary

In supervised learning, many measures are based on the concept of entropy. A major characteristic of the entropies is that they take their maximal value when the distribution of the modalities of the class variable is uniform. To deal with the case where the *a priori* frequencies of the class variable modalities are very unbalanced, we propose a decentered entropy which takes its maximum value for a distribution fixed by the user. This distribution can be the *a priori* distribution of the class variable modalities or a distribution taking into account the costs of misclassification.

Qualité des règles d'association : étude de données d'entreprise

Benoît Vaillant*, Stéphanie Menou**, Sorin Moga***,
Philippe Lenca***, Stéphane Lallich****

*Département STID / VALORIA
Université de Bretagne Sud
8, rue Montaigne, B.P. 561
56017 Vannes Cedex

**DIXID
4, rue Ampère
22300 Lannion

***UMR 2872 TAMCIC
GET / ENST Bretagne
Technopôle de Brest Iroise
CS 83818, 29238 Brest Cedex

****Laboratoire ERIC
Université Lumière - Lyon 2
5 avenue Pierre Mendès-France
69676 Bron Cedex

Résumé. L'extraction de connaissances à partir de données a pour objet la découverte de connaissances à partir de grandes quantités de données, par des méthodes d'apprentissage automatiques ou semi-automatiques, et l'utilisation industrielle ou opérationnelle de ces connaissances.

Nous explorons ici une application concrète de la fouille de données, sur un système de serveur vocal. Nous nous intéressons en particulier à l'étape cruciale de validation de motifs extraits via l'utilisation de mesures de qualité.

1 Introduction

Parmi les nombreuses définitions de l'extraction de connaissances à partir de données (ECD par la suite), on retrouve fréquemment celle proposée par Fayyad et al. (1996) : le *processus complexe permettant l'identification, au sein des données, de motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possible*. Cette définition insiste sur la notion de qualité et d'intérêt, vis-à-vis d'une utilisation finale. La notion d'intérêt varie ainsi en fonction du contexte applicatif et de l'utilisateur expert des données.

La validation de ces connaissances est une étape cruciale du processus d'ECD (Hilderman et Hamilton, 2001). Elle devient particulièrement incontournable en extraction de règles d'association face au volume de règles automatiquement produit. Partant du constat que la sélection des *bonnes* connaissances passe aussi par l'utilisation des *bonnes* mesures (Tan et al., 2002; Lenca et al., 2003; Carvalho et al., 2005), nous présentons dans Vaillant (2006) une étude systématique des mesures d'intérêt des règles d'association selon différents axes d'analyse. Parmi ces axes, nous proposons une aide multicritères à la décision afin de sélectionner les bonnes mesures (Lenca et al., 2007). Nous présentons ici une application de cette approche pour des données issues de l'interaction d'utilisateurs avec un service d'assistance téléphonique.

Nous exposons dans un premier temps, en section deux, le processus d'ECD que nous avons suivi. Dans la troisième section on présente les données étudiées ainsi que les premiers résultats bruts. La quatrième section présente brièvement les critères de décision retenus ainsi que des éléments liés à leur importance au regard de préférences exprimées par l'utilisateur. Dans la section cinq nous analysons les résultats préliminaires de l'approche aide multicritère à la décision en vue d'orienter le choix de mesure(s) de qualité(s). Enfin, nous concluons et proposons des perspectives à nos travaux.

2 Présentation du processus de fouille de données

Un processus d'ECD se décompose en plusieurs étapes. Classiquement on retrouve les étapes suivantes : ciblage des données à fouiller, nettoyage, transformation et recodage, fouille, évaluation des résultats, intégration des résultats.

Chaque étape, spécifique au problème traité, est validée par un utilisateur (tantôt un expert métier – $E_{\mathcal{R}}$, tantôt un expert de l'ECD –$E_{\mathcal{A}}$). Nous distinguons clairement plusieurs intervenants dans le processus. L'étude de cas que nous présentons permet de mettre aussi en évidence les différents points de vue, les différentes expertises et par voie de conséquence les difficultés inhérentes à ce type de collaboration.

Dans cette étude, nous avons suivi toutes les étapes, exceptée bien évidemment celle d'intégration des résultats puisque nous n'en sommes qu'aux résultats préliminaires. Nous développons particulièrement les étapes de fouille et d'évaluation des résultats.

L'étape de fouille est au cœur du processus. C'est lors de son déroulement que l'on génère les motifs à interpréter. A l'issue de cette étape, les résultats produits par les algorithmes de fouille de données ne sont pas toujours exploitables. Il est alors nécessaire de les soumettre à une évaluation. La quantité d'information pouvant être générée étant prohibitive pour une évaluation manuelle, il est courant d'automatiser un filtrage ou un ordonnancement de ces motifs, afin d'assister l'utilisateur expert des données dans la tâche de prise de décision ou d'interprétation des résultats. Ceci peut se faire au moyen de mesures de qualité, fonctions mathématiques modélisant une catégorie de connaissances désirée, et mettant celle-ci en avant parmi l'ensemble des motifs extraits.

Nous utiliserons en pratique l'implémentation proposée par Borgelt et Kruse (2002) de l'algorithme APRIORI d'Agrawal et Srikant (1994) afin d'extraire des règles d'association. Nous présentons succinctement le principe de cette fouille en section 2.1.

2.1 L'extraction de règles d'association

En extraction de règles d'association, on suppose que les données à explorer sont binaires, *i.e.* qu'on peut décrire chaque objet (ou transaction) au moyen d'un ensemble fini d'attributs booléens $\mathcal{I} = \{i_1, \ldots, i_m\}$, également appelés *items*. Pour un ensemble X d'items (on parle généralement d'itemsets) de $\mathcal{I}$, on dira qu'une transaction t contient X si et seulement si $X \subseteq t$ et on appelle support d'un itemset X le rapport entre le nombre de transactions contenant X et le nombre total de transactions.

