Extraction et Gestion de Connaissance

Brest, 25 janvier 2011

Atelier Extraction de Connaissance et Santé

Responsables

A. Baneyx, Sciences Po
S. Bringay, LIRMM, Univ. Montpellier 3
N. Souf, CERIM, Univ. Lille 2

Un Automate Cellulaire pour la Détection de Spam

Fatiha. BARIGOU *, Baghdad ATMANI**, Naouel. BARIGOU***
Laboratoire d'Informatique d'Oran, Equipe SIF
Université d'Oran, BP 1524, El M'Naouer, 31 000 Oran, Algérie
*fatbarigou@gmail.com, **atmani.baghdad@univ-oran.dz , ***barigounaouel@gmail.com

Résumé. Dans le contexte du filtrage de courriels indésirables (appelé aussi spam), nous proposons l'utilisation d'une classification supervisée booléenne à base d'automate cellulaire. Nous examinons, par des expériences sur le corpus LingSpam, la performance de cette approche en variant les méthodes de prétraitement du corpus : utilisation d'une stop-liste, racinisation, et sélection des termes.

1 Introduction

Aujourd'hui, le courriel est devenu un moyen rapide et économique pour échanger des informations. Cependant, on se retrouve assez vite submergés de quantités de messages indésirables appelé aussi spam. Pour faire face à cette charge croissante des spam, de nombreuses solutions ont vu le jour (Sanz et al, 2008). Certaines solutions sont basées sur l'en-tête du courrier électronique tel que les listes noires, les listes blanches et grises. D'autres solutions sont basées sur le contenu textuel du message tel que le filtrage à base d'apprentissage (Guzella et Caminhas, 2009). Dans ce papier, nous proposons pour la première fois une approche de détection de spam basée sur l'induction symbolique par automate cellulaire (Atmani et Beldjilali, 2007). Le principe de cette approche consiste à construire un modèle booléen à partir d'un ensemble de courriels d'apprentissage pour la classification des emails entrant en spam ou légitime. La suite de cet article est organisée de la manière suivante : la section 2 est consacrée à l'étude de l'approche proposée. La section 3 présente l'étude expérimentale pour l'évaluation de cette nouvelle solution. La section 4 présente nos conclusions et quelques orientations pour les travaux futurs.

2 Approche Cellulaire de Classification

Cette section est consacrée à l'étude de la classification supervisée à base d'automate cellulaire adoptée pour la détection de spam. Le principe de cet automate est tout d'abord décrit.

2.1 Principe de l'automate cellulaire CASI

CASI (Cellular Automata for System Induction) issue des travaux de (Atmani et Beldjilali, 2007) est une méthode cellulaire de génération, de représentation et d'optimisation des graphes d'induction (Zighed, 2000) générés à partir d'un ensemble d'exemples d'apprentissage. Ce système cellulo-symbolique est organisé en cellules où chacune d'elles, est reliée seulement avec son voisinage. Toutes les cellules obéissent en parallèle à la même règle appelée fonction de transition locale, qui a comme conséquence une transformation globale du système. Deux composants coopèrent entre eux pour la construction du modèle booléen : le COG (Cellular

Optimization and Generation) qui s'occupe de la génération du graphe d'induction cellulaire et de son optimisation et le CIE (Cellular Inference Engine), un moteur d'inférence cellulaire, qui génère un ensemble de règles cellulaires utilisées pendant la phase de filtrage. Pour se faire, ils utilisent une base de connaissances sous forme de deux couches finies d'automates finis. La première couche, CelFact[1], pour la base des faits et, la deuxième couche, CelRule[2], pour la base de règles. Le voisinage des cellules est défini par deux matrices d'incidence d'entrée R_E et de sortie R_S. La dynamique de l'automate cellulaire, utilise deux fonctions de transitions $\delta fact$ qui simule les phases de sélection et de filtrage dans un système expert et $\delta rule$ qui correspond à la phase d'exécution :

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{fact}} (EF, IF, \boldsymbol{EF}, \boldsymbol{ER} + (\boldsymbol{R_E^T} \times \boldsymbol{EF}), IR, SR)$$

$$(EF, IF, SF, ER, IR, SR) \xrightarrow{\delta_{rule}} (\boldsymbol{EF} + (\boldsymbol{R_S} \times \boldsymbol{ER}), IF, SF, ER, IR, \overline{\boldsymbol{ER}})$$

2.2 Architecture

Nous présentons dans la figure1 l'architecture de notre système à base d'automate cellulaire que nous avons baptisé CASD («Cellular Automaton for Spam Detection).

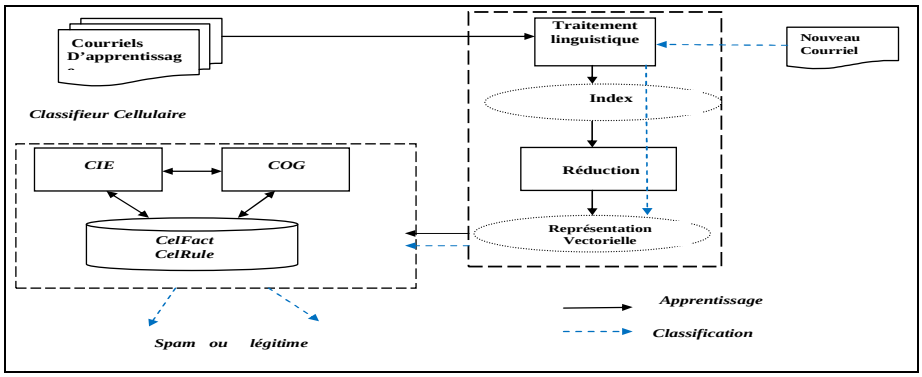

FIG.1– *Architecture de CASD*

2.3 Prétraitement Linguistique et Réduction

L'ensemble des courriels d'apprentissage doit être prétraité pour extraire les termes qui le représentent. A l'aide du module d'indexation que nous avons implémenté, nous pouvons établir une première liste de termes en procédant au découpage du texte en mots, à l'élimination des mots vides, comme nous pouvons aussi utiliser une variante de l'algorithme de Porter[3] pour la racinisation des différents mots retenus dans cette première phase. Cet ensemble de termes est par la suite réduit, par le choix de l'une des trois mesures implémentées dans CASD : l'information mutuelle (MI), le gain d'information (GI), et la statistique de Chi-2(χ^2) (Sebastiani, 2002). La sélection des termes est effectuée dans le but de choisir les éléments les plus

[1] Toute cellule de CelFact est considérée comme fait établi si sa valeur est 1, sinon, elle est considérée comme fait à établir. Elle se présente sous trois états : état d'entrée (EF), état interne (IF) et état de sortie (SF)

[2] Toute cellule de CelRule est considérée comme une règle candidate si sa valeur est 1, sinon, elle est considérée comme une règle qui ne doit pas participer à l'inférence. Elle se présente sous trois états : état d'entrée (ER), état interne (IR) et état de sortie (SR

[3] http://tartarus.org/~martin/PorterStemmer/

pertinents de tous les courriels. Le système calcule la mesure élue pour tous les termes, puis prend les K (seuil) premiers termes correspondant aux plus grandes valeurs calculées.

Comme dans la majorité des algorithmes de classification de textes, nous utilisons une représentation vectorielle (Salton et al., 1975) des courriels : le sac de mots. Ainsi chaque courriel est représenté par un vecteur d = {w_1, w_2, . . . ,$w_{|n|}$ }de $\Re^n$ où chaque coordonnée représente la présence ou l'absence (=0) du mot dans le courriel et n désigne le nombre de termes de l'index.

2.4 Apprentissage

Le processus d'apprentissage effectué par notre système est résumé dans six étapes :

1. transformation de la représentation vectorielle des courriels dans le format « arff » adopté par l'automate cellulaire,
2. production du graphe d'induction avec la méthode Sipina,
3. représentation cellulaire du graphe d'induction,
4. inférence en chainage avant : passer de la configuration G(t) vers la configuration G(t+1) en utilisant les deux fonctions de transition δfact, δrule,
5. répéter (4) jusqu'à stabilisation (i.e. G(t+1) = G(t))
6. sauvegarde du modèle booléen généré.

2.4.1 Un exemple Illustratif

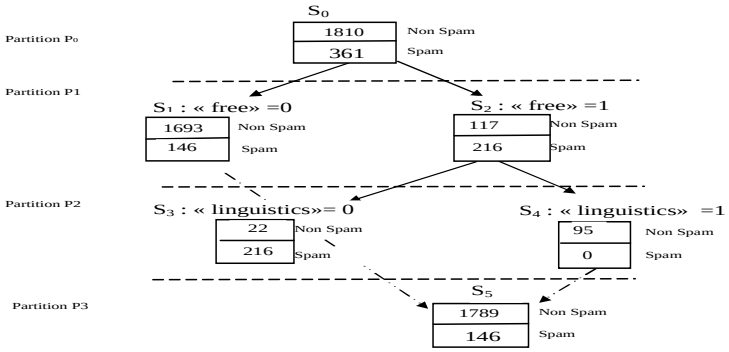

FIG. 2 – *Extrait d'un graphe d'induction : seulement les premières partitions sont présentées*

(a) Génération des règles cellulaires : un ensemble de règles est généré à partir du graphe d'induction comme illustré dans le tableau 1

Règle	Si	Prémisse	Alors	Conclusion
Règle 1	si	{S_0 }	alors	{«free » = 0, S_1}
Règle 2	si	{S_0}	alors	{«free» = 1, S_2}
Règle 3	si	{S_2}	alors	{«linguistics »=0, S_3}
Règle 4	si	{S_2}	alors	{«linguistics »=1, S_4}
Règle 5	si	{S_1, S_4}	alors	{S_5}

TAB. 1 –*Règles générées à partir du graphe de la figure 2*

(b) Représentation booléenne des règles générées : cet ensemble de règles est représenté par CelFact, CelRule, R_E et R_S dans l'automate cellulaire.

Fait n° i	
1	S_0
2	free=0
3	S_1
4	free =1
5	S_2
6	linguistics=0
7	S_3
8	linguistics=1
9	S_4
10	S_5

EF	IF	SF
1	0	0
0	1	0
0	0	0
0	1	0
0	0	0
0	1	0
0	0	0
0	1	0
0	0	0
0	0	0
CelFact		

Règle n°j	
1	R1
2	R2
3	R3
4	R4
5	R5

ER	IR	SR
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
CelRule		

TAB 2. – *Représentation cellulaire des partitions* $P_0 = \{s_0\}$, $P_1 = \{s_1, s_2\}$, $P_2 = \{s_3, s_4\}$, *et* $P_3 = \{s_5\}$.

R_E	R_1	R_2	R_3	R_4	R_5
S_0	1	1	0	0	0
«free »= 0	0	0	0	0	0
S_1	0	0	0	0	1
« free »=1	0	0	0	0	0
S_2	0	0	1	1	0
« linguistics»=0	0	0	0	0	0
S_3	0	0	0	0	0
« linguistics»=1	0	0	0	0	0
S_4	0	0	0	0	1
S_5	0	0	0	0	0

R_S	R_1	R_2	R_3	R_4	R_5
S_0	0	0	0	0	0
« free »= 0	1	0	0	0	0
S_1	1	0	0	0	0
« free»=1	0	1	0	0	0
S_2	0	1	0	0	0
« linguistics»=0	0	0	1	0	0
S_3	0	0	1	0	0
« linguistics»=1	0	0	0	1	0
S_4	0	0	0	1	0
S_5	0	0	0	0	1

TAB. 3 – *Les matrices d'incidence d'entrées/sorties pour la figure 2*

(c) Inférence : les Tableaux 2 et 3 représentent la configuration initiale G(0). Le chainage avant va permettre au modèle de passer de cette configuration aux configurations G(1), G(2)…G(i). L'inférence s'arrête après stabilisation avec une configuration finale. Le tableau 4 présente le modèle cellulaire final correspondant à la figure 2.

Fait		EF	IF	SF
1	free=0	0	1	0
2	free=1	0	1	0
3	linguistics=0	0	1	0
4	linguitics=1	0	1	0
5	S3 : class=spam	0	1	0
6	S5 : class=legitimate	0	1	0
CelFact				

Règle		ER	IR	SR
1	R1	0	1	0
2	R2	0	1	0
3	R3	0	1	0
CelRule				

R_E	R_1	R_2	R_3
«free » = 0	1	0	0
« free »=1	0	1	1
« linguistics»=0	0	1	0
« linguistics»=1	0	0	1
S_3 : class=spam	0	0	0
S_5 :class=legitimate	0	0	0

R_S	R_1	R_2	R_3
«free »= 0	0	0	0
« free »=1	0	0	0
« linguistics»=0	0	0	0
« linguistics»=1	0	0	0
S_3 :class=spam	0	1	0
S_5 :class=legitimate	1	0	1

TAB.4 – *Un extrait de la Configuration finale de l'automate*

Du modèle booléen représenté en tableau 4, des règles de classification sont déduites, par exemple la règle R1 se lit : *Si « free =0 » Alors légitime (classe majoritaire de S5).*

2.5 Classification

Cette étape utilise comme entrée le modèle élaboré depuis la phase d'apprentissage. Nous résumons les principales étapes comme suit :

1. Charger le modèle booléen : CelFact[4], CelRule, R_E, et R_S
2. Prétraiter le nouveau courriel et calculer sa représentation vectorielle : soit V.
3. Initialiser la base de faits CelFact :
   Pour chaque terme j dans CelFact faire
   Si terme $_j$ présent dans V **Alors** EF(terme$_j$ =1) ← 1
   Sinon EF(terme$_j$ =0)← 1 **Fin Si** **Fin Pour**
4. Appliquer la fonction de transition globale ∇=δfact°δrule
5. **Si** (EF (class=spam) ==1) **Alors** le courriel est classifié spam
   Sinon (EF (class= legitimate) =1) le courriel est classifié légitime **FinSi.**

3 Etude Expérimentale et Résultats

Afin d'évaluer cette approche de classification cellulaire pour le filtrage de spam, nous avons entamé plusieurs expériences sur le corpus Ling-Spam[5]. Et en nous appuyant sur la validation croisée, et en suivant les travaux effectués dans ce domaine (Androutsopoulos et al, 2000), nous mesurons le rappel de la classe spam (r), la précision de la classe spam (p), la F-mesure de la classe spam (f1) et enfin l'exactitude (e).

Soient N(LL) : le nombre de courriels légitimes classifiés légitimes; N(SS) : le nombre de courriels spam classifiés spam; N(LS) : le nombre de courriels légitimes classifiés spam et N(SL) le nombre de courriels spam classifiés légitimes, nous avons alors :

$$p = \frac{N(SS)}{N(SS)+N(LS)} \qquad r = \frac{N(SS)}{N(SS)+N(SL)} \qquad f1 = \frac{2*pr}{p+r} \qquad e = \frac{N(SS)+N(LL)}{N(SS)+N(LL)+N(SL)+N(LS)}$$

.

A travers ces expériences, nous avons constaté que la qualité de prédiction devient de plus en plus meilleure en termes de précision, rappel et exactitude à partir de 300 termes lorsqu'il y a racinisation des mots et élimination des mots vides avec les trois mesures de sélection. Nous avons constaté aussi que la mesure de sélection GI amène à une meilleure qualité de prédiction que les deux autres mesures. Enfin, nous avons constaté que l'approche proposée se stabilise à partir de 500 termes sélectionnés avec la fonction GI et amène à une qualité de prédiction intéressante : une précision = 98,1%, un rappel =84,2%, et une exactitude = 97.1%. Afin de comparer ces résultats avec les autres techniques, nous incluons les résultats des expériences réalisées sur le corpus LingSpam avec deux autres classifiers proposés dans la littérature :

NB : nous incluons les meilleurs résultats déclarés par (Androutsopoulos et al, 2000) pour l'approche bayésienne naïve. En utilisant une version lemmatisée du corpus LingSpam et l'information mutuelle (MI) comme métrique pour la sélection des termes, Ils trouvent que le classifier NB fonctionne de manière optimale avec un ensemble de termes égale à 100.

K-NN : à partir du même papier, nous incluons les meilleurs résultats déclarés pour une variante de l'algorithme du plus proche voisin. Comme dans le cas de NB, ils effectuent la sélection des termes en se basant sur la métrique MI, et obtiennent des résultats optimaux avec un plus petit nombre de termes (égale à 50) pour k=1 et K=2.