Une règle d'association est un couple (A,B), où A et B sont des itemsets non vides disjoints, *i.e.* $\mathtt{A} \neq \emptyset$, $\mathtt{B} \neq \emptyset$, et $\mathtt{A} \cap \mathtt{B} = \emptyset$. On note classiquement un tel couple sous la forme $\mathtt{A} \rightarrow \mathtt{B}$. L'itemset A est appelé *prémisse* et B *conclusion*. On définit le support d'une règle d'association

comme étant le support de l'itemset A∪B (*i.e.* la proportion de transactions contenant à la fois A et B). On définit de plus la confiance d'une règle $\mathtt{A} \rightarrow \mathtt{B}$, notée $\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B})$, comme étant le rapport entre le support de l'itemset A∪B et celui de A : $\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) = \textsc{Sup}(\mathtt{A} \cup \mathtt{B}) / \textsc{Sup}(\mathtt{A})$.

Etant donnés des seuils minimaux de support et de confiance σ_s et σ_c fixés au préalable, l'objectif des algorithmes de la famille APRIORI est d'extraire **toutes** les règles $\mathtt{A} \rightarrow \mathtt{B}$ vérifiant $\textsc{Sup}(\mathtt{A} \rightarrow \mathtt{B}) \geq \sigma_s$ et $\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) \geq \sigma_c$. Nous nous limiterons par la suite aux règles ne faisant intervenir qu'un seul item en conclusion, sans que cela ne nuise à notre application pratique (voir section 3). Il découle de cette exploration exhaustive une explosion combinatoire, le nombre de règles extrait devenant trop important pour permettre une évaluation individuelle de la pertinence de chaque règle par l'expert des données.

Plusieurs voies permettent de répondre à ce problème. Nous nous sommes intéressés à l'une d'elle : l'évaluation de la qualité des règles par des mesures objectives.

2.2 Mesures de qualité de règles d'association

Etant donné une règle $\mathtt{A} \rightarrow \mathtt{B}$, il est courant de l'analyser en se rapportant à une matrice de contingence croisant deux variables binaires A et B. On obtient une description de la règle $\mathtt{A} \rightarrow \mathtt{B}$ sous l'une des formes listées dans le tableau 1, où n_x correspond au nombre de transactions contenant X (ou fréquence absolue) et p_x à la fréquence observée de X (ou fréquence relative).

Par convention, on note multiplicativement l'union de deux itemsets afin d'alléger l'écriture (ainsi, le nombre de transactions contenant $X \cup Y$ est n_{xy}). Ces notations sont équivalentes, on passe de l'une à l'autre simplement via la relation $p_x = n_x / n$.

$A \backslash B$	0	1	total
0	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}b}$	$n_{\bar{a}}$
1	$n_{a\bar{b}}$	n_{ab}	n_a
total	$n_{\bar{b}}$	n_b	n

$A \backslash B$	0	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
total	$p_{\bar{b}}$	p_b	1

TAB. 1 – *Notations usuelles associées à une règle* $\mathtt{A} \rightarrow \mathtt{B}$

Afin de quantifier l'intérêt d'une règle, il est fréquent d'avoir recours à des mesures objectives. De telles mesures sont des fonctions définies sur la table de contingence présentée dans le tableau 1. Elles sont dites objectives (Freitas, 1999; Hilderman et Hamilton, 2000) par opposition aux mesures subjectives (Silberschatz et Tuzhilin, 1995; Liu et al., 2000; Brisson, 2004) qui, en plus d'être basées sur des comptages fréquentiels sur les données, prennent aussi en compte des connaissances spécifiques au domaine fouillé.

On peut recenser un très grand nombre de mesures de qualité dans la littérature (voir par exemple Yao et Zhong (1999); Guillet (2004) pour une liste de plus de quarante mesures de qualité). Nous nous sommes restreints dans ce travail à vingt mesures objectives, et strictement décroissantes en fonction de $n_{a\bar{b}}$. Selon nous, ces conditions sont des critères d'éligibilité pour des mesures de qualité de règles d'association (Lenca et al., 2003).

On montre dans Vaillant et al. (2004) que ces mesures mettent en avant différentes règles. Afin de guider le choix d'une mesure adaptée au contexte et aux souhaits de l'expert, nous avons étudié les vingt mesures selon neuf propriétés pertinentes dans le cadre de la fouille de

règles et faisant sens pour un expert métier (voir tableau 2, et Vaillant (2006) pour la sémantique détaillée et l'évaluation de ces propriétés sur les vingt mesures). Certaines propriétés sont considérées comme étant normatives (*i.e.* les préférences sur les modalités sont indépendantes du contexte applicatif), d'autres comme subjectives et dépendantes de préférences utilisateurs. Afin de mettre en œuvre un processus d'aide à la décision, il sera alors nécessaire que l'expert métier précise un ordre de préférence sur les modalités des propriétés subjectives.

Critère	Sémantique	Nombre de modalités	Responsabilité
g_1	non symétrie selon A et B	2	$E_{\mathcal{A}}$
g_2	décroissance avec n_b	2	$E_{\mathcal{A}}$
g_3	situation à la règle logique	2	$E_{\mathcal{A}}$
g_4	situation à l'indépendance	2	$E_{\mathcal{A}}$
g_5	situation à l'indétermination	2	$E_{\mathcal{A}}$
g_6	tolérance aux premiers contre-exemples	3	$E_{\mathcal{R}}$
g_7	prise en compte de n	2	$E_{\mathcal{R}}$
g_8	facilité à fixer un seuil	2	$E_{\mathcal{A}}$ et $E_{\mathcal{R}}$
g_9	intelligibilité	3	$E_{\mathcal{R}}$

TAB. 2 – *Propriétés de mesures de qualité*

3 Contexte applicatif

Nous appliquons un processus d'ECD à des données fournies par la société DIXID, concernant un service d'assistance. Plus précisément, un utilisateur ayant un problème avec une application professionnelle sur son ordinateur peut appeler le serveur vocal concerné. Le serveur vocal lui demande alors quel est son problème afin de l'orienter vers un service spécifique d'assistance, dont les opérateurs sont formés pour répondre au mieux aux problèmes posés (au total, 9 services d'assistance différents existent, avec plusieurs opérateurs pour chaque service). Dans l'idéal, l'utilisateur expose son problème en prononçant le nom de l'application concernée (mais bien sûr cette situation est "idéale") et il est orienté vers le "bon" opérateur. Un appel dure en moyenne près de 40 secondes (avec une amplitude de 1 à 141 secondes). Nos objectifs sont multiples. Nous souhaitons analyser le système mis en œuvre et en juger les performances. De plus, la société DIXID travaillant dans le domaine de l'ergonomie, nous souhaitons voir s'il est possible d'améliorer celle du système vocal. C'est en vue d'atteindre ce dernier objectif que nous nous sommes intéressés en première approche aux règles concluant sur le résultat des appels (voir tableaux 3 et 7).

Les données à notre disposition portent sur 2006 appels dans leur globalité, et les caractérisent selon 8 attributs discrets listés dans le tableau 3. Il n'y a pas de données manquantes. D'autres données, quantitatives, décrivant de façon précise les différentes interactions entre l'utilisateur et le service vocal n'ont pas été prises en compte pour le moment. Pour autant, malgré les difficultés liées à leur exploitation, l'expert des données pense que ces dernières sont riches d'information.

On peut d'ores et déjà remarquer que certaines variables sont obligatoirement liées (par exemple, un appel avec résultat "raccrochage après flash" aura évidemment une modalité "oui"

Variable	modalités (fréquence)	sémantique
NB_FEEDBACK_OK	0 (680), 1 (1184), 2 (123), 3 (17), 4 (2)	Nombre de prompts du serveur montrant une compréhension correcte de la demande de l'utilisateur
NB_FEEDBACK_NOK	0 (1675), 1 (251), 2 (57), 3 (15), 4 (4), 5 (3), 6 (1)	Nombre de prompts du serveur montrant une compréhension incorrecte de la demande de l'utilisateur
EXPOSITION_PB	aucune_precision (14), avec_un_nom (1760), ne_dit_rien (204), probleme_hors_sujet (3), sans_nom (25)	Manière d'exposer le problème.
REPETITION	oui (283), non (1723)	L'utilisateur répète mot à mot, au moins une fois, une demande.
REFORMULATION	oui (181), non (1825)	L'utilisateur explique un même problème en utilisant des termes différents, au moins une fois.
NON_REPONSE	0 (1792), 1 (158), 2 (54), 4 (2)	Nombre de non-réponse de l'utilisateur à une question explicite posée par le serveur.
DIFFUSION_FLASH	oui (41), non (1965)	Lorsque l'utilisateur expose son problème, un message "flash" concernant une certaine application est automatiquement diffusé.
RESULTAT_APPEL	bon_operateur (1369), erreur_d_operateur (244), operateur_par_defaut (160), raccrochage_apres_description (102), raccrochage_immediat (28), raccrochage_sans_description (42), raccrochage_sans_rien_dire (39), raccrochage_apres_flash (22)	Résultat de l'appel.

TAB. 3 – *Attributs décrivant les données* DIXID

pour l'attribut DIFFUSION_FLASH...). De plus, certaines modalités de plusieurs attributs sont très peu fréquentes. Afin d'ignorer le moins d'information possible, nous avons ainsi fixé le seuil de support le plus bas possible, soit 1% avec l'implémentation d'APRIORI que nous utilisons. Nous fixons le seuil de confiance à 50%, afin de générer des règles ayant plus d'exemples que de contre-exemples. Enfin, nous avons précisé que les règles pouvaient faire intervenir entre 2 et 8 attributs, l'implémentation utilisée se limitant par défaut à une taille maximale de 5. Cette extension volontaire des règles à la plus grande taille possible est un choix en première approche, visant à vérifier qu'il n'existe pas de règle de grande taille porteuse d'information. La redondance entre règles apportée par ce choix, ainsi que la difficulté d'interprétation des règles longues incitera à revenir au paramètre par défaut d'APRIORI dans les études futures.

Avec ces paramètres, APRIORI génère 8921 règles, sur la base des 2006 objets. Si l'on se restreint aux règles faisant intervenir l'attribut RESULTAT_APPEL, qui est celui qui nous intéresse si l'on souhaite analyser l'ergonomie du serveur vocal, il reste 6061 règles à évaluer.

En étant encore plus strict et en ne considérant que les règles faisant intervenir cet attribut en conclusion, on retient 626 règles, ce qui est toujours trop volumineux pour une étude manuelle exhaustive. Les 626 règles retenues par la suite peuvent ainsi être vues comme un résultat d'un apprentissage supervisé sur les différentes modalités de l'attribut RESULTAT_APPEL.

Comme mentionné précédemment, certaines des règles générées sont sans grand intérêt, comme `DIFFUSION_FLASH='oui' -> RESULTAT_APPEL='raccrochage_apres_flash'` (support : 1.1%, confiance : 53.7%).

4 Choix de mesure de qualité

Afin de guider l'analyse de l'ensemble de règles généré, nous procédons à un filtrage par des mesures de qualité. En se basant sur la prise en compte des préférences de l'utilisateur expert des données, nous sélectionnons trois mesures parmi les vingt étudiées, et retenons les 20 règles les mieux classées par chacune d'entre elles.

La modélisation des préférences du décideur comporte deux aspects. Dans un premier temps, il est nécessaire de préciser l'ordre dans lequel les modalités des propriétés subjectives sont à considérer, afin de transformer ces propriétés en critères de choix. Puis, chaque critère est affecté d'un poids relatif à son importance, telle que exprimée par les différents intervenants. Un troisième aspect pourrait être considéré dans la méthode d'aide multicritère à la décision que nous utilisons, mais il ne fait pas sens dans notre cas : le choix d'une fonction de préférence (à valeurs dans $[0, 1]$), permettant de rendre commensurable les différents critères. N'ayant que des critères à deux ou trois modalités, nous avons retenu le modèle usuel de préférence.