[4] Le EF de chaque cellule est initialisé à 0: aucun fait n'est établit
[5] Ling-Spam Corpus, http ://www.aueb.gr/users/ion/data/lingspam*public:tar:gz*

Classifier	Mesure de Sélection	Nbr Termes	Spam Précision	Spam Rappel	Spam F-mesure	Exacti-tude
NB	MI	100	99,02	82,35	89,92	96,926
TiMBL(1)	MI	50	95,92	85,27	90,28	96,890
TiMBN(2)	MI	50	97,10	83,19	89,61	96,753
CASD	**GI**	**500**	**98,10**	**84,20**	**90,62**	**97,100**

TAB. 5– *Résultats avec les meilleures configurations sur le corpus LingSpam*

Le tableau 5 présente les meilleurs résultats obtenus en utilisant notre classifieur CASD à côtés de ceux publiés précédemment et cité ci-dessus. Les résultats indiquent une amélioration des performances lorsque le classificateur CASD est utilisé. Il est clair qu'il surpasse NB et Knn en exactitude et en F-mesure.

4 Conclusion

Dans ce papier, nous avons proposé l'utilisation d'une nouvelle approche basée sur un automate cellulaire pour la détection de spam. Nos premières évaluations indiquent que l'approche proposée est intéressante. Bien que les résultats obtenus soient encourageants, beaucoup de points sont susceptibles d'être étudiés dans le cadre de travaux futurs. Nous devons approfondir nos expériences pour bien discerner les avantages et faiblesses de cette approche, cela nous permettra de mieux comprendre les situations où l'approche deviendra plus intéressante. Nous devons aussi mener une comparaison plus poussée de cette approche avec d'autres algorithmes d'apprentissage utilisés dans le filtrage de spam et en considérant d'autres corpus comme Spam Assassin et critères d'évaluation.

Références

Androutsopoulos, I., Koutsias, J (2000), "An Evaluation of Naive Bayesian Networks.", In: Machine Learning in the New Information Age. Barcelona Spain 9-17

Atmani B. et Beldjilali B. (2007). Knowledge Discovery in Database: Induction Graph and Cellular Automaton, Computing and Informatics Journal, 26, 171-197.

Guzella T. S., Caminhas W. M. (2009), "Review: A review of machine learning approaches to spam filtering", Expert Systems with Applications, 36(7), 10206-10222.

Salton G., Wong A., Yang C. S., (1975) "A vector space model for automatic indexing", Communications of the ACM, 18(11), 613-620

Sanz . E.P, Hidalgo J M G, Perez J C C, (2008), "Email spam filtering", in Zelkowitz M. (Ed.), Advances in computers, vol.74, 45-114.

Sebastiani F. (2002) Machine Learning in Automated Text Categorization, ACM Computing.

Zighed. (2000). Graphe d'induction: Apprentissage et data mining. HERMES.

Summary

Spam, also known as junk mail quickly became a major problem on the Internet. To address this growing burden of this type of spam, we propose the use of a supervised classification based on Boolean cellular automata to automatically classify incoming emails as spam or legitimate. To evaluate the performance of this new approach, we conduct a series of experiments on the corpus LingSpam.

Extraction et formalisation des connaissances pour l'aide au choix et aux réglages du Fauteuil Roulant Manuel

Véronique Delcroix[*,**], Karima Sedki[*,**]
François-Xavier Lepoutre[*,**] Anne-Pascale Maquinghen[*,**]

[*]Université Lille Nord de France, F-59000 Lille, France 2 UVHC
[**]LAMIH FRE CNRS 3304, F-59313 Valenciennes, France
{Veronique.Delcroix, Karima.Sedki, François-Xavier.Lepoutre,
Anne-Pascale.Maquinghen}@univ-valenciennes.fr

Résumé. Dans cet article, nous proposons un modèle fondé sur un réseau bayésien ayant une structure particulière pour représenter les connaissances relatives au problème du choix d'un Fauteuil Roulant Manuel (FRM). Un tel choix se fait en fonction de plusieurs critères et de nombreuses contraintes. Les connaissances sont relatives à la personne concernée (caractéristiques, usages, environnement, contraintes, etc.) et aux objets de choix (les FRM) qui sont décrits par plusieurs paramètres. La construction du modèle nécessite de regrouper de nombreuses connaissances d'experts, concernant les paramètres et la façon dont ils influent sur le choix. L'avantage du modèle est qu'il permet de représenter la plupart des paramètres des connaissances concernant le problème en prenant en compte un niveau d'incertitude. Une fois construit, le réseau bayésien peut-être utilisé d'une manière facile et efficace. Le principe de modélisation repose sur la définition de trois indices pour chaque critère : l'indice de qualité d'un FRM pour le critère, l'indice importance du critère pour la personne concernée par le choix et l'indice de satisfaction pour ce critère concernant le choix d'un FRM pour la personne. Le système proposé permet, d'une part, de proposer un FRM pour une personne, et, d'autre part, d'évaluer les niveaux des satisfaction procurés par le FRM choisi en fonction de chaque critère.

1 Introduction

Le Fauteuil Roulant Manuel - FRM - est une aide technique à la mobilité très connue et emblématique du handicap. C'est un dispositif mécanique de conception rustique, car il ne comporte pas de pièces compliquées. C'est cependant un objet complexe car il est composé de plusieurs centaines de pièces. Il comporte couramment des dizaines de réglages et d'options. Dans ce cadre, les professionnels médicaux doivent préconiser l'usage du FRM à des personnes souffrants de déficiences diverses, prescrire un type de FRM particulier, et aussi préciser les réglages et options, qui conviendront "au mieux" en fonction des besoins reconnus. Les distributeurs doivent ensuite trouver chez leurs fournisseurs le ou les FRM(s) qui correspond(ent) à ces prescriptions. Dans les faits, il semble que la moitié des FRMs utilisés soient peu ou mal

adaptés à la personne ce qui peut entraîner limitation d'autonomie, gêne, douleur, escarres, Troubles Musculo Squelettiques. En réponse à cette situation et dans le cadre de l'appel à projet de l'Agence Nationale de la Recherche sur les Technologies de la Santé (ANR-TECSAN 2006), un travail a été réalisé pour formaliser les connaissances et réaliser un système d'aide au choix et aux réglages du FRM (projet SACR-FRM). Les connaissances expertes dans ce domaine sont particulièrement des travaux scientifiques concernant des points de vue particuliers : biomécanique de la propulsion (Mulroy et al. (2004); Koontz et al. (2007); Guillaume et al. (2010); Yoshimasa et al. (2010)), biomécanique des transferts (Gagnon et al. (2009); Debril et al. (2009)). Il existe aussi des documents plus généraux, essentiellement anglo-saxons, sur les critères de qualité des FRMs (Guillon et al. (2009); Tomlinson (2000)) et des règles d'adaptation particulières (OMS; Axelson et al. (1994)). En France, le cours du CERAH [1] à destination des distributeurs (CERAH) est une bonne synthèse en 180 pages qui est réservé aux participants de leurs stages de formation. Un autre acteur français de référence dans ce domaine, est le service d'Aide au Choix du Fauteuil Roulant de la Fondation Garches [2] qui propose une approche individualisée fondée sur des essais. De plus, chaque ergothérapeute a son approche particulière fondée sur sa sensibilité, ses catégories de patients, son histoire. Nous avons tenté de regrouper et d'homogénéiser ces connaissances et pratiques françaises dont les principaux acteurs étaient participants du projet SACR-FRM. Une partie fondamentale du travail a consisté à identifier, homogénéiser et lister les caractéristiques de la personne, de son environnement et de son projet de vie, susceptibles d'orienter un choix. De la même manière, les caractéristiques (ou paramètres) du FRM (de l'ordre de 200) ont été recensées. Une troisième liste a été établie qui contient les critères de performance, de coût et de sécurité des FRM. Pour chacun de ces critères, des indices ont été construits : l'un relatif aux caractéristiques du FRM (appelé Indice Qualité), l'autre relatif à l'importance de ce critère pour la personne (appelé Importance). Un troisième indice de satisfaction (appelé Satisfaction) est déterminé.

Le formalisme des Réseaux Bayésiens (RB) (Jensen (1996); Naïm et al. (2007)) permet de modéliser les connaissances extraites et calculer ces indices en gérant l'incertitude au niveau des paramètres et de leurs liens. Ce formalisme a été retenu car il permet d'une part de visualiser sous forme de graphe les liens de causalité et d'influence et, d'autre part, de prendre en compte des relations déterministes et non déterministes entre les caractéristiques, à l'aide de distributions de probabilités conditionnelles associées aux variables du problème. Les réseaux Bayésiens permettent de prendre en compte l'incertitude concernant la valeur des variables et la façon dont une variable influence la valeur d'une autre variable. Ce type de modèle graphique probabiliste permet une modélisation fine de la connaissance des experts en termes de préconisation d'un FRM pour une personne donnée. Il intègre les contraintes et les critères du problème d'optimisation. Une fois construit, le réseau bayésien peut-être utilisé simplement et efficacement pour chaque nouveau cas de décision.

Le fonctionnement du réseau bayésien consiste à propager dans le réseau un ensemble d'informations connues ou imposées, vers un ensemble de variables cibles. Ce calcul des probabilités *a posteriori* (ou inférence) peut se faire pour n'importe quel ensemble de variables du réseau, aussi bien pour les observations que pour les variables d'intérêt. Dans le système d'aide à la décision qui nous intéresse, le réseau bayésien permet d'une part de proposer un

1. Centre d'Etude et de Recherche sur l'Appareillage pour le Handicap
2. Centre d'essai des fauteuils roulants. http ://www.handicap.org/ ?Centre-d-essai-des-fauteuils

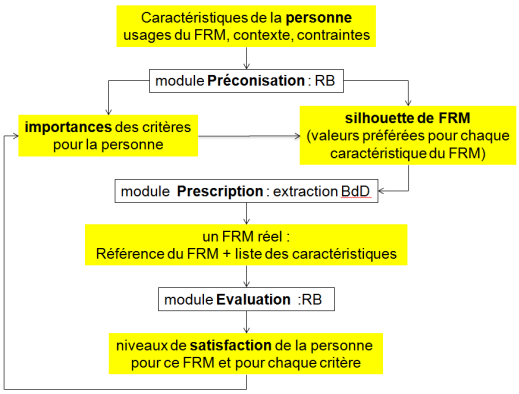

FIG. 1 – *Les étapes de l'utilisation du système d'aide au choix d'un FRM.*

ensemble de FRM les mieux adaptés à une personne, mais aussi d'évaluer un FRM sélectionné pour une personne, en calculant un indice de satisfaction pour chacun des critères de choix.

La démarche pour utiliser ce système (figure 1) consiste simplement à renseigner les caractéristiques de la personne, ses usages du FRM et tous les éléments de son contexte susceptibles d'influer sur son choix. Le premier module du système exploite le réseau bayésien pour obtenir d'une part les indices d'importances pour chaque critère pour la personne et d'autre part une préconisation de FRM qui se traduit par des ensembles de valeurs préférées pour chaque caractéristique du FRM. Le deuxième module se base sur ce résultat pour extraire un ou plusieurs FRM réels d'une base de données. Le dernier module utilise de nouveau le RB pour calculer les indices de satisfactions pour chaque critère concernant le FRM choisi pour la personne concernée.

Cet article a pour objet de présenter les différentes parties citées ci-dessus. La partie 2 concerne les listes de variables : les caractéristiques de la personne, du FRM et la liste des critères. Les partie 3 et 4 présentent les indices et le réseau bayésien, graphe et probabilités, puis la démarche complète est décrite partie 5 sur un exemple de fonctionnement. Une discussion termine l'article, à propos des avantages, inconvénients et limites des réseaux bayésiens pour traiter des problèmes tels que celui de l'aide au choix et au réglage d'un FRM.

2 Les paramètres du problème

Cette partie donne un aperçu significatif de l'ensemble des paramètres qui interviennent lors du choix d'un FRM. Ces listes ont été élaborées au cours du projet SACR-FRM avec l'aide des différents partenaires. Elles concernent d'une part la personne pour qui le choix est effectué, d'autre part le FRM, et enfin les différents critères de choix. Chacun de ces paramètres est représenté par un nœud dans le réseau bayésien qui constitue le modèle de raisonnement.

Les caractéristiques de la personne : Un premier groupe concerne les caractéristiques directes de la personne : âge, poids, taille et mensurations, amputations, capacités et incapacités

FIG. 2 – *Illustration de la diversité des FRMs aux niveaux de leurs structure, réglages et options.*

au niveau des membres supérieurs et inférieurs : capacité de marche, de propulsion, de prise manuelle, force des membres supérieurs, mais aussi évaluations des douleurs à différents niveaux et de la fatigabilité. Un second groupe de caractéristiques traduit le projet de vie de la personne dans ce FRM, son mode de vie et ses usages du FRM : niveau d'activité de la personne, temps passé par jour dans ce FRM, pourcentage d'utilisation du FRM en intérieur, périmètre de déplacement, présence d'un aidant (fréquence, force), caractéristiques du logement et des lieux d'usage (passages difficiles, changement de niveau, sols difficiles, ...), déplacements en voiture (fréquence, type de voiture et de chargement du FRM, aidant). Il est à noter que les différents experts ne sont pas accordés sur une liste unique de paramètres pertinents. La liste que nous avons élaborée se veut une synthèse des paramètres utiles pour le choix du FRM.

Les caractéristiques du FRM : La figure 2 illustre la diversité des FRMs disponibles sur le marché, au niveau de la structure, des réglages et des accessoires. La liste ci-dessous donne un aperçu de ces éléments et de la complexité de l'ensemble.

Le FRM peut-être propulsé par un accompagnateur ou par la personne elle-même, par propulsion podale ou manuelle. Des options des FRM permettent un ou plusieurs de ces modes de propulsion, comme la présence et le type des mains courantes, levier pendulaire, poignées, systèmes de freinage, etc. Le châssis peut être pliant ou rigide, parfois conçu spécifiquement pour une activité de sport. Il supporte un poids maximum de l'utilisateur, il peut inclure un système de réglage de l'inclinaison de l'assise, un dispositif anti-bascule, un levier de basculement, etc. Le dossier et le siège sont caractérisés notamment par le type de structure (rigide, plaque avec coussins, toile), les dimensions, les possibilités de réglages et d'inclinaisons de divers éléments. Il existe différents accessoires pour le maintien et le positionnement d'une partie du corps et aussi un dispositif permettant de passer de la position assise à la position debout (verticalisateur). Les accoudoirs présentent différentes formes, dimensions et possibilités de réglages. Les ensembles repose-pieds (potences et palettes) peuvent être fixes, escamotables ou amovibles, en une ou deux parties, inclinables, réglables suivant différents axes. Divers accessoires améliorent le maintien de la jambe et du pied. Les deux roues motrices et les deux roues directrices varient par leurs types, leurs dimensions, leur position, les types de revêtement, la possibilité de démontage et les systèmes associés, les accessoires pour la propulsion

et la direction. Ces caractéristiques sont normalisées et listées dans un document officiel de nomenclature, établi par la CNSA [3] et le CERAH, pour déterminer les conditions de prise en charge financière. Ce document est cependant encore en évolution.

Les critères de choix et les contraintes : Selon le Dictionnaire de la Langue Française un critère est un "Principe, élément considéré pour évaluer, analyser, juger quelque chose". Un critère conduit naturellement à une optimisation (le plus performant, le plus économique, le plus sûr, etc.). Les critères sont souvent contradictoires ce qui impose de réaliser des compromis. Les critères que nous avons retenus sont les suivants : confort de repos, confort de propulsion, confort de franchissement d'obstacle, confort de transfert frontal et latéral, confort de chargement en voiture, confort de l'aidant, manœuvrabilité, stabilité arrière, avant et latérale, résistance au roulement, solidité.

Une contrainte, par définition, doit être respectée. Elle décrit des relations directes entre des caractéristiques de la personne et celles du FRM. Elles conduisent à éliminer définitivement les FRM qui ne répondent pas à ces relations. Par exemple, il est indispensable que la largeur du siège soit supérieure ou égale à la largeur du bassin de la personne.