Après explicitation des propriétés, l'expert des données a ordonné les modalités des propriétés contextuelles comme suit, exprimant ainsi ses préférences sur chaque critère pris individuellement :

g_6 N'étant pas dans une situation où les contre-exemples sont critiques, on préférera une décroissance linéaire de la mesure en fonction des contre-exemples autour de 0^+, puis une décroissance faible, et enfin une décroissance forte.

g_7 La prise en compte d'une modélisation statistique étant souhaitée, les mesures construites par cette approche seront préférées aux mesures descriptives.

g_8 On peut souhaiter opérer une validation statistique des règles retenues par la mesure. Une mesure se prêtant à un tel calcul sera donc préférée à une autre ne s'y prêtant pas.

g_9 L'interprétation de la valeur prise par la mesure ayant de l'importance afin de communiquer les résultats, plus la mesure est simple à comprendre plus elle est préférée.

Pour déterminer les poids de chaque critère, nous avons exploré deux scénarios. Le premier (SC1) consiste à ne reposer que sur l'avis de l'expert des données pour le fixer. Cette approche n'a pas abouti à des résultats satisfaisants à notre avis, mais permet de confirmer une assertion courante. Elle illustre aussi les difficultés liées à l'expression de besoins de l'utilisateur, particulièrement lorsque ce dernier n'est pas expert de ECD. Le second (SC2) consiste à prendre en compte l'avis de l'expert des données sur les propriétés subjectives, et à fixer les poids des propriétés normatives selon l'avis de l'expert de la fouille de données. Les jeux de poids ainsi déterminés sont listés dans le tableau 4. L'importance de chaque critère est évaluée sur une échelle entière entre 0 (sans importance) et 10 (forte importance).

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
SC1	8	6	2	0	7	4	2	7	3
SC2	9	7	6	6	5	4	2	7	3

TAB. 4 – *Jeux de poids utilisés pour sélectionner les mesures de qualité*

5 Résultats

A partir de ces préférences exprimées, nous avons utilisé la méthode d'aide multicritère PROMETHEE (Brans et al., 1984). Cette méthode de surclassement permet de constituer un flot agrégé sur les mesures afin de prendre en compte les préférences utilisateur. A partir de ce flot de surclassement, on peut alors disposer d'un ordre sur les mesures, de celles répondant au mieux aux souhaits exprimés à celles y répondant le moins. Le tableau 5 présente les trois meilleures mesures pour chaque jeu de poids, ainsi que les deux suivantes. La méthode PROMETHEE permet également de procéder à une étude de stabilité du jeu de poids. Le tableau 6 donne l'intervalle dans lequel on peut faire varier le poids associé à chaque critère pris individuellement, sans que cela ait un impact sur les trois premières mesures du rangement.

	rang 1	rang 2	rang 3	rang 4	rang 5	...
SC1 :	CONF (0.24)	LOE (0.17)	CONFCEN (0.16)	MOCO (0.14)	TEC (0.10)	
SC2 :	LOE (0.24)	FB (0.17)	INTIMP (0.16)	CONFCEN (0.15)	CONV (0.13)	

TAB. 5 – *Rangements et flots de préférence agrégés des premières mesures*

	g_1	g_2	g_3	g_4	g_5
SC1 :	$[0, +\infty]$	$[5.2, 8.8]$	$[1.8, 4.2]$	$[0, 2.8]$	$[4.2, 7.8]$
SC2 :	$[0, +\infty]$	$[4.45, +\infty]$	$[5.45, +\infty]$	$[3.45, +\infty]$	$[0, 7]$

	g_6	g_7	g_8	g_9
SC1 :	$[1.5385, +\infty]$	$[0, 5.45]$	$[3.55, +\infty]$	$[1.6667, 3.3333]$
SC2 :	$[1.3846, 4.5556]$	$[1.45, 2.25]$	$[5, +\infty]$	$[2.8148, 3.4074]$

TAB. 6 – *Stabilité des jeux de poids*

Le fait que la confiance soit classée première pour SC1 peut paraître surprenant, ou décevant. En effet, cette mesure ne prenant en compte que deux grandeurs et étant un filtre utilisé par APRIORI, on peut s'attendre à ne pas obtenir de résultats satisfaisants. Toutefois, ce résultat confirme une assertion courante : la confiance est très souvent utilisée en pratique pour sa simplicité et son intelligibilité. LOE est très bien classée dans les deux scénarios, elle réalise un bon compromis, comme déjà remarqué dans Lenca et al. (2007).

Lors de la sélection des 20 meilleures règles, CONF $= \frac{n_{ab}}{n_a}$ et LOE $= 1 - \frac{n n_{a\bar{b}}}{n_a n_{\bar{b}}}$ en sélectionnent effectivement 20, les autres mesures en retenant plus, à cause d'ex-aequo : 32 règles

sont retenues par CONFCEN $= \frac{nn_{ab}-n_an_b}{nn_a}$ et FB $= \frac{n_{ab}n_{\bar{b}}}{n_bn_{a\bar{b}}}$ (ce sont de plus les mêmes, les différences d'évaluation des règles par ces mesures n'apparaissent qu'en suite), et 98 règles sont retenues par INTIMP $= P\left[N(0,1) \geq \frac{n_an_b-nn_{ab}}{\sqrt{n}p_ap_{\bar{b}}}\right]$. Dans le cas de cette dernière, on observe la non-discriminance de la mesure. Il est à noter que de très nombreuses règles retenues sont des spécialisations d'autres règles. Ceci est dû au fait que nous ayons "forcé" l'algorithme à explorer totalement l'espace de recherche. En se limitant à 5 items par règle, un nombre important de telles règles ne seraient pas apparues.

Les seules modalités de conclusion retenues par nos mesures sont `bon_operateur` et `operateur_par_defaut`, ce qui ne nous permet donc de n'évaluer que les points positifs du système. En étudiant les règles non-supervisées, CONF, LOE et FB évaluent 2523 règles au maximum. Ce sont les règles n'ayant pas de contre-exemples. En revanche INTIMP ne sélectionne "que" 434 règles, qui ne sont pas nécessairement logiques ce qui cette fois illustre l'avantage d'avoir recours à un modèle statistique.

6 Conclusions et perspectives

Nous avons illustré dans cet article la mise en pratique d'une méthode d'aide multicritère à la décision afin de guider le choix d'une mesure de qualité en vue de quantifier la qualité de règles d'association, selon des préférences exprimées par des utilisateurs experts. Les résultats produits attestent de la difficulté de cette tâche, tant dans le choix des paramètres de l'algorithme de fouille de données que dans la retranscription du système de préférence expert.

Les règles sélectionnées par les mesures ont été soumises à l'expert des données afin d'affiner les paramètres choisis dans cette première approche. Il est apparu que certaines des mesures retenues n'étaient pas assez discriminantes, alors que d'autres retenaient le même ensemble de règles. Ces aspects expérimentaux pourraient être inclus au processus de sélection de mesure. En outre, des informations plus précises relatives aux différents appels sont à notre disposition. L'exploitation de ces données pourrait permettre de mieux analyser les appels. On se heurte toutefois à des problèmes de temporalité ou de causalité liés à la nature même de cette information.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *Proceedings of the 20th Very Large Data Bases Conference*, pp. 487–499. Morgan Kaufmann.

Borgelt, C. et R. Kruse (2002). Induction of association rules : APRIORI implementation. In *Proceedings of the 15th Conference on Computational Statistics*, Heidelberg, Germany. Physika Verlag.

Brans, J.-P., B. Mareschal, et P. Vincke (1984). PROMETHEE *: A New Family of Outranking Methods in MCDM*. IFORS 84.

Brisson, L. (2004). Mesures d'intérêt subjectif et représentation des connaissances. Technical Report ISRN I3S/RR-2005-35-FR, Université de Nice.

Carvalho, D., A. A. Freitas, et N. Ebecken (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. In *Knowledge Discovery in Databases : Proceedings of PKDD 2005, LNAI 3731*, pp. 453–461. Springer Verlag.

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Freitas, A. (1999). On rule interestingness measures. *Knowledge-Based Systems journal*, 309–315.

Guillet, F. (2004). Mesure de la qualité des connaissances en ECD. Tutoriel de la 4e Conf. Extraction et Gestion des Connaissances. 60 pages.

Hilderman, R. J. et H. J. Hamilton (2000). Applying objective interestingness measures in data mining systems. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pp. 432–439. Springer-Verlag.

Hilderman, R. J. et H. J. Hamilton (2001). *Knowledge Discovery and Measures of Interest*, Volume 638 of *The International Series in Engineering and Computer Science*. Kluwer.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de données)* (1), 123–134.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2007). On selecting interestingness measures for association rules : user oriented description and multiple criteria decision aid. *European Journal of Operational Research*, To appear.

Liu, B., W. Hsu, S. Chen, et Y. Ma (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems 15*(5), 47–55.

Silberschatz, A. et A. Tuzhilin (1995). On subjective measures of interestingness in knowledge discovery. In *Knowledge Discovery and Data Mining*, pp. 275–281.

Tan, P.-N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on KDD*, pp. 32–41.

Vaillant, B. (2006). *Mesurer la qualité des règles d'association : études formelles et expérimentales*. Ph. D. thesis, École nationale supérieure de télécommunications de Bretagne/Université de Bretagne Sud.

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. In *Discovery Science*, pp. 290–297.

Yao, Y. Y. et N. Zhong (1999). An analysis of quantitative measures associated with rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 479–488.

Summary

Knowledge discovery in databases aims at extracting information from large datasets through the use of automated tools, in order to re-use this information in industrial or decision oriented applications. In this paper, we explore a datamining process applied to the use of a vocal answering machine. We particularly focus on the crucial knowledge validation step, relying in our approach on the use of interestingness measures.

TAB. 7 – *Ensemble des* 20 *meilleures règles pour* CONF *(20 règles)*

id	prémisse	conclusion	CONF	n_a	n_b	n_{ab}	$n_{\bar{a}b}$
451	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98611	72	1369	71	1
264	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98611	72	1369	71	1
261	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom'	RESULTAT_APPEL='bon_operateur'	0.98611	72	1369	71	1
106	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non'	RESULTAT_APPEL='bon_operateur'	0.98611	72	1369	71	1
577	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and NON_REPONSE=0 and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98551	69	1369	68	1
453	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and NON_REPONSE=0 and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98551	69	1369	68	1
449	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and NON_REPONSE=0	RESULTAT_APPEL='bon_operateur'	0.98551	69	1369	68	1
262	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and NON_REPONSE=0	RESULTAT_APPEL='bon_operateur'	0.98551	69	1369	68	1
578	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and REFORMULATION='non' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98333	60	1369	59	1
454	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and REFORMULATION='non' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98333	60	1369	59	1
450	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and REFORMULATION='non'	RESULTAT_APPEL='bon_operateur'	0.98333	60	1369	59	1
263	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and REFORMULATION='non'	RESULTAT_APPEL='bon_operateur'	0.98333	60	1369	59	1
620	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and NON_REPONSE=0 and REFORMULATION='non' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98276	58	1369	57	1
579	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and NON_REPONSE=0 and REFORMULATION='non' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.98276	58	1369	57	1
576	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and EXPOSITION_PB='avec_un_nom' and NON_REPONSE=0 and REFORMULATION='non'	RESULTAT_APPEL='bon_operateur'	0.98276	58	1369	57	1
452	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and REPETITION='non' and NON_REPONSE=0 and REFORMULATION='non'	RESULTAT_APPEL='bon_operateur'	0.98276	58	1369	57	1
267	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and EXPOSITION_PB='avec_un_nom' and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.97030	101	1369	98	3
110	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and DIFFUSION_FLASH='non'	RESULTAT_APPEL='bon_operateur'	0.97030	101	1369	98	3
107	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0 and EXPOSITION_PB='avec_un_nom'	RESULTAT_APPEL='bon_operateur'	0.97030	101	1369	98	3
28	NB_FEEDBACK_OK=2 and NB_FEEDBACK_NOK=0	RESULTAT_APPEL='bon_operateur'	0.97030	101	1369	98	3

Quelques erreurs et abus courants dans l'interprétation des résultats de la fouille des données