3 Les liens entre les paramètres : indices de qualité, d'importance et de satisfaction des critères

Les paramètres du problème étant listés, la suite du travail d'extraction des connaissances a consisté à définir trois indices pour chaque critère, chacun représenté par un nœud dans le RB. Un indice de qualité qui représente et quantifie la qualité du FRM selon ce critère. Cet indice est une fonction logico-mathématique de caractéristiques essentiellement du FRM, mais aussi, éventuellement, de la personne, de son activité et de son environnement. Un indice d'importance qui représente et quantifie l'importance de ce critère pour la personne. Cet indice est une autre fonction de caractéristiques de la personne. Un troisième indice dit de satisfaction permet d'effectuer une optimisation multicritères par le réseau Bayésien (cf. partie 4), fondée sur l'hypothèse que le "meilleur" fauteuil roulant sera celui dont les caractéristiques généreront les indices de qualité les plus élevés pour les critères dont les indices d'importances sont aussi les plus élevés. Ces trois indices sont représentés par une valeur dans l'intervalle $[0, 1]$; plus l'indice de qualité (resp. d'importance, de satisfaction) est élevé, plus le niveau de qualité du FRM (resp. d'importance, de satisfaction) est élevé. Nous illustrons cela sur le critère de manœuvrabilité, qui correspond à la qualité d'un FRM qui peut être propulsé ou poussé aisément dans un environnement encombré (couloir, ascenseur, parking, terrain de sport, ...).

Indice de qualité de la manœuvrabilité du FRM : L'augmentation de cette manœuvrabilité est fonction des : **(i)** déport vers l'avant des roues arrière par rapport au châssis, parce qu'elle diminue le poids sur les roues avant et diminue l'inertie de rotation autour d'un axe vertical, **(ii)** diminution du diamètre des roues avant qui réduit les forces nécessaires pour faire pivoter le FRM, **(iii)** diminution de la longueur hors tout du FRM qui diminue l'inertie de rotation autour d'un axe vertical et **(iv)** diminution de la hauteur du dossier qui réduit les gênes au niveau des

3. Caisse Nationale de Solidarité pour l'Autonomie : http ://www.cnsa.fr/

bras et avant-bras lors de la propulsion. L'influence de ces paramètres sur l'indice de qualité du FRM est illustré sur le graphe de la figure 3 (partie droite). Certains paramètres comme la longueur du FRM dépendent d'autres paramètres et de leur réglages, comme l'angle de la potence. Cet indice de qualité est une somme de termes relatifs à chaque facteur. En effet, ces facteurs sont indépendants et aucun d'eux n'a de valeur qui rendrait la manœuvrabilité nulle quel que soit les autres facteurs.

Indice d'importance de la manœuvrabilité du FRM : L'importance de cette manœuvrabilité pour une personne est fonction des **(i)** Poids de la personne : plus la personne est lourde plus un FRM manœuvrable est souhaitable. **(ii)** Niveau d'activité : plus la personne est active plus un FRM manœuvrable est souhaitable. A contrario, une personne sans aucune activité a peu besoin d'un FRM manœuvrable. **(iii)** Difficultés de circulation dans l'environnement ordinaire : ascenseur, porte, couloir, toilette, **(iv)** Capacités de marche résiduelle et de transfert assis debout qui permettent de quitter le FRM dans les passages difficiles et diminue ce besoin de manœuvrabilité. La partie gauche de la figure 3 illustre ces dépendances. La formule mathématique sera aussi une somme pondérée de ces paramètres. (Dans d'autres cas, par exemple le critère de confort de repos, les formules pourront être multiplicatives pour refléter le fait qu'un paramètre "mauvais" peut annuler le confort). L'indice de satisfaction relatif à ce critère est déterminé par inférence ce qui est expliqué dans les paragraphes suivants.

4 Modélisation des connaissances par un réseau bayésien

Dans les deux parties précédentes, nous avons listé les variables, puis explicité une partie des liens d'influence entre ces variables. Nous présentons maintenant la structure générale du modèle pour expliquer comment l'ensemble des connaissances des experts est représenté dans le réseau et comment il est ensuite utilisé. Nous avons utilisé le logiciel Netica (Norsys (1998)) pour définir et utiliser le RB.

Structure du réseau bayésien : La figure 4 donne la structure générale du graphe du réseau bayésien utilisé pour l'aide au choix d'un FRM. Chaque zone regroupe un ensemble de nœuds du graphe, correspondant chacun à un paramètre du problème décrit dans la section 2. Une flèche entre deux zones signifie qu'un nœud (une variable) d'une zone peut avoir une influence directe sur un nœud de l'autre zone. Il existe aussi des flèches entre les nœuds d'une même zone, en particulier pour les caractéristiques de la personne et celles du FRM (voir figure 3).

A chaque critère est associé un nœud "indice importance" et un nœud "indice qualité". Les caractéristiques qui influent sur la définition de ces variables sont des nœuds parents des nœuds indice ou importance. Ainsi la valeur du nœud fils (indice de qualité ou d'importance) dépend de la valeur des nœuds parents.

La zone satisfaction contient les indices de satisfaction pour chaque critère. La satisfaction d'un critère représente le niveau de satisfaction de la personne à propos d'un FRM choisi sur ce critère. Le niveau de satisfaction dépend de l'indice d'importance du critère et de l'indice de qualité du FRM. Plus le critère est jugé important, plus il faut que l'indice de qualité du FRM soit élevé pour obtenir une *bonne* satisfaction. La figure 3 montre le graphe complet du RB pour le critère de manœuvrabilité.

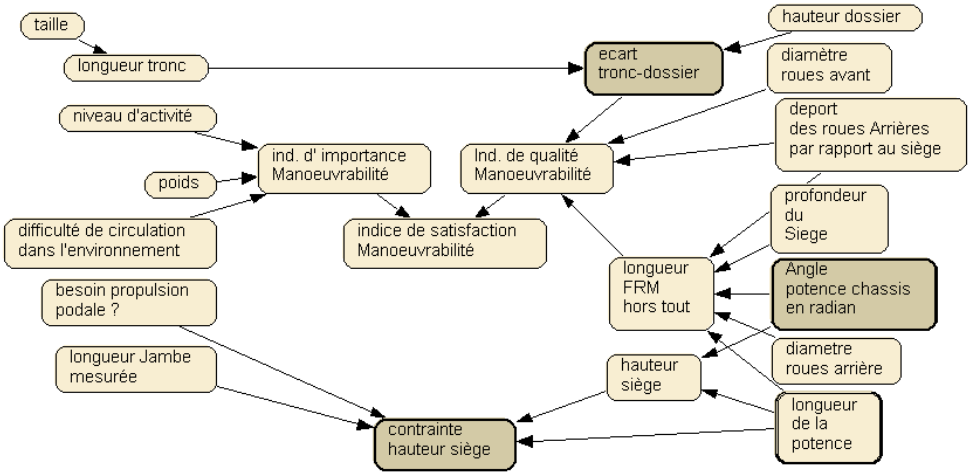

FIG. 3 – *Graphe du Réseau Bayésien pour le critère de manœuvrabilité (les nœuds fonçés sont déterministes).*

Le graphe du réseau bayésien pour tous les critères est obtenu par assemblage des réseaux bayésiens associés à chaque critère. Différents critères peuvent dépendre de caractéristiques communes. Cet exemple illustre aussi la gestion d'une contrainte sur la hauteur de siège et la longueur des potences. Les contraintes sont traduites dans le réseau bayésien suivant le même principe que les critères, en ajoutant un nœud spécifique, dont les parents sont les paramètres de la contrainte. Le détail de cette contrainte est donné plus bas.

Une fois défini la structure du graphe du réseau bayésien (les paramètres et les liens entre eux), il importe de définir les probabilités sur chaque nœud. Ces deux composantes permettent ensuite de réaliser l'inférence dans le RB. Ce principe consiste à mettre à jour les probabilités des variables d'intérêt et constitue la base du raisonnement.

Définition des probabilités et inférence dans le RB : La construction complète du réseau bayésien nécessite de définir une distribution de probabilités pour chaque nœud, sur son ensemble de définition. Deux cas se présentent suivant que le nœud considéré a des parents ou non. Un nœud sans parent est assorti d'une distribution de probabilité *a priori* qui représente l'état de la connaissance sur cette variable sans autre information. Lorsqu'un nœud a un (des) nœuds parents, sa valeur dépend de la valeur des nœuds parents. Par exemple la *longueur du tronc* dépend de la *taille de la personne*. On définit les probabilités sur ces deux nœuds en décrivant les distributions de probabilités $P(taille)$ $P(longueurTronc|taille)$ (cf. figure 5). Sur la partie droite de cette figure, on a entré dans le RB la taille d'une personne. Les distributions de probabilités sur les nœuds indice d'importance et de qualité sont obtenus à l'aide des formules qui traduisent la façon dont ces indices dépendent des nœuds parents. Les probabilités sur les nœuds indice de satisfaction sont déduit de la formule : $indSatisfaction = indQualite^{indImportance}$.

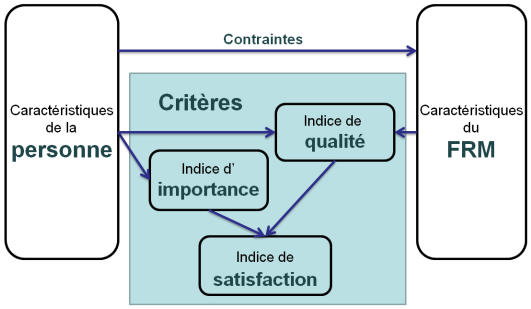

FIG. 4 – *Structure générale du graphe du Réseau Bayésien pour l'aide au choix d'un FRM.*

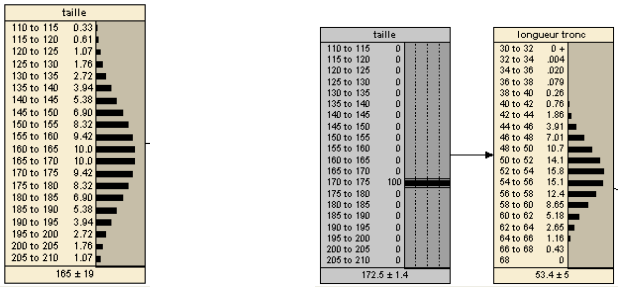

FIG. 5 – *Distributions de probabilités* a priori $P(taille)$ *et* $P(longueurTronc|taille)$.

L'inférence dans le réseau bayésien permet de calculer l'influence d'un ensemble d'informations connues noté Obs sur la valeur d'un nœud cible N. Il est possible de calculer la probabilité *a posteriori* $P(N|Obs)$ quels que soient les nœuds N et Obs du réseau. En particulier, la première inférence utilisée dans notre système est faite en fixant les valeurs des caractéristiques de la personne considérée, et en imposant les valeurs des nœuds satisfaction et contraintes. L'inférence dans le RB permet alors de calculer l'importance des critères pour cette personne et les probabilités *a posteriori* pour chaque caractéristique du FRM. Ce mécanisme est illustré dans la suite sur un cas précis pour la gestion d'une contrainte et d'un critère.

Gestion des contraintes : Les contraintes sont des relations entre des caractéristiques de la personnes et celle du FRM qui doivent impérativement être respectées. A la différence des critères, on ne leur attribue pas un degré d'importance. Si un FRM ne vérifie pas une contrainte, il est simplement rejeté. La gestion des contraintes repose sur un modèle simplifié de gestion des critères. L'exemple de la figure 6 illustre la gestion d'une contrainte sur la hauteur du siège du FRM et des potences : si une personne a besoin de la propulsion podale, alors la hauteur du siège doit correspondre exactement à la hauteur de la jambe (du talon au dessous du genou), sinon, c'est la potence qui doit être de même longueur que la jambe. Sur la figure 6 les nœuds en gris sont ceux dont on connaît ou on impose la valeur. La valeur du nœud contrainte est fixée

de sorte que la contrainte soit respectée : seules les valeurs de hauteur de siège qui respectent la contrainte ont une probabilité non nulle.

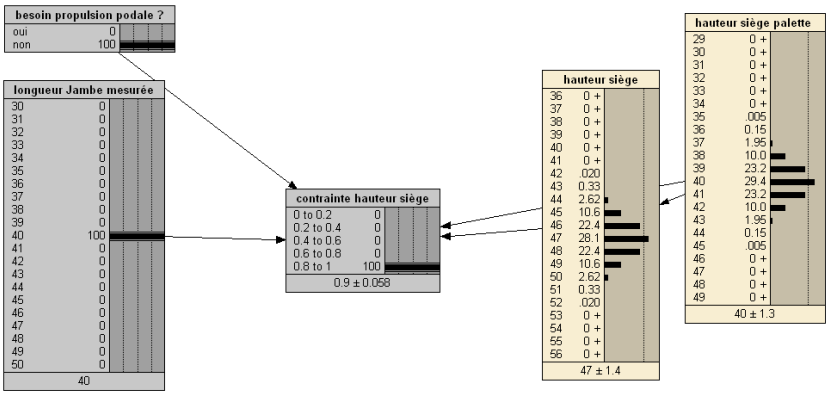

FIG. 6 – *Extrait du réseau bayésien pour gérer une contrainte.*

5 Processus d'utilisation du Système Interactif d'Aide à la Décision

Le réseau bayésien est un modèle de raisonnement qui permet de répondre à plusieurs objectifs selon les nœuds qui sont définis comme cibles, c'est-à-dire ceux pour lesquels on souhaite mettre à jour les probabilités en fonction des valeurs fixées. Afin de sélectionner un FRM adéquat pour une personne donnée, nous suivons le processus de décision décrit sur le schéma de la figure 1. Pour illustrer ce processus nous présentons un cas d'utilisation simplifié sur les différentes étapes. La personne considérée est un jeune homme actif, qui se déplace fréquemment avec son FRM. Il vit dans un appartement où la circulation est difficile. Voici les étapes du processus de choix :

Étape 1 : Introduction dans le système les caractéristiques de la personne. Dans le réseau bayésien, cela correspond à des observations. Si certaines informations ne sont pas connues, la distribution de probabilité *a priori* sur le nœud sera exploitée. Dans notre exemple, ces caractéristiques sont : *niveau d'activité* très élevé, le *poids* entre 80 et 90 kg et une assez grande *difficulté de circulation dans l'environnement*. Ces informations sont visibles sur la figure 7 (la partie gauche du modèle).

Étape 2a : Calcul des indices d'importance. Appliquer la première inférence dans le réseau bayésien et récupérer les indices d'importance pour chaque critère. Pour notre cas d'illustration, il apparaît que le critère de la manœuvrabilité est très important pour la personne concernée. Elle a donc besoin d'un FRM très manoeuvrable (c'est à dire ayant un indice de qualité élevé pour ce critère.

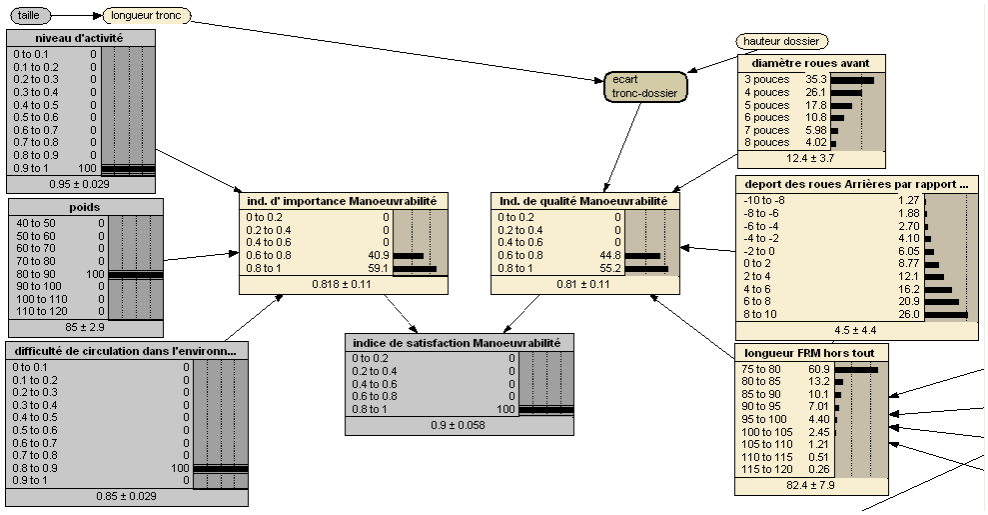

FIG. 7 – *Extrait du réseau bayésien pour le critère manœuvrabilité.*

Étape 2b : Calcul de la silhouette de FRM. Cette étape exploite un deuxième résultat de l'inférence précédente. Il faut auparavant fixer les valeurs des indices de satisfaction à la valeur la plus élevée (probabilité 1). Cette étape permet d'associer à chaque caractéristique du FRM une distribution de probabilités sur l'ensemble de ses valeurs. Les valeurs dont la probabilité est plus élevée sont les plus compatibles avec une valeur élevée de l'indice de satisfaction. Dans notre exemple, il est recommandé pour cette personne de choisir un FRM avec une *longueur hors tout* entre 75 et 80 cm, *diamètre roue avant* entre 3 et 5 pouces et un *déport des roues arrières* entre 6 et 10 pouces. Ce résultat est observé sur la figure 7 (partie droite du modèle).

Étape 3 : Sélection d'un FRM réel de la base de données. Nous avons utilisé la base de données du CERAH, contenant des références de FRM et les valeurs de caractéristiques disponibles pour chaque référence. Cette étape se fait en deux temps : sélection des FRM compatibles avec la silhouette de choix, c'est à dire dont toutes les caractéristiques présentent une probabilité supérieure à une seuil (bas) dans la silhouette de FRM. Puis calcul du FRM "le plus semblable" à la silhouette.

Étape 4 : Calcul des indices de satisfaction du FRM sélectionné pour chaque critère. Cela nécessite de rentrer la valeur de chaque caractéristique du FRM comme observation dans le réseau bayésien. Cette étape permet à la personne concernée par le choix d'obtenir la qualité du FRM vis-à-vis de chaque critère. Cela permet aussi de comparer deux FRM satisfaisants mais ne proposant pas le même compromis entre les différents critères.

6 Discussion et Conclusion

Le système proposé pour l'aide au choix d'un Fauteuil Roulant Manuel repose sur une modélisation fine des connaissances à l'aide d'un réseau bayésien. Quels sont les avantages, inconvénients et limites des réseaux bayésiens pour traiter des problèmes d'aide au choix multicritères tels que celui du choix d'un FRM ? La structure particulière du réseau proposé exige du modélisateur qu'il oriente l'extraction des connaissances suivant différents axes : une fois listés les paramètres susceptibles d'influer sur le choix du FRM, un effort d'abstraction doit permettre de définir les critères qui vont orienter le choix. La suite du travail de construction du modèle exige de transformer l'expression première du savoir des experts (dans notre cas, ceux qui savent quel type de FRM choisir suivant la personne et son besoin) de façon à décliner l'influence des paramètres suivant les différents critères, soit en terme d'importance de ce critère pour la personne, soit en terme de qualité du FRM suivant un critère. Un avantage majeur de cette proposition réside dans le fait de mettre à disposition de l'utilisateur un ensemble important de connaissances expertes regroupées dans le réseau bayésien, tout en préservant une utilisation simple du système. En contrepartie, la tâche de construction du réseau bayésien est conséquente, aussi bien en termes de structure du graphe, qu'en termes d'agrégation des paramètres pour définir les indices d'importance et de qualité des critères. Un autre avantage du système proposé est la possibilité d'évaluer finement l'adéquation d'un FRM pour une personne. Au cours de ce travail, nous avons utilisé une base de données de FRM du CERAH. Cette base étant depuis plusieurs années en pleine restructuration suite à un important travail de nomenclature des fauteuils roulants, nous avons du adapter l'ancienne base pour l'exploiter. Enfin la taille du modèle informatique notamment du au fait de la forte connexité du graphe devient un facteur pénalisant en terme de temps de calcul. Ce constat (prévisible) nous conduit à envisager de distribuer le réseau bayésien dans un système multi-agent. Cette perspective nécessite de proposer un mécanisme d'inférence dans le RB distribué, ce qui n'est pas sans difficulté vu la structure du graphe du système multi-agents.

Remerciements

Nous remercions pour le soutien qu'ils ont apporté à la réalisation de ce travail dans le cadre du projet SACR-FRM, l'ANR-TECSAN, l'IFR25 Handicap, le campus international CISIT, la Délégation Régionale à la Recherche et à la Technologie, le Ministère de l'Éducation Nationale, de la Recherche et de la Technologie, la Région Nord-Pas de Calais et le Centre National de la Recherche Scientifique.

Références

Axelson, P., J. Minkel, MAPT, et D. Chesney (1994). A guide to wheelchair selection. In *the Paralyzed Veterans of America*.

CERAH. Les véhicules pour personne handicapé. Document de formation du cerah. 57147 woippy.

Debril, J.-F., P. Pudlo, P. Gorce, et F.-X. Lepoutre (2009). Experimental platform to assess the weight relief lifting and sitting pivot transfer movements. In *7th IFAC Symposium on Modelling and Control in Biomedical Systems*, Aalborg, Denmark, août.

Gagnon, D., A. Koontz, S. J. Muiroy, D. A. Nawoczenski, E. Butier-Forsiund, A. Granstrom, S. Nadeau, et B. M. L. (2009). Biomechanics of sitting pivot transfers among individuals with a spinal cord lnjury : A review of the current knowledge. In *Spinal Cord lnj Rehabil*, Volume 15(2), pp. 33–58.

Guillaume, D., R. Dumas, D. Pradon, P. Vaslin, F.-X. Lepoutre, et L. Chèze (2010). Upper limb joint dynamics during manual wheelchair propulsion. *J. Clin. Biomech., doi:10.1016/j.clinbiomech.2009.12.011*.

Guillon, B., S. Bouche, B. Bernuz, et D. Pradon ([26-170-B-10]. 2009). Fauteuils roulants : description, utilisation, critères de choix. Em consulte, kinésithérapie-médecine physique-réadaptation.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. UCL Press.

Koontz, A., Y. Yang, R. Price, M. Tolerico, C. Digiovine, S. Sisto, R. Cooper, et M. Boninger (2007). Multisite comparison of wheelchair propulsion kinetics in persons with paraplegia. *Journal of Rehabilitation Research & Development 44*, 25–34.

Mulroy, S., S. Farrokhi, C. Newsam, et J. Perry (2004). Effects of spinal cord injury level on the activity of shoulder muscles during wheelchair propulsion : an electromyographic study. In *Archives of physical medicine and rehabilitation*, Volume 85, pp. 925–34.

Naïm, P., P.-H. Wuillemin, P. Leray, O. Pourret, et A. Becker (2007). *Réseaux bayésiens* (3 ed.).

Norsys (1998). Netica application, norsys software corp., http://www.norsys.com.

OMS. Guide pour les services des fauteuils roulants manuels dans les régions à faible revenus. In *Bibliothèque de l'OMS. ISBN 978 92 4 154748 2.140p*.

Tomlinson, J. (2000). Managing maneuverability and rear stability of adjustable manual wheelchairs: an update. In *Phys Ther*, Volume 80, pp. 904–911.

Yoshimasa, S. J., E. Watelain, F.-X. Lepoutre, A. Thevenon, et P. MD (2010). Effects of wheelchair mass on the physiologic responses, perception of exertion, and performance during various simulated daily tasks. In *Arch Phys Med Rehabil*, Volume 91, pp. 1248–1254.

Summary

In this paper, we propose a model based on a particular structure of Bayesian network that can represent knowledge about a problem of choice of a Manual Wheelchair (MWC). The choice is guided by several criteria and several constraints. The knowledge is related to the user (his/her characteristics, needs, use of the MWC, etc.) and to the alternatives (the MWC) that are described by several attributes. Building the model requires knowledge of experts about the parameters of the choice and the way they influence the choice. The advantage of the model is that it allows modeling all relevant knowledge of the problem. Once the bayesian network is built, it can be used in an easy and effective way. The reasoning is based on three indices for each criterion: the index of quality of a MWC for the criterion, the index of importance of that

criterion regarding the user and index of satisfaction concerning the choice of a MWC for the person. The framework that we propose helps to select an adequate MWC for a given person, and then to evaluate it for each criterion.

Automatic methods for mapping Biomedical terminologies in a Health Multi-Terminology Portal

Tayeb Merabti*, Julien Grosjean *
Hocine Abdoune** Michel Joubert** Stefan Darmoni*

*CISMeF, Rouen University Hospital, Normandy &
TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France
nom.prenom@chu-rouen.fr,
http://www.cismef.org
**LERTIM EA 3283, Marseilles, Faculty of Medicine, Marseilles, France
nom.prenom@ap-hm.fr
http://cybertim.timone.univ-mrs.fr

Abstract. Terminology mapping is an important and crucial task to improve semantic interoperability between health care applications and resources. In 2009, CISMeF created a Health Multi-Terminological Portal (HMTP) to search concepts among all the health terminologies available in French (or in English and translated in French) included in this portal and to browse it dynamically. To map terminologies in the HMTP, two methods are used: (1) conceptual method which exploits various features of the UMLS, (2) lexical method based on natural language processing in French and English. A total of 199,786 mappings were performed between at least two French terms using conceptual method, whereas 266,139 mappings were performed using lexical methods. These mappings were all integrated in the HMTP developed by CISMeF. Conceptual and lexical methods were used to translate some English terminologies into French such as MEDLINEPlus, FMA and SNOMED CT.

1 Introduction

Biomedical terminologies and ontologies have proliferated during the past decade. Due to this proliferation, different health care systems use different biomedical terminologies. In this context, tools and methods to map biomedical terminologies are needed to solve data interoperability problems. The process of terminology mapping consists of identifying relationships or identical (or approximately identical) concepts between terminologies Wang et al. (2008). A number of algorithms and approaches have been proposed to create an automatic mapping between health terminologies. For example, Rocha et al. (1994), Cimino and Barnett (1990) both proposed a frame-based approach to perform mappings between health terminologies. Other approaches were proposed using UMLS (Unified Medical Languages Systems) Lindberg et al. (1993) as a knowledge resource to perform mappings between terminologies. For example, Fung and Bodenreider (2005), Bodenreider et al. (1998) described an algorithm

to map between any two terminologies in the UMLS making use of synonymy, explicit mapping relations and hierarchical relationships. However, approaches using UMLS are limited to the biomedical terminologies already incorporated into UMLS. In Merabti et al. (2010b) we proposed a lexical algorithm to map a French terminology not included in UMLS to UMLS. The objective of this paper is to describe two types of approaches to map biomedical terminologies in English and French whether or not included into UMLS Lindberg et al. (1993). These two approaches are currently implemented into a Health Multi-Terminological Portal (HMTP) Darmoni et al. (2010) developed by the CISMeF team Darmoni et al. (2000).

2 Material

2.1 Unified Medical Language Systems

UMLS Lindberg et al. (1993) integrates over 2 million concepts (2,200,159 in the 2010AA version) from 148 biomedical vocabularies. The UMLS is made up of three main knowledge components, but, for our purpose, we retained only the Metathesaurus: a very large, multipurpose, and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships between them. Each concept has a unique identifier in the Metathesaurus (Concept Unique Identifier, CUI). Currently, a number of 181,168 French terms (47,719 preferred terms) are integrated in the UMLS.

2.2 CISMeF BackOffice & HMTP

The CISMeF BackOffice Darmoni et al. (2010) is a multi-terminological server developed by CISMeF to integrate and manage multiple terminologies. The HMTP[1] is a "Terminological Portal" connected to the CISMeF BackOffice to search concepts among all the health terminologies available in French (or in English and translated in French)[2] included in this portal and to browse it dynamically. A number of 27 terminologies and classifications were included in the CISMeF BackOffice, and therefore in HMTP. Some terminologies and classifications are included in the UMLS meta-thesaurus (n=9) but the majority are not (n=18). Table 1 displays the number of descriptors and relationships included in the HMTP.

Terminologies	27
Concepts	$\succ$ 867,791
Synonyms	$\succ$ 1,837,761
Definitions	223,654
Relations and hierarchies	2,990,365

TAB. 1 – *Main figures of the Health Multi-Terminology Portal.*

[1] `http://pts.chu-rouen.fr/pts_site`
[2] Bilingual graphical user interface (French and English) was developed.

3 Methods

Two automatic mapping approaches are implemented in the HMTP: conceptual and lexical approach. The conceptual approach uses the UMLS metathesaurus to map the terminologies included into the UMLS. The lexical approaches use some natural language processing tools to map terminologies whether or not included into the UMLS.

3.1 Conceptual Approach

This approach implies that each term to be mapped must be included into the Metathesaurus Joubert et al. (2009). The principle of the method is based on the conceptual construction of the UMLS metathesaurus. Three types of mapping are provided using this method: "Exact Mapping", "Broader Mapping" and/or "Narrow Mapping" and "Close Mapping" (see table 2 for some examples), this mapping method is inspired by SKOS (Simple Knowledge Organization System) definitions of mapping properties[3], SKOS language is also used to represent French health terminologies into the French Health Multi-terminological Server Darmoni et al. (2009). The mapping approach is as follows: suppose two terms t1 and t2 of two terminologies T1 and T2, respectively, suppose CUI1 and CUI2, the respective projections of t1 and t2 in the Metathesaurus, then t1 and t2 are mapped if :

- CUI1=CUI2, this corresponds to the "Exact Mapping",
- There is parent of t1 or t2 which maps t2 or t1 respectively, this corresponds to "Broad Mapping" and/or "Narrow Mapping": these are used to state mapping links through hierarchies,
- There is an explicit mapping between CUI1 and CUI2, this corresponds to the non-transitive "Close Mapping": two concepts are sufficiently similar that they can be used interchangeably.

The algorithm is carried out sequentially and stops if a candidate mapping is found. As an application of this, even if an explicit mapping comes from other terminologies in UMLS, e.g. ICD-9-CM and SNOMED CT Imel (2002) not part of terminologies under consideration, explicit mappings between two terminologies can be "reused" for other terminologies by means of the UMLS concept structure Fung and Bodenreider (2005).

3.2 Lexical Approach

In this approach, some Natural Language Processing (NLP) tools (English and French) are used to link terms from different sources. Lexical approach allows to find a term in the target terminology that is the most lexically similar to a given term in a source terminology. Two lexical algorithms were used in French and English to map all HMTP terminologies.

[3]World Wide Web Consortium Simple Knowledge Organization System: `www.w3.org/2004/02/skos`

Type of relation	**Source term** (Terminology)	**Target Term** (Terminology)
Exact Mapping	Congenital bladder anomaly (MedDRA)	Congenital anomaly of the bladder, nos (SNMI)
Close Mapping	Diseases of lips (ICD10)	Ulcer of lip (SNMI)
BT-NT Mapping	Hepatic insufficiency (MeSH)	Liver disease, nos (ICPC2)

TAB. 2 – *Examples for each type of conceptual mapping.*

3.2.1 French based approach

This approach uses a French NLP tools and mapping algorithms developed by the CISMeF team to map French health terminologies Merabti (2010); Merabti et al. (2010a,b). These tools were used in a previous work and extended to link terms in multiple French Health terminologies:

- Remove stop words: frequent short words that do not affect the phrases such “a”, “Nos”, “of”, etc are removed from all terms in all terminologies.
- Stemming, a French stemmer provided by the “lucene” software library which proved to be the most effective for the F-MTI automatic indexing tools using several health terminologies Pereira (2007), as compared to the stemming tools developed by the CISMeF team.

Mapping used by this approach may provide three types of correspondences between all terms:

- Exact correspondence: if all words composing the two terms are exactly the same.
- Single to multiple correspondence: when the source term cannot be mapped by one exactly target term, but can be expressed by a combination of two or more terms.
- Partial correspondence: In this type of mapping only a part of the source term will be mapped to one or more target terms.

Examples for each type of mapping are given in Table 3. In this work, only the exact correspondence was described.

3.2.2 English based approach

In this approach we use some NLP tools in English developed by the NLM Browne et al. (2003). These NLP tools are a series of tools designed to aid users in analyzing and indexing natural language texts in the medical field McCray et al. (1994); Peters et al. (2010). They include essentially some tools like:

- LVG (Lexical Variant Generation): a Multi-function tool for lexical variation processing;

Type of correspondance	**Source term** (Terminology)	**Target Term(s)**(Terminology)
Exact	Syndrome de Marfan “Marfan Syndrome”(MeSH)	Syndrome de Marfan “Marfan’s Syndrome” (MedDRA)
Single to Multiple	Albinisme surdité “Albinism- deafness syndrome” (ORPHANET)	Albinisme “Albinism” (MeSH) and (+) Surdité “Deafness” (SNMI)
Partial	Chromosome 14 en anneau “Ring chromosome 14” (ORPHANET)	Chromosome humain 14 “Chromosome 14” (MeSH)

TAB. 3 – *Examples of the three types of mappings using the French lexical approach.*

- Norm[4]: a program used to normalize English terminologies (UMLS terminology);
- WordInd: a tool used to tokenized terms into word.