Jean-Hugues Chauchat*, Annie Morin**

*Université de Lyon, ERIC-Lyon2, 5 avenue Pierre-Mendès-France 69676 Bron Cedex France
jean-hugues.chauchat@univ-lyon2.fr
**IRISA, Université de Rennes 1, Rennes Cedex 35042
amorin@irisa.fr

Résumé. L'interprétation des résultats de la fouille des données (data mining) est souvent erronée car on néglige le processus de constitution du corpus des données utilisées. A l'aide de la théorie des sondages, on donne ici des solutions dans les cas où premièrement il faut redresser les résultats car les proportions des classes à apprendre sont différentes dans la base d'apprentissage et dans l'univers de référence, et deuxièmement il faut corriger l'estimation de la qualité des résultats car les individus (cas) de la base ont été collectés en grappes ou par sondages à deux degrés. Par ailleurs, la fouille des données met en évidence des corrélations et non des causes. Les modèles produits sont prévisionnels et non causaux car on travaille souvent sur des données de seconde main, non issues d'un plan d'expérience.

1 Introduction

Pour éviter les critiques usuelles concernant la fouille des données (data mining), la pêche dans les données (data fishing), sans méthode et sans but (Jensen, 2006), on présente quelques erreurs courantes d'interprétation des résultats de la fouille des données. Cette analyse s'appuie sur le corps de réflexion de la statistique : - la théorie des sondages pour le redressement des échantillons stratifiés et l'estimation de la variance avec des échantillons en grappes ou à plusieurs degrés, - l'épistémologie des modèles pour souligner la différence de nature entre les données de seconde main et celles issues d'un plan d'expériences contrôlé.

2 Prise en compte du processus de constitution du corpus

On rencontre souvent des erreurs d'extrapolation dues à la non prise en compte du processus de constitution du corpus des données fouillées.

2.1 Taux de sondage inégaux dans les classes

On utilise souvent un corpus d'apprentissage correspondant à un échantillon stratifié selon les classes à apprendre. C'est à dire que la population de référence (l'univers) est divisée

en sous-populations correspondant chacune à une classe ; dans chaque classe, on a extrait un échantillon aléatoire, par tirages équiprobables et indépendants.

Souvent les classes ont, dans la population, des effectifs très déséquilibrés tandis que dans la base d'apprentissage, leurs effectifs sont quasi équilibrés : le taux de sondage est inégal d'une classe à l'autre et les résultats doivent être redressés, comme cela est connu en théorie de l'échantillonnage. Exemple : la base contient autant de malades que de non malades (50% et 50%), alors qu'il y a x% (par exemple 4%) de malades dans l'univers ; dans tous les calculs de fréquence rapportés à l'univers, on doit appliquer le théorème de Bayes, ou ce qui revient au même, on doit donner un poids relatif de (ω = x% / 50%, soit 4%/50%=0,08) à chaque malade de l'échantillon, et (100%-x%)/50% soit 96%/50%=1,92) à chaque non-malade de l'échantillon. Ce redressement est rarement fait ; c'est le cas en particulier des résultats produits par SAS Entreprise Miner. D'une façon générale, ce module de SAS ne demande pas la répartition des classes dans l'univers de référence.

Pour ceux qui ne sont pas familiers avec la théorie de l'échantillonnage, le théorème de Bayes permet de calculer ces redressements comme le montre la figure 1. Soit un univers découpé en J classes, chacune de fréquence $\pi_{j.}, j = 1, ..., J$, et un échantillon stratifié comprenant $n_{j.}$ individus par classe, $\sum_j n_{j.} = n$. Si en fin d'apprentissage, la feuille l d'un arbre contient n_{jl} individus de la classe j, on peut estimer la fréquence dans l'univers d'une classe k $P_l(k) = \pi_{k/l}$; en effet, la probabilité a priori d'une classe j est sa fréquence $\pi_{j.}$ dans l'univers et la probabilité conditionnelle de la feuille l dans la classe j est estimée par $P_j(l) = n_{jl}/n_{j.}$.

Le théorème de Bayes s'écrit :

$$P_l(k) = [\pi_{k.} \times n_{kl}/n_{k.}] \,/\, \left[\sum_j \pi_{j.} \times n_{jl}/n_{j.}\right] = [\omega_k \times n_{kl}] \,/\, \left[\sum_j \omega_{j.} \times n_{jl}\right]$$

où $\omega_j = \pi_{j.}/\left(n_{j.}/n\right)$ est le poids de chaque individu de l'échantillon de la classe j (les n se simplifient).

Dans l'exemple proposé ici, la classe d'intérêt représente 11% du sous-ensemble de l'univers caractérisé par la feuille, ce qui est plus que les 4% de l'univers entier, mais moins que les 75% observé dans le sous-ensemble de l'échantillon caractérisé par la feuille.

2.2 Données arrivées en grappes, ou par sondage à deux degrés

Souvent le processus de constitution du corpus disponible fait que les individus (les cas) sont arrivés « en grappes », ou par sondage « à deux degrés », (exemples : - tous les patients, ou un échantillon de patients, d'un échantillon d'hôpitaux ; - toutes les carottes d'un échantillon de forages ; - un échantillon de phrases prononcées par un échantillon de locuteurs) : les procédures de validation par ré-échantillonnage doivent tenir compte de ce procédé particulier de « sondage » par lequel le corpus a été constitué ; par exemple en « validation croisée », les sous-corpus placés successivement en « échantillon test » doivent être constitués de grappes entières : on doit faire l'apprentissage sur les malades d'un sous-ensemble d'hôpitaux et tester sur les malades des autres hôpitaux.