In this work we basically used a normalization program (“Norm”). The Normalization process involves stripping genitive marks, transforming plural forms into singular, replacing punctuation, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words into alphabetic order. In the English base approach, only the exact correspondence was used. This type of mapping is easy to evaluated in English, however, the “not exact” correspondence will be useful for the translation of English terms to French.

4 Results

4.1 Conceptual Approach

A number of 199,786 mappings exists between at least two French terms from UMLS (25,833 (Exact Mapping), 69,085 (Close Mapping) and 104,868 (Broader and /or Narrower Mapping)). In contrast, from the 25,833 terms mapped “Exactly”, 15,831 come from SNOMED International where only 296 come from ICPC2 (Table 4). The three types of mappings (“Exact”,“Broader” and/or “Narrow” and “Close”) are included into the HMTP (see Figure 1).

4.2 Lexical Approach

A number of 266,139 mappings exists between at least two terms from HMTP (English and French). Most of these mappings were evaluated in previous work Merabti et al. (2010a);

[4]National Library of Medicine: Lexical Tools:
`http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2010/docs/userDoc/index.html`

Terminology	Number of terms mapped
ICD10	3,282 (35%)
ICPC2	296 (39%)
MedDRA	5,700 (28%)
MeSH	10,637 (40%)
SNOMED Int.	15,831 (14%)
WHO-ART	1,392 (81%)

TAB. 4 – *Number of terms from each terminology exactly mapped (conceptual approach).*

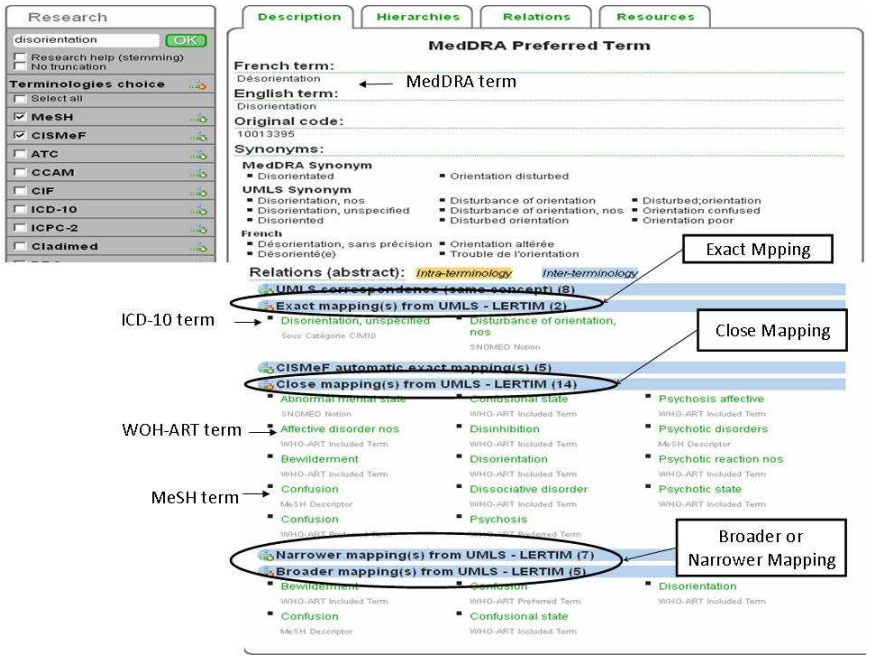

FIG. 1 – *The three types of conceptual approach integrated into the HMTP (Example of the MedDRA term "Disorientation").*

Merabti (2010); Merabti et al. (2010b). For example, in Merabti et al. (2010a) the "Exact mapping" between ORPHANET and some French terminologies was evaluated and considered "relevant" in 98% of cases. Table 5 displays a fragment of the entire matrix mapping between all terminologies of the HMTP. Some of these mappings were evaluated such as ORPHANET to MeSH, MedDRA, ICD10, WHO-ART or WHO-ATC to MeSH Darmoni et al. (2010). Nevertheless, most of the mappings are not yet evaluated. Terminologies included in the HMTP in English and French were mapped using the two lexical approaches. For example, the terminologies: MeSH, SNOMED International, ORPHANET and ATC were mapped using English and French lexical approches. However, some terminologies were mapped using English (SNOMED CT, PSIP Taxonomy) or French (CISMeF, DRC) lexical approache alone.

All exact mappings are integrated into the HMTP (example of figure 2).

	FMA	MedDRA	MeSH	ORPHANET	SNOMED Int	WHO-ART
CCAM	0	110	305	0	430	5
CISMeF	9	99	517	11	222	17
CISP2	7	138	219	30	254	109
CLADIMED	35	24	258	3	259	4
Codes for drugs	0	24	1,455	3	302	0
FMA		119	1,745	32	5,777	3
ICD10	10,209	2,380	3,827	947	7,474	1,134
IDIT	0	79	75	0	0	0
IUPAC	58	32	726	8	317	11
LPP	0	0	36	0	22	0
MedDRA	119		3,728	885	5,360	1,278
MEDLINEPlus	34	314	675	138	448	170
MeSH	1,745	3,728		1,805	15,127	1,417
ORPHANET	32	885	1,805		1,635	284
SNOMED Int	5,777	5,360	15,127	1,635		1,747
UNIT	0	0	77	0	0	0
WHO-ART	3	1,278	1,417	284	1,747	
WHO-ATC	61	58	3,533	0	1,581	4
WHO-ICF	178	9	294	2	222	7
WHO-ICPS	1	13	159	0	114	6

TAB. 5 – *Fragment of the entire matrix mapping from HMTP.*

4.3 Comparing the two approaches

As shown in Table 6, the lexical approach was able to find 8,680 more mappings for MeSH and 50,116 for SNOMED International than the conceptual approach. For example, the mapping between the MeSH term “Oral Hygiene” and the SNOMED International “Dental hygienist” was found only by the lexical approach.
The conceptual approach founded 95 more mappings for MeSH and 192 for SNOMED International than lexical approach. For example, the mapping between the MeSH term “Acute-phase proteins” and the SNOMED International term “acute phase reactant” was found only by the conceptual approach.

5 Discussion

The aim of this study was to proposed conceptual and lexical methods to map several biomedical terminologies whatever or not included into UMLS. Methods developed can be

FIG. 2 – *Mapping of the MeSH term "myocardial infarction" according to the lexical approach in HMTP (Exact correspondence).*

Terminology	MeSH	SNOMED International
Number of terms mapped by the two approaches	10,542 (41%)	15,639 (14%)
Number of terms mapped only by the conceptual approach	95(0.3%)	192(0.17%)
Number of terms mapped only by the lexical approach	8,680(33%)	50,116(45%)

TAB. 6 – *The number of MeSH and SNOMED International terms mapped according to each approach.*

applied to map English or French terms. The results obtained through these methods are different according to the type of terminology and the number of target terms used to map the source terminology. For example, using the conceptual approach, only 10,637 MeSH terms were mapped, whereas 19,222 MeSH terms including the MeSH Supplementary Concepts (n= 186,702) were mapped using a lexical approach. The difference between this two figures can be easily explained by the difference of the target terms used by the two approaches. However, the number of mappings also differs according to the type of terminology. For example, in the lexical approach, there are 1,635 mappings between ORPHANET (terminology for rare dis-

ease) and SNOMED International when 15,127 mappings were obtained between MeSH and SNOMED International (see Table 5).
These methods are also used to translate some of English terminologies to French (SNOMED CT Joubert et al. (2009), MEDLINEPlus Deléger et al. (2010)). Lexical approaches are limited in the management of the ambiguous acronyms. For example, the acronym "CMT" corresponds to "Charcot-marie-tooth disease" in MeSH and "Thyroid neoplasms". Another limit of the lexical approach concerns terms lexically close but with a different meaning such as "left" (gauche) and "Gaucher disease" (maladie de Gaucher). Difference in knowledge representation and terminological differences can also cause some problems in the lexical mappings as stressed in Bodenreider and Zhang (2006). For example, there is a mapping between the MeSH term "Marfan syndrome" and the SNOMED International term "Arachnodactyly" because there is a shared synonyms "Dolichostenomelia" between the two terms. However, the same mapping between the ORPHANET term"Marfan syndrom" and the SNOMED international term "Arachnodactyly" was evaluated as false by an ORPHANET expert because "Arachnodactyly" corresponds as a sign of the "Marfan syndrom". In perspective, we are working on a third approach based on statistical method (co-occurence).

6 Conclusion

Automatic mapping between biomedical terminologies integrated in the HMTP in English and French was achieved. These mappings were also used to translate English terminologies to French such as FMA, MEDLINEPlus and SNOMED CT.

7 Acknowledgements

Multi-terminology portal was supported in part by the grants InterSTIS project (ANR-07-TECSAN-010); ALADIN project (ANR-08-TECS-001); L3IM project (ANR-08-TECS-00); PSIP project; (Patient Safety through Intelligent Procedures in medication -FP7-ICT-2007-); PlaIR project, funded by FEDER.

References

Bodenreider, O., S. J. Nelson, W. T. Hole, and H. F. Chang (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In *Proc. AMIA Symp. 1998*, pp. 815–819.

Bodenreider, O. and S. Zhang (2006). Comparing the representation of anatomy in the FMA and SNOMED CT. In *AMIA Annu Symp Proc*, pp. 46–50.

Browne, A., D. G, A. Aronson, and M. AT (2003). Umls language and vocabulary tools. In *AMIA Annu Symp Proc*, pp. 798.

Cimino, J. and G. Barnett (1990). Automated translation between terminologies using semantic defintions. *MD Comput 7*, 104–109.

Darmoni, S., J. Grosjean, T. Merabti, B. Dahamna, I. Kergouraly, L. Soualmia, and B. Thirion (2010). Health multi-terminology portal: semantics added-value for quality-controlled health gateway. *Journal of Biomedical Semantic*. Submitted.

Darmoni, S., M. Joubert, B. Dahamna, J. Delahousse, and M. Fieschi (2009). Smts: a French Health Multi-Terminology Server. In *Proc. AMIA Symp. 2009*.

Darmoni, S., J.-P. Leroy, F. Baudic, M. Douyère, J. Piot, and B. Thirion (2000). CISMeF : a structured health resource guide. *Methods of Information in Medicine 39*, 30–35.

Darmoni, S., C. Letord, T. Merabti, , S. Sakji, and I. Kergourlay (2010). ATC to PubMed: a bibliographic tool for drugs. In *Proc. AMIA Symp 2010*. In Press.

Deléger, L., T. Merabti, T. Lecroq, M. Joubert, P. Zweigenbaum, and S. Darmoni (2010). A Twofold Strategy for Translationg a Medical Terminology into French. In *Proc. AMIA Symp. 2010*. In press.

Fung, K. and O. Bodenreider (2005). Utilizing UMLS for semantic mapping between terminologies. In *Proc AMIA Symp*, pp. 266–270.

Imel, M. (2002). A closer look: the SNOMED clinical terms to ICD-9-CM mapping. *J AHIMA 73*(6), 66–9; quiz 71–2.

Joubert, M., H. Abdoune, T. Merabti, S. Darmoni, and M. Fieschi (2009). Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. In *Proc. AMIA Symp. 2009*, pp. 291–295.

Lindberg, D., B. Humphreys, and A. McCray (1993). The Unified Medical Language System. *Methods Inf Med 32(4)*, 281–291.

McCray, A., S. Srinivasan, and A. Brown (1994). Lexical methods for managing variation in biomedical terminologies. In *Annual Symposium on Computer Applications in Medical Care*, pp. 235–239.

Merabti, T. (2010). *Methods to map health terminologies: contribution to the semantic interoperability between health terminologies*. Ph. D. thesis, University of Rouen.

Merabti, T., M. Joubert, T. Lecroq, A. Rath, and S. Darmoni (2010a). Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH. *Biomedical Engineering and Research 31*(4), 221–225.

Merabti, T., P. Massari, M. Joubert, E. Sadou, T. Lecroq, H. Abdoune, J. Rodrigues, and S. Darmoni (2010b). Automated approach to map a French terminology to UMLS. In *MedInfo2010*, Volume 160(pt 2), Cap Town, South Africa, pp. 1040–4.

Pereira, S. (2007). *Muti-Terminology indexing of concepts in health*. Ph. D. thesis, University of Rouen.

Peters, L., J. Kapusnik-Uner, and O. Bodenreider (2010). Methods for managing variation in clinical drug names. In *Proc Annu Symp AMIA 2010*. In press.

Rocha, R., B. Rocha, and S. Huff (1994). Automated translation between medical vocabularies using a frame-based interlingua. In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pp. 690–694.

Wang, Y., J. Patrick, G. Miller, and J. O'Hallaran (2008). A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. *BMC Medical Informatics and Decision*

Making 8(Suppl 1), pp. S5.

Résumé

Terminology mapping is an important and crucial task to improve semantic interoperability between health care applications and resources. In 2009, CISMeF created a Health Multi-Terminological Portal (HMTP) to search concepts among all the health terminologies available in French (or in English and translated in French) included in this portal and to browse it dynamically. To map terminologies in the HMTP, two methods are used: (1) conceptual method which exploits various features of the UMLS, (2) lexical method based on natural language processing in French and English. A total of 199,786 mappings were performed between at least two French terms using conceptual method, whereas 266,139 mappings were performed using lexical methods. These mappings were all integrated in the HMTP developed by CISMeF. Conceptual and lexical methods were used to translate some English terminologies into French such as MEDLINEPlus, FMA and SNOMED CT.

Réseau d'Observation en Télécardiologie : Modèle de Données Basé sur HL7

Radja Messai*, Julie Jacques**, Régis Beuscart*

*CERIM, E.A. 2694, Faculté de Médecine, Université de Lille2, Lille, France
radja.messai@univ-lille2.fr
http://cerim.univ-lille2.fr
**Alicante, Seclin, France
julie.jacques@alicante.fr
http://www.alicante1.moonfruit.fr

Résumé. La télécardiologie constitue un moyen priviligié pour le suivi des patients possédant un dispositif cardiologique implantable. Cependant ce type de dispositifs génère des alertes sur la base de seuils préétablis et qui sont indépendants de la situation clinique du patient ou du traitement qu'il suit. Le projet ANR Akenaton a pour objectif principal de créer une plateforme pour l'intégration des données médicales pour la prise en compte du contexte clinique du patient dans la prise de décision face à une alerte provenant de son dispositif médical. Pour réaliser cet objectif nous avons créé un modèle de données qui définit le schéma physique de l'entrepôt de données pour l'intégration des données cliniques du patient basé sur le modèle d'information de référence du standard HL7 V3.

1 Introduction

Les dispositifs médicaux implantables utilisés dans le domaine de la cardiologie, tels que les défibrillateurs ou les stimulateurs, constituent une avancée notable dans la thérapie des maladies cardiaques chroniques. Cependant, un suivi régulier et rigoureux est encore nécessaire pour assurer la sécurité des patients. La télécardiologie est une approche assez récente dans le suivi des patients et qui s'appuie sur des services de télécommunication pour transmettre les données liées aux dispositifs médicaux aux centres de contrôle d'une manière régulière. Après traitement et décodage des informations, des messages d'alertes sont envoyés aux cardiologues, s'il y a lieu, par e-mail, SMS, fax, voire téléphone. Le cardiologue pourra ensuite consulter le détail de l'alerte sur un site sécurisé.

Cependant, ces alertes sont générées sur la base de seuils préétablis ce qui peut générer un nombre d'alertes important qui n'ont parfois pas à avoir lieu en regard du contexte clinique ou du traitement suivi par le patient. Le projet Akenaton est un projet ANR qui a pour but de créer une plateforme d'intégration des données médicales et des données liées aux dispositifs pour améliorer le suivi des patients par la prise en compte de leurs contextes cliniques en utilisant des outils de représentation des connaissances et d'aide à la décision (Burgun et al., 2010).

Ce papier s'intéresse à la création du modèle de données qui a été utilisé pour intégrer les données liées au patient et au dispositif implantable. Ces données sont d'une nature hétérogène et proviennent de plusieurs sources différentes. Elles représentent les diagnostics, les traitements et les résultats de laboratoire liés au patient ainsi que les alertes qui proviennent du dispositif implantable. Le modèle de données abrite également les données nécessaires pour le processus de raisonnement et de la prise de décision.

Le modèle de données vise à devenir un entrepôt de données qui abriteraient des données provenant de plusieurs sources. Cela pose le problème d'importation de données qui ne possèdent pas le même schéma. Actuellement il y a plusieurs efforts de standardisation des données médicales pour faciliter l'échange de ce type de données. Parmi ces efforts on peut citer celui de l'organisation HL7 qui vise à définir des mécanismes pour le codage et l'échange des données médicales. Si un modèle de données est basé sur HL7 on facilitera l'échange de données avec tous les systèmes qui implémentent une interface HL7. Pour cette raison nous avons basé le développement de notre modèle de données sur le modèle d'information de référence de HL7 nommé le RIM (Reference Information Model).

2 HL7 V3 Reference Information Model

L'organisation HL7 [1], accréditée par l'ANSI (American National Standards Institute), a pour mission de développer des standards dans le domaine de la santé. Les membres de HL7 incluent les principales organisations de santé dans le monde. Les standards HL7 définissent la transmission et l'échange des données médicales entre les applications informatiques, les systèmes et les organisations. La suite de standards HL7 Version 2.X (Version 2.7 est la plus récente) est considérée comme le cheval de trait dans l'échange de données médicales et est la plus largement implémentée à travers le monde.

En 1997, HL7 a commencé à travailler sur le standard HL7 Version 3 (HL7 V3) avec une approche de développement des messages significativement différente des versions précédentes. Toutes les spécifications des messages du standard HL7 V3 sont dérivées du Modèle d'information de référence RIM (Reference Information Model). Le RIM est un mdoèle statique des informations médicales qui couvrent d'une manière abstraite tous les aspects cliniques et administratifs d'une organisation de santé (Schadow, 2006). Dans le processus de développement d'un message HL7 V3, le RIM est contraint pour couvrir uniquement les informations nécessaires pour un message particulier. Même si le RIM n'est pas initialement conçue pour la conception des base de données, il fournit un modèle intégré pour les données médicales et nous pensons qu'il peut constituer une base adéquate pour l'architecture d'un entrepôt de données.

Le modèle HL7 RIM est constitué des classes de base suivantes :

- Act : représente une action qui s'est passée, qui est entrain de se passer ou qui va se passer.
- Entity : représente un objet physique ou un être tels que des personnes, des lieux ou des dispositifs.
- Role : représente le rôle que les entités jouent en participant à des actes médicaux.
- RoleLink : représente une relation entre deux rôles.

1. http ://www.hl7.org

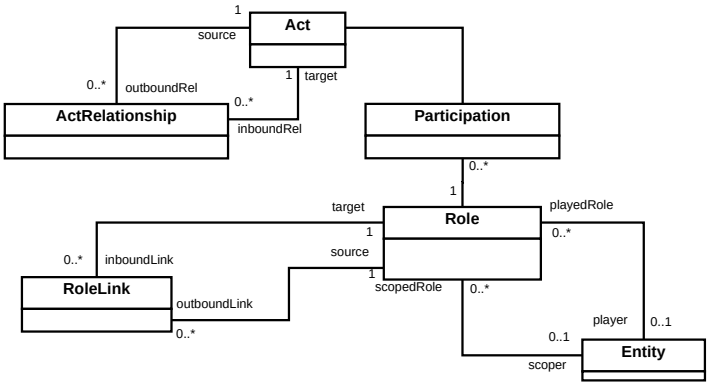

FIG. 1 – *Classes de base du RIM*

– Participation : représente une relation entre un rôle et un acte (par exemple, le contexte d'un acte, tel que : qui l'a réalisé, pour qui, ou où il a été réalisé).
– ActRelationship : représente la relation entre deux actes (par exemple, la relation entre la demande d'une analyse de sang et le résultat de cette analyse.

La figure 1 montre un modèle UML simplifié correspondant aux classes de base du RIM. Ces classes possèdent des sous-classes qui désignent des concepts plus spécifiques, par exemple, la classe *LivingSubject* est une sous-classe de la classe *Entity* et la classe *Person* est une sous-classe de la classe *LivingSubject*.

Le RIM définit un ensemble prédéfini d'attributs pour chaque classe. Chaque attribut a un type de données spécifique. Ces attributs et types de données deviennent des balises XML dans les messages HL7 V3. Une spécification donnée de messages utilise un sous-ensemble d'attributs RIM énumérant chaque élément et spécifiant combien de fois cet élément peut être répété. Ce processus s'appelle l'affinage du RIM. Le listing 1 montre un extrait d'un message HL7 V3.

```
<value code="493.91"
        codeSystem="2.16.840.1.113883.6.2"
        codeSystemName="ICD9-CM"
        displayName="Asthma">
        <originalText>
                Alergic asthma
        </originalText>
</value>
```

Listing 1 – *Extrait d'un message HL7 V3*

Cet extrait du message a pour but d'envoyer une information sur un diagnostic lié à un patient. Cette donnée possède un type de données CV (Coded Value) qui est un type de données propre à HL7. Il permet de représenter une information en utilisant les élements suivants : code, codeSystem, codeSystemName, codeSystemVersion, displayName, originalText.

Nous décrivons dans les sections suivantes notre approche pour implémenter un modèle physique basé sur le RIM. Ce modèle permettra d'acquérir facilement des données prove-

nant de messages HL7 V3 et d'en produire. L'intégration des données basée sur des standards d'échange permet un partage large des données à travers les systèmes ce qui permet de les réutiliser pour des fins différentes de celles de leur production.

3 Implémentation du data model basé sur le RIM

3.1 Implémentation des classes de bases de HL7

Implémenter le modèle de données basé sur le RIM signifie de concevoir un modèle physique de base de données. Il est important de créer un modèle physique qui correspond étroitement au modèle logique basé sur le RIM pour que l'importation de données à partir de messages HL7 soit la plus simple et la plus directe possible. Le modèle de données est inspiré du travail publié dans (Eggebraaten et al., 2007) et adapté à nos besoins spécifiques de la télécardiologie.

Toutes les classes de base du RIM, à l'exception de la classe *Observation* que nous discuterons plus tard, sont modélisées en utilisant l'approche Entité-Relation. Chaque classe correspond à une table physique dans le modèle. Pour maintenir les relations entre les classes telles qu'elles sont définies dans le RIM, des clés étrangères sont ajoutées aux classes appropriées.

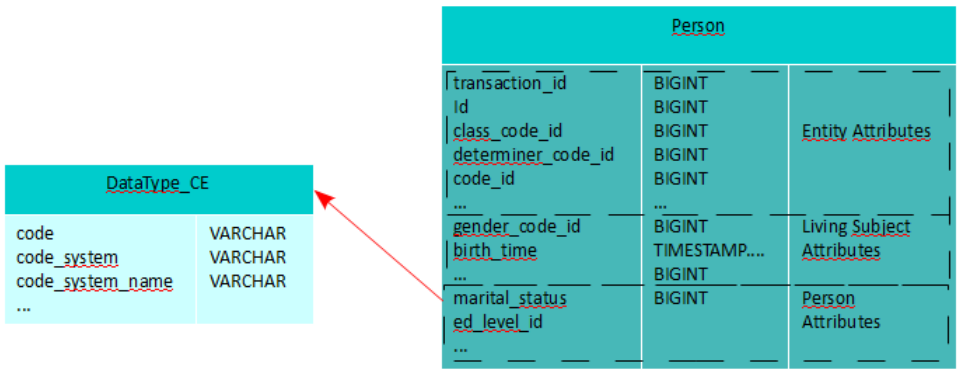

FIG. 2 – *Classe personne : attributs hérités et types de données*

Les classes dans le RIM peuvent hériter des attributs de leurs classes mères. Dans le schéma physique, nous construisons une table unique de chaque classe dans le modèle avec tous les attributs hérités de la classe mère. La figure 2 montre la structure de la classe *Person* avec les attributs hérités des deux classes mères *Entity* et *LivingSubject*. Les attributs des classes RIM possèdent des types de données complexes créés par HL7. Ces types de données peuvent être assimilés à des enveloppes qui englobent un ensemble de valeurs concernant un attribut donné. Par exemple, l'attribut *marital status* possède un type de données *coded value* qui permet de représenter les informations concernant l'attribut sous la forme d'un quadruplet : *code*, *code_system*, *version* et *display_name*. Ce type de codage permet de représenter l'information sans ambiguité ce qui est la finalité du standard HL7. Dans notre modèle de données nous

avons attribuer à chaque attribut une valeur qui représente une clé étrangère vers une table qui représente le type de données en question et qui contient toutes les informations concernant l'attribut. La figure 2 illustre la structure de la table avec un lien de l'attribut *marital_status_id* vers la table du type de données correspondant et qui contient les informations sur la situation familiale de la personne.

3.2 Implémentation de la classe *Observation* du RIM

La classe *Observation* qui est une sous-classe de la classe *Act* dans le RIM codent plusieurs types de données dans les messages HL7, tels que les diagnostics, les résultats de laboratoire, les allergies et les signaux vitaux. Une instance observation possède dans la majorité des cas un identificateur ID, un code et une valeur. Le code identifie l'observation en question (ex., glucose) et la valeur représente ce qui a été observé (ex., 35 mg). L'attribut valeur (*value* dans le RIM) a un type de données *Any*, ce qui signifie qu'il peut être de n'importe quel type de données défini dans HL7. En pratique le type de données est contraint au moment de l'affinage du RIM pour la production d'un message pour un besoin spécifique. Cependant, le modèle de données est conçu pour abriter toutes les valeurs possibles d'observation. Pour cela, nous avons construit une table observation par type de données, par exemple la table *OBSERVATION_INT* a pour objectif de stocker les observations de type INT et ainsi de suite. La figure 3 montre un exemple de tables d'observation par type de données.

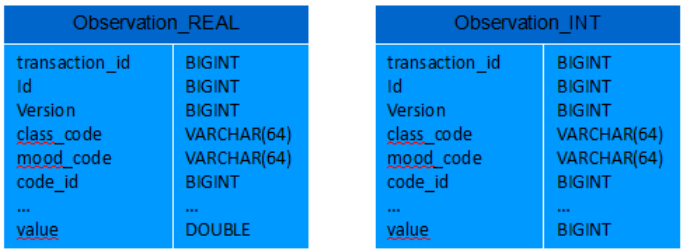

FIG. 3 – *Classe Observation : Modélisation des types de données*

4 Conclusion

Le modèle d'information de référence RIM de HL7 est un modèle qui couvre tous les aspects d'informations administratives ou cliniques d'une organisation de santé. Les experts de l'industrie de santé ont créé le RIM pour implémenter des méthdologies de création de messages utilisés dans l'échange des informations médicales. En dérivant un modèle physique du RIM nous bénéficions de l'expertise de l'organisation HL7 et permettons le traitement des messages HL7 V3 dans notre base de données. Ce type de modèle de données permet d'importer facilement des données provenant de messages HL7. Il permet également de produire directement des messages HL7 ce qui permet d'atteindre une interopérabilité assez facilement

avec des systèmes de santé du moment qu'ils implémentent une interface HL7. Cependant les données qui ne sont pas codées selon le standard HL7 doivent passer par un prétraitement pour déterminer le type de données HL7 adéquat à utiliser pour coder une telle information. Par la suite, l'avantage est que ces données deviennent échangeables avec tous systèmes compatibles HL7.

Ce travail répond principalement à la problématique de l'interopérabilité syntaxique. Cependant, un travail est encore nécessaire pour résoudre la question de l'interopérabilité sémantique. L'utilisation des systèmes d'encodage internationaux tel que LOINC pour la réprsentation des résultats de laboratoire et la normalisation des données sont la solution pour atteindre l'interopérabilité sémantique (Khan, 2006). En France, peu de données cliniques sont normalisées et une grande quantité est sous format non-structuré. Beaucoup de travail reste à faire pour la réutilsation des données existentes et surtout développer des systèmes qui prévoient l'encodage de l'information d'une manière standardisée (Wurtz, 2005).

5 Remerciements

Ce travail a été financé par le projet Akenaton (ANR-07-TECSAN-001-06).

Références

Burgun, A., A. Rosier, L. Temal, O. Dameron, P. Mabo, P. Zweigenbaum, R. Beuscart, D. Delerue, et C. Henry (2010). Supporting medical decision in telecardiology : A patient-centered ontology-based approach. In *MedInfo*.

Eggebraaten, T. J., J. W. Tenner, et J. C. Dubbels (2007). A health-care date model based on the hl7 reference information model. *IBM Systems Journal 46*(1), 5–18.

Khan, A. N., G. S. P. M. C. R. D. R. A. C. J. . B. J. (2006). Standardizing laboratory data by mapping to loinc. *J Am Med Inform Assoc 13*(3), 353–355.

Schadow, G. (2006). Hl7 reference information model compendium. Booklet.

Wurtz, R. (2005). Elr, loinc, snomed, and limitations in public health. White paper.

Summary

Telecardiology is a privileged means for monitoring patients with implantable cardiac devices. However this type of device generates alerts based on predefined thresholds that are independent of the patient's clinical situation or his treatment. The main objective of the ANR Akenaton project is to create a platform for the integration of medical data to take into account the clinical context of the patient and make appropriate decisions face to an alert from his medical device. To achieve this goal we created a HL7 V3 RIM based data model that defines the physical schema of the data warehouse for integrating patient clinical data.

Prédiction du grade d'un cancer du sein par la découverte de motifs séquentiels contextuels dans des puces à ADN

Julien Rabatel[*,**], Mickaël Fabrègue [*], Sandra Bringay[*,***], Pascal Poncelet[*], Maguelonne Teisseire[****]

[*]LIRMM, Université Montpellier 2, CNRS
161 rue Ada, 34392 Montpellier Cedex 5, France
[**]Tecnalia, Cap Omega, Rond-point Benjamin Franklin - CS 39521
34960 Montpellier, France
[***]Dpt MIAP, Université Montpellier 3, Route de Mende
34199 Montpellier Cedex 5, France
[****]TETIS, 500 rue Jean-François Breton
34093 Montpellier Cedex 5, France

Résumé. Le cancer du sein reste de nos jours un problème de santé majeur et un véritable défi pour les biologistes et les professionnels de santé. Les puces à ADN permettent aujourd'hui d'étudier selon un jour nouveau les problématiques associées à cette maladie. Dans cet article, nous proposons de traiter les données issues des puces à ADN par le biais de l'extraction de motifs séquentiels contextuels (séquences de gènes ordonnés selon leur niveau d'expression associées à un contexte). L'objectif est de proposer une aide au diagnostic du grade d'une tumeur. Notre approche tient à la fois compte de l'information contenue dans les puces à ADN (exprimée par le biais de motifs séquentiels), mais également d'informations additionnelles d'ordre contextuel (e.g., âge du patient, taille de la tumeur, etc.) et qui sont associées aux données de puces à ADN lorsque celles-ci sont publiées en ligne. L'approche proposée a été évaluée sur des données réelles.

1 Introduction

De nos jours, le cancer du sein représente un problème de santé majeur. Selon le site `breastcancer.org`, environ 1 femme sur 8 aux Etats-Unis est susceptible de développer un cancer du sein au cours de sa vie, et on estime à plus de 200 000 le nombre de nouveaux cas pour l'année 2010. De nouveaux outils de diagnostic sont continuellement créés pour détecter des types spécifiques de cancer et de nouveaux traitements sont également développés pour s'appliquer à ces différents types de cancer. L'objectif est ainsi de réduire les effets indésirables tels que des disfonctionnements cardiaques ou des ménopauses prématurées par exemple en proposant aux patients des traitements ajustés aux spécificités de leur cancer. Cependant, malgré les nombreuses avancées dans le domaine, jusqu'à 50% des femmes touchées développeront des métastases distantes, qui restent malheureusement incurables.

Les trois principaux défis associés aux cancer du sein aujourd'hui sont les suivants : *(1)* comment diagnostiquer un cancer du sein le plus tôt possible (dépistage) et identifier le type d'une tumeur, *(2)* comment prédire la réaction d'un patient à un traitement donné en fonction des informations disponibles sur les historiques des précédents patients (sur leur type de tumeur et sur les résultats des traitements qu'ils ont reçu), et *(3)* comment proposer un choix de thérapie pour un patient donné en fonction des prédictions précédentes et en étant capable de l'informer sur ses chances de rémission, sur le développement prévisible de la maladie ainsi que les conséquences possible des traitements proposés.