Nous devons insister ici sur le fait que ces grappes (clusters) sont liées au processus de constitution du corpus (le statisticien dirait à la méthode d'échantillonnage dans l'univers de référence), et non à un traitement de classification non supervisée (clustering) sur le corpus

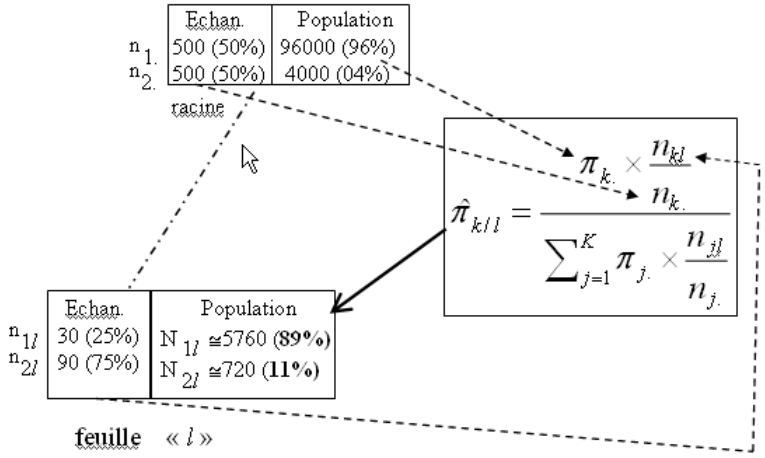

FIG. 1 – *Redressement des résultats par le théorème de Bayes*

déjà constitué. Ce processus de constitution de la base est une "méta-donnée" qu'on ne peut pas déduire des données elles-mêmes.

Les travaux de Chauchat et Rakotomalala (Chauchat et al., 2002) et (Rakotomalala et al., 2006) détaillent l'adaptation nécessaire des procédures de validation croisée ; les résultats de l'apprentissage sur des données simulées montrent que, à taille de corpus donnée :

- le vrai taux d'erreur (en généralisation) augmente avec la dispersion relative des grappes, et décroît avec la taille des grappes ;
- la validation croisée standard, ignorant l'effet de grappes, sous-estime fortement ce taux ; le biais relatif augmente avec la taille des grappes ;
- le biais d'estimation du taux d'erreur croît avec l'effet de grappes : ce dernier est maximum quand tous les individus d'une grappe sont identiques ;
- la validation croisée tenant compte de l'effet de grappe sur-estime légèrement le vrai taux d'erreur ; ceci était attendu car la validation croisée n'utilise, à chaque étape, qu'une fraction de l'échantillon disponible pour construire le modèle de prédiction.

Pellegrino et al. (2002) présentent une application sur des données réelles utilisées pour la reconnaissance automatique de la langue parlée. L'objectif de l'apprentissage est de reconnaître l'une des 5 langues parlées à partir du signal physique d'une phrase prononcée par un locuteur inconnu. Le corpus rassemble les enregistrements de 150 phrases dans chaque langue, soit $5 \times 150 = 750$ cas. Pour plusieurs méthodes d'apprentissage, la validation croisée standard donne des estimations du taux d'erreur autour de 15%. Mais ces estimations sont biaisées car les 750 phrases (cas) n'ont pas été choisies indépendamment. Le corpus utilisé a été constitué avec un échantillon de 50 locuteurs, 10 dans chaque langue, chacun disant entre 10 et 20 phrases (15 en moyenne). Ici l'ensemble des phrases dites par un locuteur constitue donc une grappe et on dispose de 50 grappes choisies au hasard indépendamment dans l'univers des

locuteurs et non de 750 phrases choisies indépendamment dans l'univers de ce que l'on pourrait entendre. La validation croisée tenant compte des grappes donne un taux d'erreur estimé autour de 20% au lieu de 15%.

Algorithme	Estimation "standard" biaisée	Estimation respectant les grappes
Analyse discriminante linéaire	15%	20%
Perceptron multicouche	16%	21%
Maximisation de la vraisemblance pour un mélange de gaussiennes dans $\mathbb{R}^5$		20%

Le statisticien dirait qu'il y a un fort "effet de grappe" car le rythme de la parole diffère certes d'une langue à l'autre et d'une phrase (cas) à l'autre, mais aussi d'un locuteur (grappe) à l'autre. Comme l'objectif est d'apprendre à reconaître la langue parlée par une nouvelle personne, différente de celles du corpus disponible, l'estimation réalisée sans tenir compte des grappes est "biaisée" ; le résultat de 15% de taux d'erreur en généralisation est trop optimiste.

Pour construire un estimateur non-biaisé du taux d'erreur, il faut connaître a priori la partition des 750 "cas" (les phrases) selon les locuteurs. Soulignons que cette information ne peut être déduite des données elles-mêmes.

3 Dans l'interprétation, on confond souvent corrélation et causalité

3.1 L'éventuelle causalité ne peut être établie que par un raisonnement extérieur aux données elles-mêmes

Il faut sans cesse rappeler que les résultats de la fouille des données ne mettent en évidence que des corrélations (Rossman, 1994). L'éventuelle causalité ne peut être établie que par un raisonnement extérieur aux données elles-mêmes. Voici un exemple d'une telle erreur : à l'issue d'un stage de fin de Master dans une grande banque de dépôt, un étudiant présente un score permettant de repérer les clients susceptibles de passer à la concurrence. En résumé, le score est de la forme :

(nombre de réclamations du client – nombre de cartes de crédit et de polices d'assurance souscrites récemment).

Ce score est raisonnable a posteriori et le directeur de la banque est très satisfait du travail. Il en conclut devant nous qu'une campagne marketing poussant les clients à souscrire des polices d'assurance va les fidéliser. Il confond donc corrélation et causalité, modèle prévisionnel et modèle causal. La fouille des données établit un modèle prévisionnel qui permet de conclure que (toutes choses égales par ailleurs) les clients qui ont fait des réclamations et qui n'ont rien souscrit récemment quittent plus fréquemment la banque, et ces corrélations sont symétriques : les clients qui quittent la banque ont plus souvent fait des réclamations et ont moins souvent souscrit de nouveaux contrats. Mais ces résultats ne permettent pas de conclure que l'un ou l'autre de ces comportements soit la cause du départ ; ceci signifierait qu'on pourrait réduire

les départs en empêchant les clients de réclamer ou bien en leur faisant souscrire de nouveaux contrats.

A notre avis (raisonnement extérieur aux données), la vraie cause de la fidélité est la satisfaction du client, qui concourt à la fois à faire acheter de nouveaux services, et à réduire ses réclamations ; et cette satisfaction peut baisser fortement si on appelle chez lui un client mécontent pour essayer de lui vendre un nouveau produit ; mieux vaudrait sans doute répondre correctement à ses réclamations !