Dans cet article, nous proposons d'exploiter les connaissances qui ont été rendues publiques par différentes équipes de biologistes travaillant sur le cancer du sein et qui sont disponibles sur Internet pour commencer à apporter des solutions aux trois défis précédents. Ces équipes ont mutualisé les résultats issus de l'analyse de puces à ADN appliquées sur les tissus de différents types de cancer du sein. Les puces à ADN sont des outils puissants permettant de dresser un véritable portrait génétique d'un échantillon biologique (ici des échantillons de tumeurs) en comparant l'expression de milliers de gènes dans différents tissus, cellules, ou conditions. Simon et Dobbin (2003) décrivent trois moyens d'exploiter les puces à ADN :

– **Comparer des classes** consiste à identifier des variations (e.g., dans l'expression des gènes) entre plusieurs classes. Il est possible de comparer des tissus normaux à des tumeurs (Alon et al. (1999)), ou des tumeurs qui réagissent bien à une thérapie à celles qui ne réagissent pas (Rosenwald et al. (2002)), ou encore de distinguer différents types de tumeurs (Dougherty (2001); Sotiriou et al. (2003)). Il est ainsi possible d'identifier des gènes ayant des comportements différents dans différentes classes et que l'on suppose jouer un rôle important dans une maladie. Ces informations peuvent alors être utilisées dans un but de prédiction.

– **Découvrir des classes** consiste à découvrir de nouvelles catégories dans une population (e.g., de nouvelles catégories de tumeurs). Par exemple, Sotiriou et al. (2003) a utilisé les données issues de l'analyse de puces à ADN appliquées à des tissus issus de tumeur du sein pour décrire différents types de cancer du sein.

– **La classification** exploite les résultats précédents pour affecter une nouvelle puce à ADN à une classe connue (e.g., pour associer cette nouvelle puce à un type de tumeur dans Van De Vijver et al. (2002)). Si le classifieur construit est suffisamment fiable, ses résultats peuvent être utilisés pour assister les professionnels de santé dans leur prise de décision lors d'un dépistage, par exemple.

Le nombre de données générées à partir de l'analyse de puces à ADN est considérable. Les méthodes statistiques et les méthodes de fouille de données jouent alors un rôle important dans la découverte de nouvelles connaissances. Cependant, leur mise en place est difficile du fait du grand nombre de valeurs mesurées par puce comparé au nombre de puces testées. Ce problème est connu sous le nom de « fléau de la dimension » (*curse of dimensionality*) (Dougherty (2001)). Par ailleurs, il existe des corrélations entre les expressions des gènes qui ne sont pas toujours bien connues et les valeurs d'expression des gènes sont également souvent entachées de bruit lié aux dispositifs expérimentaux. Pour toutes ces raisons, le problème de

A

Puce	G_1	G_2	G_3	G_4	G_5
$P1$	7.3	6.6	6.6	9.5	8.1
$P2$	5.6	7.4	5.6	5.3	7.9
$P3$	5.7	5.2	8.7	6.8	6.2

B

Puce	Séquence
$P1$	$\langle(G_2G_3)(G_1)(G_5)(G_4)\rangle$
$P2$	$\langle(G_4)(G_1G_3)(G_2)(G_5)\rangle$
$P3$	$\langle(G_2)(G_1)(G_5)(G_4)(G_3)\rangle$

FIG. 1 – **A** - *Puces à ADN.* **B** - *Séquences correspondantes.*

la classification de puces à ADN est différent des problèmes de classification connus et les méthodes traditionnelles rencontrent peu de succès (Zupan et al. (2000)).

L'objectif de l'étude menée ici est de développer une méthode pour extraire puis exploiter des séquences de gènes ordonnés selon leur niveau d'expression ainsi que des informations relatives au contexte lié à chaque patient (e.g., son âge) dans un but de classification, afin d'aider les professionnels de santé à diagnostiquer un type de cancer et à choisir la meilleure thérapie possible. Les séquences sont générées à partir des données issues de l'analyse de puces à ADN et sont basées sur une technique largement utilisée en fouille de données : l'extraction de motifs séquentiels. Salle et al. (2009) décrit un algorithme efficace pour extraire de telles séquences et montre l'intérêt de ces motifs pour distinguer différentes classes. Par exemple, le motif $\langle(17aag\_ovca\_dn)(tgz\_adip\_up)\rangle\ 80\% - Gr1,\ 40\% - Gr2$ signifie *« Pour 80% des tumeurs de Grade 1 et 40% des tumeurs de Grade 2, le niveau d'expression du gène* $17aag\_ovca\_dn$ *est inférieur à celui du gène* tgz_adip_up *»*. Les motifs séquentiels ont déjà été utilisés avec succès pour la classification de textes (Jaillet et al. (2006)). Nous allons montrer dans la suite leur pertinence pour la classification de tumeurs malgré la complexité des données et la teneur interdisciplinaire de ces travaux.

La suite de l'article est organisée de la manière suivante. Nous introduisons dans la section 2 l'extraction de motifs séquentiels contextuels dans les puces à ADN. La section 3 décrit la méthode de classification développée. L'évaluation de notre approche est exposée dans la section 4. Enfin, nous concluons et discutons les perspectives de ces travaux dans la section 5.

2 Extraction de motifs séquentiels dans les puces à ADN

La première étape de nos travaux concerne l'extraction des motifs séquentiels dans les puces à ADN. Dans un premier temps, nous décrivons comment les puces à ADN peuvent être traduites sous-forme d'une base de séquences, sur laquelle les sous-séquences fréquentes seront extraites. Ensuite, nous montrons comment les informations additionnelles liées à chaque patient (e.g., l'âge, la taille de la tumeur, etc.) peuvent être exploitées de manière à affiner nos connaissances.

2.1 Puces à ADN

La figure 1-A présente la structure des puces à ADN que nous manipulons. Soit G_1, G_2, ..., G_5 les gènes traités, et $P1, P2, P3$ des puces à ADN. Une puce contient une valeur d'expression pour chacun des gènes. Par exemple, la valeur de l'expression du gène G_1 dans la puce $P1$ est 7.3.

2.2 Motifs séquentiels et puces à ADN

Les motifs séquentiels traditionnels, introduits par Agrawal et Srikant (1995), peuvent être considérés comme une extension du concept d'itemsets fréquents de Agrawal et al. (1993) en considérant les estampilles temporelles associées aux items. La fouille de motifs séquentiels vise à l'origine à extraire des ensembles d'items fréquemment associés au cours du temps. En considérant l'étude des achats dans une boutique, un motif séquentiel pourrait par exemple être : «*40 % des clients achètent une télévision, puis plus tard achètent un lecteur DVD*».

Salle et al. (2009) propose une réprésentation des puces à ADN permettant l'extraction de motifs séquentiels dans le but de mieux caractériser différentes classes de puces. La découverte de motifs séquentiels dans les puces à ADN est définie comme suit.

Soit $\mathcal{G}$ un ensemble de ***gènes*** distincts. Un ***itemset*** est un sous-ensemble de gènes, noté $I = (g_1 g_2 ... g_n)$, tel que les gènes $g_1, ... g_n$ ont la même expression notée $exp(I)$. Une ***séquence*** s est une liste ordonnée d'itemsets notée $\langle I_1 I_2 ... I_k \rangle$, telle que $exp(I_1) < exp(I_2) < ... < exp(I_k)$. Ainsi, les gènes dans une séquence sont organisés en fonction de l'ordre de leur expression dans une puce donnée.

Exemple 1. *La figure 1-B montre, pour chaque puce, la séquence correspondante. Ainsi, la puce $P1$ est traduite par la séquence $\langle (G_2 G_3)(G_1)(G_5)(G_4) \rangle$.*

Soit $s = \langle I_1 I_2 ... I_m \rangle$ et $s' = \langle I'_1 I'_2 ... I'_n \rangle$ deux séquences. La séquence s est une ***sous-séquence*** de s', noté $s \sqsubseteq s'$, si $\exists i_1, i_2, ... i_m$ avec $1 \leq i_1 < i_2 < ... < i_m \leq n$ tel que $I_1 \subseteq I'_{i_1}$, $I_2 \subseteq I'_{i_2}$, ..., $I_m \subseteq I'_{i_m}$.

Exemple 2. *Soit les séquences $s = \langle (\mathbf{G_3})(\mathbf{G_4}) \rangle$ et $s' = \langle (G_2 \mathbf{G_3})(G_1)(G_5)(\mathbf{G_4}) \rangle$. s est une sous-séquence de s', i.e., $s \sqsubseteq s'$.*

Une ***base de séquences*** $\mathcal{B}$ est une relation $\mathcal{R}(ID, S)$, où un élément $id \in dom(ID)$ est un identifiant de séquence, et $dom(S)$ est l'ensemble des séquences. La ***taille*** de $\mathcal{B}$, notée $|\mathcal{B}|$, est le nombre de tuples dans $\mathcal{B}$. Un tuple $\prec id, s \succ$ ***supporte*** une séquence α si α est une sous-séquence de s, i.e., $\alpha \sqsubseteq s$. Le ***support*** d'une séquence α dans la base de séquences $\mathcal{B}$ est la proportion de tuples dans $\mathcal{B}$ supportant α, i.e. :

$$sup_{\mathcal{B}}(\alpha) = \frac{|\{\prec id, s \succ | \ (\prec id, s \succ \in \mathcal{B}) \wedge (\alpha \sqsubseteq s)\}|}{|\mathcal{B}|}.$$

Etant donné un nombre réel $minSup$ le seuil de ***support minimum***, tel que $0 < minSup \leq 1$, une séquence α est un ***motif séquentiel*** dans la base de séquences $\mathcal{B}$ si son support dans $\mathcal{B}$ est supérieur ou égal à $minSup$, i.e., $sup_{\mathcal{B}}(\alpha) \geq minSup$. La séquence α est alors dite ***fréquente dans*** $\mathcal{B}$.

Exemple 3. *Considérons la base de séquences $\mathcal{B}$ formée par les séquences de la figure 1-B, et un support minimum $minSup$ fixé à 1. Une séquence est un motif séquentiel si son support est supérieur ou égal à $minSup \cdot |\mathcal{B}| = 3$. La séquence $s = \langle (G_2)(G_5) \rangle$ est un motif séquentiel. En effet, son support est $sup_{\mathcal{B}}(s) = 3/3$ et est égal à $minSup$. En revanche, $s' = \langle (G_1)(G_4) \rangle$ n'est pas un motif séquentiel : son support est $sup_{\mathcal{B}}(s') = 2/3$ et est inférieur à $minSup$.*

2.3 Informations contextuelles

La découverte de motifs séquentiels dans les puces à ADN représentées sous forme de séquences permet d'extraire des connaissances précises sur différentes classes de données. Par exemple, il est possible de révéler que le motif $\langle(G_2)(G_5)\rangle$ est plus fréquemment associé aux cancers de grade 1 qu'aux autres grades. Ainsi, les motifs séquentiels extraits aideront à construire le profil génétique des différents grades de cancer, information utile pour effectuer une classification.

Cependant, les puces à ADN liées au cancer du sein sont fréquemment associées à des informations additionnelles. Par exemple, il sera possible de connaître l'âge du patient atteint, la taille de la tumeur détectée, etc. Ces informations, d'ordre contextuel, doivent être prises en compte afin de permettre une caractérisation plus fine des différents types de cancer, et ainsi permettre une classification plus fiable. Par exemple, nous cherchons à extraire des informations du type *« la séquence $\langle(G_3)(G_1G_4)\rangle$ est spécifique aux cancers de grade 1 chez les patients âgés de moins de 60 ans, tandis que la séquence $\langle(G_2)(G_3)\rangle$ est commune à tous les cancers de grade 2, peu importe l'âge du patient. »*

Rabatel et al. (2010) propose une définition formelle de telles informations contextuelles, ainsi qu'un algorithme d'extraction de motifs séquentiels contextuels (i.e., liés aux contextes auxquels ils sont spécifiques).

Les concepts liés à l'extraction de motifs séquentiels contextuels sont brièvement présentés dans l'exemple suivant.

Exemple 4. *Soit deux **dimensions contextuelles** Age (l'âge du patient lors du diagnostic) et Grade (le grade du cancer diagnostiqué). Nous noterons $dom(Age) = \{jeune, âgé\}$ le domaine de la dimension Âge et $dom(Grade) = \{1, 2\}$ le domaine de la dimension Grade. Un **contexte** est noté $[a, g]$ où $a \in dom(Age)$ et $g \in dom(Grade)$.*

Ainsi, le contexte $[jeune, 1]$ contient toutes les puces associées à un patient jeune atteint d'un cancer de grade 1. En introduisant une valeur joker $*$, nous pouvons également définir des contextes plus généraux. Par exemple, le contexte $[*, 1]$ correspond aux puces associées à un cancer de grade 1, pour un âge quelconque (i.e., $[*, 1]$ contient toutes les puces des contextes $[jeune, 1]$ et $[âgé, 1]$). Le contexte $[*, *]$ est quant à lui le contexte le plus général. Il représente l'ensemble des puces (i.e., pour un âge et un grade quelconque).

Les différents contextes définis peuvent ainsi être organisés selon la hiérarchie présentée dans la figure 2 (où j et a signifient respectivement $jeune$ et $âgé$).
*Dans cette hiérarchie, un motif séquentiel s est dit **spécifique à un contexte** c ssi :*

1. *il est fréquent dans c ainsi que dans tous les contextes descendants de c (s est alors dit **c-général**),*
2. *il n'existe pas de contexte plus général que c qui vérifie la première condition.*

Par exemple, le motif séquentiel s est spécifique à $[*, 1]$ ssi s est $[*, 1]$-général (i.e., il est fréquent dans $[jeune, 1]$ et $[âgé, 1]$) et s n'est pas $[*, *]$-général.

3 Classification de puces à ADN

Nous avons montré dans la section précédente comment extraire des motifs séquentiels contextuels dans des données issues de puces à ADN. Désormais, nous visons à exploiter ces

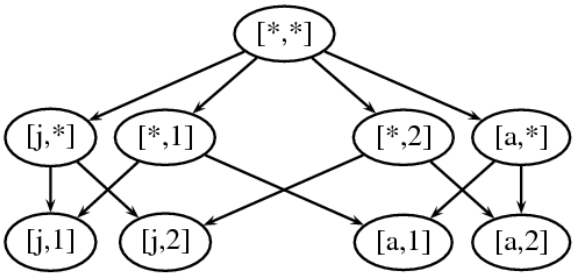

FIG. 2 – *Une hiérarchie de contextes.*

motifs pour améliorer la classification des puces. Le processus de classification de puces à ADN est constitué de trois étapes : *(i)* l'extraction des motifs séquentiels spécifiques pour chaque contexte considéré, *(ii)* la sélection des motifs les plus discriminants pour chaque contexte, et *(iii)* la prise de décision, i.e., l'association d'un grade à une puce à ADN donnée.

3.1 Définition du problème de classification

Dans cet article, le problème de classification que nous proposons de résoudre consiste en l'estimation du grade d'un cancer, à partir d'une puce à ADN liée à un patient. Cette méthode pourrait bien sûr s'appliquer pour d'autres objectifs de classification si l'on considère d'autres critères de classification des types de cancer que les grades comme proposé par Sotiriou et al. (2003). Pour ce faire, nous proposons d'utiliser les informations contextuelles disponibles afin de tirer profit de la précision des motifs séquentiels extraits.
Parmi les dimensions contextuelles définies dans la section 2, nous distinguons deux ensembles :

- **La dimension de classement.** Il s'agit de la dimension pour laquelle on souhaite faire une prédiction, désigner une valeur. Dans notre cas, la dimension *Grade* est la dimension de classement, i.e., nous ne connaissons pas sa valeur et nous voulons la prédire.
- **Les dimensions guides.** Il s'agit des dimensions dont nous connaissons la valeur, et sur lesquelles nous souhaitons nous appuyer pour mieux estimer la valeur des dimensions de classement. Par exemple, la dimension *Age* est une dimension guide dans le problème de classification des puces à ADN.

Soit $\mathcal{D}$ l'ensemble des dimensions contextuelles considérées. $D_1^g, ..., D_n^g$ sont les dimensions guides dans $\mathcal{D}$ et D^c est la dimension de classement dans $\mathcal{D}$.

Soit un contexte $c = [d_1^g, ..., d_n^g, d^c]$, où d_i^g est la valeur associée à la dimension D_i^g pour $i \in \{1, ..., n\}$ et d^c est la valeur associée à la dimension D^c. Le **guide** de c, noté $(d_1^g, ..., d_n^g)$, représente l'ensemble des valeurs associées à c sur les dimensions guides de $\mathcal{D}$.

Par la suite, nous nous intéresserons au cas où le guide d'un contexte est limité à une seule dimension. Ainsi, un contexte sera de la forme $[d^g, d^c]$.