Un débat du même type agite actuellement les médecins gériatres : la fouille des dossiers médicaux montre que les personnes âgées sous-alimentées sont plus souvent malades. Cette corrélation est établie mais elle ne dit pas ce qu'il faut faire : - est-ce la sous-alimentation qui provoque la maladie (et alors il faut obliger les personnes âgées à manger) ? - est-ce la maladie qui coupe l'appétit (et alors il faut soigner les malades pour qu'ils retrouvent l'appétit) ?

3.2 Seuls les résultats d'expériences « randomisées » peuvent fournir une preuve statistique, ou un début de preuve

On sait en statistique que seuls les résultats d'expériences « randomisées » peuvent fournir une preuve, ou un début de preuve, d'une causalité posée en hypothèse a priori. Lors d'une telle expérience, les individus sont tirés au hasard dans l'univers, puis affectés au hasard aux différentes modalités de la variable d'intérêt. Cette méthode a été créée par sir Ronald Fisher dans les années 1920-1930 (Fisher, 1935) pour tester l'efficacité des types d'engrais, de semences et de modes de culture, ainsi que leurs interactions, sur la quantité récoltée ; les combinaisons étaient réparties au hasard sur chaque champ utilisé. Notons que Fisher a mis conçu les plans d'expérience après avoir tenté sans succès d'exploiter (on dirait aujourd'hui de fouiller) les données de 90 ans de relevés météo et culturaux de la station agronomique de Rothamsted en Angleterre. Ces plans d'expérience ont été étendus en gestion de la qualité industrielle dans la "méthode Taguchi" (Pillet, 1997). En recherche pharmaceutique, les plans d'expérience randomisés et en double aveugle constituent la seule méthode acceptée par les autorités européennes et américaines pour prouver l'efficacité d'un médicament et permettre la délivrance de l'AMM (Autorisation de Mise sur le Marché)

Or on souligne souvent dans la définition du « data mining » que ces méthodes travaillent sur des données de seconde main, c'est à dire recueillies dans un autre but que la fouille elle-même. On est donc loin de l'expérience randomisée et des tests d'hypothèses causales qu'elle permettrait. Autrement dit, la fouille des données est utile pour repérer des corrélations et construire des modèles prévisionnels du type : les patients (les clients) qui ont telles caractéristiques ont plus de chance d'avoir telle maladie (comportement). Mais la fouille par elle-même ne peut pas établir des relations causales du type : en agissant sur telle caractéristique on améliorera la santé des patients ou ou modifiera le comportement du client.

4 Conclusion

En conclusion, avant de fouiller les données, il faut chercher à savoir comment la base de données disponible a été constituée à partir de l'univers de référence. Pour cela il faut pouvoir interroger celui qui a constitué la base. Notons que cela est souvent impossible pour les bases "benchmark", telles celles de UCI. De plus, la fouille des données apporte une aide

à la décision, mais ne construit pas la décision elle-même ; celle-ci doit tenir compte de méta-données concernant le corpus analysé et de la « connaissance - métier » du décideur. Ces réflexions sont issues de l'observation d'interprétations souvent abusives des résultats de la fouille de données, tant dans le monde académique que dans celui des entreprises.

Références

Chauchat, J.-H., R. Rakotomalala, et F. Pellegrino (2002). Estimation du taux d'erreur sur données en grappes - application à la reconnaissance de la parole. In *"Revue Extraction des Connaissances et Apprentissage "*, Volume 1-4, pp. 269–280.

Fisher, R. (1935). *The Design of Experiments*. Springer Verlag. (Paperback - Aug 9, 1990 Statistical Methods, Experimental Design, and Scientific Inference : A Re-issue of Statistical Methods for Research Workers, The Design of Experiments, and Statistical Methods and Scientific Inference by R. A. Fisher, F. Yates, and J. H. Bennett).

Jensen, D. (2006). Data snooping, dredging and fishing : The dark side of data mining. a sigkdd99 panel report. In *SIGKDD99*. http ://www.acm.org/sigs/sigkdd//explorations/issues/1-2-2000-01/jensen.pdf.

Pellegrino, F., J.-H. Chauchat, R. Rakotomalala, et J. Farinas (2002). Can automatically extracted rhythmic units discriminate among languages ? In *Speech Prosody 2002 , Aix-en-provence, France, 11/04/02-13/04/02*. Congres : http ://www.lpl.univ-aix.fr/sp2002/.

Pillet, M. (1997). *Les plans d'expériences par la méthode Taguchi*. Paris : Les Editions d'organisation.

Rakotomalala, R., J.-H. Chauchat, et F. Pellegrino (2006). Accuracy estimation with clustered dataset. In C. Peter, P. J. Kennedy, J. Li, S. J. Simoff, et G. J. Williams (Eds.), *Fifth Australasian Data Mining Conference (AusDM2006)*, Volume 61 of *CRPIT*, Sydney, Australia, pp. 17–22. ACS.

Rossman, A. J. (1994). Televisions, physicians, and life expectancy. *Journal of Statistics Education 2-2*, 679–696.

Summary

The interpretation of datamining results are often wrong because the way the database has been built is not taken into account. We use sampling theory to provide some solutions to this problem: first, when the proportions of classes to learn are not the same in the learning database and in the universe, we need to weight the results. Besides, when data have been collected in clusters or through a two-stage survey, we must correct the estimation of the quality of results. Then, datamining points out correlation and not causality.

Index des auteurs

– B –
Briand, H., 35

– C –
Cadot, M., 15
Chauchat, J.-H., 65
Clément, D., 5

– D –
David, J., 35

– G –
Gras, R., 35
Guillet, F., 35

– L –
Laboisse, B., 5
Lallich, S., 45, 55
Lecornu, L., 25
Lelu, A., 15
Lenca, P., 45, 55

– M –
Menou, S., 55
Moga, S., 55
Morin, A., 65

– S –
Solaiman, B., 25

– V –
Vaillant, B., 45, 55

– Z –
Zemirline, A., 25