Exemple 5. *Etant donné un patient âgé chez qui un cancer a été diagnostiqué, le problème de classification consiste à répondre à la question « Quel est le grade de son cancer (i.e., la valeur de la dimension de classement Grade), étant donné la puce à ADN de ce patient et son âge (i.e., la valeur de la dimension guide Age) ? »*

En d'autres termes, d'après la hiérarchie de contextes de la figure 2, il s'agira de décider, parmi les contextes $[\hat{a}g\acute{e}, 1]$ *et* $[\hat{a}g\acute{e}, 2]$*, quel est le plus proche du patient.*

3.2 Sélection des motifs séquentiels

Chaque contexte de la hiérarchie est associé à un ensemble de motifs séquentiels spécifiques. Cependant, tous ces motifs n'ont pas le même intérêt dans un but de classification. Ainsi, nous sélectionnons les k meilleurs motifs spécifiques pour chaque contexte.

3.2.1 Contextes conflictuels

L'utilisation d'une dimension guide permet de limiter le nombre de contextes possibles lors du processus de classification. Considérons l'exemple 5. L'âge du patient étant connu, il ne reste que deux contextes dans lesquels il peut être classé : $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$. En effet, les autres contextes ne respectent pas la contrainte d'âge que nous devons prendre en compte (e.g., $[*, 1]$ ou $[jeune, 2]$), ou ne contiennent pas de valeur pour la dimension de classement (e.g., $[\hat{a}g\acute{e}, *]$). Ces deux contextes sont conflictuels, i.e., le processus de classification doit être capable de les départager. Cette notion est importante car elle permettra par la suite de sélectionner les motifs qui seront les plus utiles pour la classification, i.e., qui seront les plus discriminants en considérant uniquement les contextes conflictuels.
De manière plus formelle, la notion de contextes conflictuels est définie comme suit.
Soit deux contextes $c = [d^g, d^c]$ et $c' = [d'^g, d'^c]$. Le contexte c' est **en conflit avec** c ssi :

- $d^g = d'^g$, i.e., les guides de c et c' sont identiques ;
- $d^c \neq *$ et $d'^c \neq *$, i.e., une valeur est bien définie sur la dimensions de classement ;
- $d^c \neq d'^c$, i.e., les deux contextes sont distincts.

Par la suite, nous noterons $conf(c)$ l'ensemble des contextes en conflit avec c.

3.2.2 Sélection

Soit un contexte c donné, nous cherchons l'ensemble des k motifs qui seront les plus pertinents dans un but de classification, i.e., les motifs les plus discriminants relativement aux autres contextes.

Nous définissons par conséquent une **mesure d'intérêt d'un motif** m pour caractériser un contexte c, notée $disc_c(m)$ et définie comme suit :

$$disc_c(m) = sup_c(m) - \max_{x \in conf(c)} sup_x(m).$$

Revenons sur cette définition. Son principe général est de considérer l'écart minimal qui existe entre le support de m dans le contexte c, et son support dans les autres contextes. Ainsi, un motif m sera jugé pertinent s'il est fréquent dans c et peu fréquent dans tous les autres contextes. De plus, notons que cette définition ne considère que les contextes en conflit avec c pour déterminer la pertinence de m. En effet, il est inutile de tenir compte des contextes avec lesquels c n'est pas en conflit. Pour nous en convaincre, considérons le contexte $[\hat{a}g\acute{e}, 1]$ et un motif m. Il est inutile de mesurer si m est discriminant par rapport à $[jeune, 2]$ car jamais le classifieur n'aura à comparer ces deux contextes. Le seul contexte qui sera comparé à $[\hat{a}g\acute{e}, 1]$ est $[\hat{a}g\acute{e}, 2]$, i.e., le seul contexte en conflit avec $[\hat{a}g\acute{e}, 1]$.

Exemple 6. *Le tableau 1 montre, pour les motifs m_1, m_2 et m_3 spécifiques à $[\hat{a}g\acute{e}, 1]$, leur support dans les contextes $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$, ainsi que la mesure d'intérêt de chacun d'eux dans $[\hat{a}g\acute{e}, 1]$. Ainsi, si nous souhaitons sélectionner les deux meilleurs motifs pour décrire ce contexte, m_2 et m_1 (par ordre d'intérêt décroissant) seront sélectionnés.*

motifs	$[\hat{a}g\acute{e}, 1]$	$[\hat{a}g\acute{e}, 2]$	$disc_{[\hat{a}g\acute{e},1]}(m)$
m_1	1	0.8	0.2
m_2	0.9	0.65	0.25
m_3	0.9	0.8	0.1

TAB. 1 – *Sélection des meilleurs motifs pour le contexte $[\hat{a}g\acute{e}, 1]$*

3.3 Prise de décision

A ce stade, chaque contexte de notre hiérarchie est associé à un ensemble de k motifs séquentiels. Désormais, nous considérons une nouvelle puce à ADN dont la séquence correspondante S est associée à une valeur d^g sur la dimension guide D^g. Il s'agit donc de classer S dans un des contextes dits **candidats**, i.e., de la forme $[d^g, d^c]$, tel que d^c est une valeur sur la dimension de classement D^c, en exploitant les motifs séquentiels associés à chacun de ces contextes.

Score dans un contexte. Soit un contexte candidat c et la séquence S à classer. La séquence S obtient un score dans c défini comme suit :

$$score_c(S) = |\{m \in c | m \sqsubseteq S\}|.$$

En d'autres termes, le score de S dans un contexte c est le nombre de motifs associés à c qui sont inclus dans S.

Classement. Connaissant le score de S dans chaque contexte candidat, le processus de classification consiste à prendre une décision finale : *« Dans quel contexte candidat est classé S ? »*

Nous proposons ici de retenir le contexte candidat dans lequel S a obtenu le meilleur score [1]. La valeur d^c associée à S par le classifieur sur la dimension de classement D^c est telle que :

$$score_{[d^g, d^c]}(S) = \max_{x \in dom(D^c)} sup_{[d^g, x]}(S).$$

Exemple 7. *Considérons la séquence $S = \langle (G_4)(G_5G_6)(G_3)(G_1)(G_2) \rangle$ associée à une nouvelle puce à ADN dont nous voulons estimer le grade. Le patient est $\hat{a}g\acute{e}$. Par conséquent, nous recherchons les contextes candidats : $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$, i.e., qui respectent la valeur $\hat{a}g\acute{e}$ sur la dimension guide Age et qui possédent une valeur sur la dimension de classement $Grade$.*

Le tableau 2 présente pour les contextes $[\hat{a}g\acute{e}, 1]$ et $[\hat{a}g\acute{e}, 2]$, les k meilleurs motifs de chaque contexte (pour $k = 3$).
Nous calculons le score de chaque contexte :

1. Dans le cas où le meilleur score est obtenu par plusieurs contextes candidats, alors le classifieur retourne les différentes valeurs possibles.

$[\mathbf{\hat{a}g\acute{e}}, \mathbf{1}]$	$[\mathbf{\hat{a}g\acute{e}}, \mathbf{2}]$
$\langle(G_1G_3)\rangle$	$\langle(G_4)(G_3)\rangle$
$\langle(G_5)(G_2)\rangle$	$\langle(G_5G_6)(G_2)\rangle$
$\langle(G_2)(G_4)(G_3)\rangle$	$\langle(G_3)(G_2)\rangle$

TAB. 2 – *Les motifs associés aux contextes* $[\hat{a}g\acute{e}, 1]$ *et* $[\hat{a}g\acute{e}, 2]$.

- $score_{[\hat{a}g\acute{e},1]}(S) = 1$
- $score_{[\hat{a}g\acute{e},2]}(S) = 3$

Par conséquent, le contexte retenu est $[\hat{a}g\acute{e}, 2]$ *car il s'agit du meilleur score, et le grade estimé de la puce à ADN est* 2*.*

4 Résultats expérimentaux

4.1 Description des données

Les données exploitées pour évaluer l'approche présentée proviennent de plusieurs enregistrements issus du NCIB [2] (National Center for Biotechnology Information). Un tri des enregistrements a été nécessaire afin de sélectionner un nombre suffisant de puces adéquates avec notre approche (i.e., qui contiennent toutes les informations contextuelles souhaitées). Le jeu de données utilisé pour les expérimentations est constitué de 649 puces à ADN, chacune étant associée à cinq dimensions contextuelles.

Gènes considérés. Les puces à ADN fournissent l'expression de milliers de gènes pour chacun des patients. Or, la plupart de ces gènes ne sont pas connus pour avoir une implication dans le cancer du sein. Nous nous sommes donc appuyés sur les 128 gènes identifiés par Sotiriou et al. (2006) pour leur implication dans cette maladie.

Dimensions contextuelles utilisées.

- **Grade** : le grade d'une tumeur constitue une des méthodes de diagnostic les plus anciennes et les plus utilisées. Elle permet d'établir trois grades de malignité grâce à l'évaluation de trois critères que sont la différenciation, le degré d'anisocaryose, ainsi que le nombre de mitoses. Valeurs : *grade 1, 2 ou 3*.
- **Age** : l'âge du patient lors du diagnostic. Valeurs : *50- (moins de 50 ans), 50-62 (entre 50 et 62 ans), 62+ (plus de 62 ans)*.
- **Taille** : la taille de la tumeur (en cm). Valeurs : *1.85- (moins de 1.85 cm), 1.85-2.45 (entre 1.85 et 2.45 cm), 2.45+ (plus de 2.45 cm)*
- **ER** : Récepteur des œstrogènes. Certaines tumeurs sont à récepteurs d'œstrogènes positifs (valeur *1*), c'est à dire que l'exposition des cellules cancéreuses à cette hormone va favoriser leur croissance. Au contraire, chez certaines personnes, le taux de croissance de la tumeur n'est pas affecté par cette exposition. Ces tumeurs sont à récepteurs d'œstrogènes négatifs (valeur *0*).

2. `http://www.ncbi.nlm.nih.gov`

– **Node** : invasion des ganglions lymphatiques. Cette dimension nous donne sur l'état des ganglions lymphatique, elle sera positive (valeur *1*) en cas d'invasion métastatique de ceux-ci, négative (valeur *0*) sinon.

4.2 Résultats

Les expérimentations effectuées visent à répondre à deux questions principales :

1. Quel est l'apport de l'utilisation d'une dimension guide dans les résultats de classification ?
2. En fonction des caractéristiques d'un nouveau patient, quel guide choisir pour maximiser les résultats ?

Dans un premier temps, nous construisons et évaluons les classifieurs obtenus à partir de chaque dimension guide, mais également lorsqu'aucune dimension guide n'est utilisée.

La qualité des classifieurs est évaluée par le biais du rappel et de la précision pour chaque contexte. La *précision* est le nombre de séquences correctement affectées à un contexte, divisé par le nombre total de séquences affectées à ce contexte. Le *rappel* est le nombre de séquences correctement affectées à un contexte, divisé par le nombre total de séquences appartenant réellement à ce contexte. Ces deux mesures sont utilisées pour calculer une *F-mesure* combinant rappel et précision. La F-mesure dans un contexte c est définie de la manière suivante :

$$F = 2 \cdot \frac{précision \cdot rappel}{précision + rappel}.$$

Guide		**Rappel**	**Précision**	**F-mesure**
sans guide		0.66	0.65	0.66
Age	50-	0.78	0.77	0.78
	50-62	0.80	0.78	0.79
	62+	0.81	0.80	0.80
Taille	1.85-	0,76	0,71	0.73
	1.85-2.45	0.81	0.80	0.81
	2.45+	0.76	0.73	0.74
ER	0	-	-	-
	1	0.71	0.70	0.70
Node	0	0.72	0.68	0.70
	1	0.89	0.91	0.90

TAB. 3 – *Tableau récapitulatif des performances des classifieurs construits (pour* $minSup = 0.8$ *et* $k = 5000$*)*

Le tableau 3 présente de manière résumée les performances obtenues pour chaque guide possible. Par exemple, le classifieur construit pour prédire le grade de cancer pour les personnes âgées de plus de 62 ans obtient une F-mesure de 0.8. Notons qu'aucun classifieur n'a été construit pour la valeur 0 sur la dimension ER. En effet, le contexte correspondant à cette valeur et au grade 1 ne contenant que 5 puces à ADN, il est impossible d'en extraire les motifs

séquentiels, et par conséquent de construire le classifieur associé. La première ligne montre les résultats obtenus sans utiliser de guide, i.e., sans exploiter les informations contextuelles. La F-mesure obtenue est alors 0.66. Nous constatons que les résultats sont toujours meilleurs lorsqu'un guide est utilisé, peu importe celui-ci. Ainsi, l'exploitation des informations contextuelles associées aux puces apporte une véritable amélioration des résultats.

De plus, ces résultats permettent de mettre en évidence les guides les plus avantageux pour la classification. Considérons l'exemple d'une puce associée aux informations contextuelles suivantes :

- Age : 53 ans,
- Taille de la tumeur : 2.5 cm,
- ER : 1,
- Node : 1.

Chacune de ces informations peut être utilisée comme guide. L'expert peut par conséquent choisir parmi les classifieurs correspondants celui qui sera le plus apte à classer correctement la puce. Dans ce cas, la dimension *Node* est celle qui lui offrira la meilleure F-mesure (0.90).

5 Conclusion

Dans cet article, nous avons présenté une approche de classification des puces à ADN reposant sur l'extraction de motifs séquentiels dans les puces à ADN ainsi que sur la prise en compte d'informations contextuelles liées aux patients. Nous avons montré l'intérêt de telles informations pour améliorer les performances de la classification afin de proposer des outils d'aide au diagnostic plus fiables. En exploitant ces connaissances, l'expert est en mesure de choisir le classifieur qui maximisera les chances de réussite pour le diagnostic.

Cependant, l'approche proposée ne permet de prendre en compte qu'une seule dimension contextuelle à la fois. Parmi les perspectives ouvertes par ces travaux, la première consiste à développer une méthode permettant de prendre en compte toutes les informations contextuelles afin de maximiser les performances en classification. De plus, le processus de choix d'un classifieur en fonction des informations contextuelles connues pourrait être automatisé (par exemple sous la forme d'un arbre de décision). Enfin, la méthode définie pourrait être exploitée pour d'autres maladies que le cancer du sein.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2).

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. S. P. Chen (Eds.), *Eleventh International Conference on Data Engineering*. IEEE Computer Society Press.

Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, et A. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America 96*(12), 6745.

Dougherty, E. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics 2*(1), 28–34.

Jaillet, S., A. Laurent, et M. Teisseire (2006). Sequential Patterns for Text Categorization. *International Journal of Intelligent Data Analysis (IDA) 10*(3).

Rabatel, J., S. Bringay, et P. Poncelet (2010). Contextual Sequential Pattern Mining. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*. IEEE Computer Society.

Rosenwald, A., G. Wright, W. Chan, J. Connors, E. Campo, R. Fisher, R. Gascoyne, H. Muller-Hermelink, E. Smeland, J. Giltnane, et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine 346*(25), 1937.

Salle, P., S. Bringay, et M. Teisseire (2009). Mining Discriminant Sequential Patterns for Aging Brain. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine : Artificial Intelligence in Medicine*, pp. 365–369. Springer-Verlag.

Simon, R. et K. Dobbin (2003). Experimental design of DNA microarray experiments. *Biotechniques 34*(Suppl 1), 16–21.

Sotiriou, C., S. Neo, L. McShane, E. Korn, P. Long, A. Jazaeri, P. Martiat, S. Fox, A. Harris, et E. Liu (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America 100*(18), 10393.

Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, et al. (2006). Gene expression profiling in breast cancer : understanding the molecular basis of histologic grade to improve prognosis. *JNCI Cancer Spectrum 98*(4), 262.

Van De Vijver, M., Y. He, L. van't Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton, et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine 347*(25), 1999.

Zupan, B., J. Demsar, M. Kattan, J. Beck, et I. Bratko (2000). Machine learning for survival analysis : a case study on recurrence of prostate cancer. *Artificial intelligence in medicine 20*(1), 59–75.

Summary

Breast cancer is a major health problem and a challenge for biologists and health professionals. DNA microarrays now provide new tools to study the problems associated with this disease. In this paper, we propose to process data from DNA microarrays by discovering contextual sequential patterns (gene sequences ordered according to their expression level associated with a context). The goal is to provide aid in the diagnosis of a tumor grade. Our approach takes into account both the information contained in DNA microarrays (expressed through sequential patterns) but also of additional contextual information (e.g., patient age, tumor size, etc.) generally being associated with microarray data. The proposed approach has been evaluated on real data.