Extraction de connaissances d'adaptation par analyse de la base de cas

Fadi Badra, Jean Lieber, Amedeo Napoli

Extraction de connaissances d'adaptation par analyse de la base de cas

Fadi Badra*, Jean Lieber*, Amedeo Napoli*

*LORIA (UMR 7503 CNRS–INPL–INRIA-Nancy 2–UHP)
BP 239, 54506 Vandœuvre-lès-Nancy, FRANCE
{badra,lieber,napoli}@loria.fr

Résumé. En raisonnement à partir de cas, l'adaptation d'un cas source pour résoudre un problème cible est une étape à la fois cruciale et difficile à réaliser. Une des raisons de cette difficulté tient au fait que les connaissances d'adaptation sont généralement dépendantes du domaine d'application. C'est ce qui motive la recherche sur l'acquisition de connaissances d'adaptation (ACA). Cet article propose une approche originale de l'ACA fondée sur des techniques d'extraction de connaissances dans des bases de données (ECBD). Nous présentons CABAMAKA, une application qui réalise l'ACA par analyse de la base de cas, en utilisant comme technique d'apprentissage l'extraction de motifs fermés fréquents. L'ensemble du processus d'extraction des connaissances est détaillé, puis nous examinons comment organiser les résultats obtenus de façon à faciliter la validation des connaissances extraites par l'analyste.

1 Introduction

Raisonner à partir de cas consiste à résoudre un problème à l'aide d'une base de cas, dans laquelle un cas représente un problème déjà résolu accompagné de sa solution (Riesbeck et Schank (1989)). Un système de raisonnement à partir de cas (RÀPC) sélectionne un cas dans la base de cas, puis adapte la solution associée. L'adaptation nécessite des connaissances spécifiques au domaine d'application. L'acquisition de connaissances d'adaptation a pour but d'extraire ces connaissances, ce qui peut être réalisé soit directement auprès d'un expert du domaine (d'Aquin et al. (2006)), ou encore par analyse de la base de cas (voir par exemple Hanney et Keane (1996), McSherry (1998), Craw et al. (2006)).

Un cas est généralement représenté par un couple $(\texttt{pb}, \texttt{Sol}(\texttt{pb}))$ dans lequel pb représente un énoncé de problème et $\texttt{Sol}(\texttt{pb})$ une solution de pb. L'ensemble des *cas sources* $(\texttt{srce}, \texttt{Sol}(\texttt{srce}))$ d'un système de RÀPC constitue la *base de cas* BC. Lors d'une session particulière de RÀPC, le problème à résoudre est appelé *problème cible*, dénoté par cible. Une inférence à partir de cas associe à cible une solution $\texttt{Sol}(\texttt{cible})$, compte tenu de la base de cas BC et de bases de connaissances additionnelles, en particulier $\mathcal{O}$, l'*ontologie du domaine*, qui introduit les concepts et les termes utilisés pour représenter les cas.

Le processus de RÀPC est principalement composé d'une étape de remémoration et d'une étape d'adaptation. La *remémoration* sélectionne $(\texttt{srce}, \texttt{Sol}(\texttt{srce})) \in \texttt{BC}$ tel que srce est jugé similaire à cible. Le but de l'étape d'*adaptation* est ensuite de résoudre cible en

modifiant Sol(srce) de façon adéquate. Un *problème d'adaptation* est donné par un triplet (srce, Sol(srce), cible), et une solution d'un problème d'adaptation est une solution Sol(cible) du problème cible. Une étape de mémorisation d'un cas peut venir compléter le processus.

Le modèle d'adaptation adopté est une forme d'*analogie transformationnelle* (Carbonell (1983)) :

1. (srce, cible) $\mapsto$ Δpb, où Δpb encode les similarités et dissimilarités entre des problèmes srce et cible.
2. (Δpb, CA) $\mapsto$ Δsol, où CA est un ensemble de connaissances d'adaptation et Δsol encode les similarités et dissimilarités entre Sol(srce) et la solution Sol(cible) à construire pour cible.
3. (Sol(srce), Δsol) $\mapsto$ Sol(cible), Sol(srce) est modifié en Sol(cible) selon Δsol.

L'étape d'adaptation est dépendante du domaine d'application car elle nécessite des connaissances spécifiques au domaine. Ces connaissances doivent être acquises[1]. C'est l'objet de l'*acquisition de connaissances d'adaptation* (ACA).

Dans la section suivante nous rappelons les différentes étapes du processus d'ECBD et détaillons la façon dont celles-ci sont effectuées dans notre système CABAMAKA. Puis, dans la section 3, nous nous intéressons à la définition d'indices de qualité pour classer les règles d'adaptation obtenues. Enfin, dans la section 4, nous montrons comment le système peut être amélioré pour extraire des dépendances qualitatives entre variables.

2 CABAMAKA

CABAMAKA (acronyme de *case base mining for adaptation knowledge acquisition*) reprend les idées principales présentées dans Hanney et Keane (1996). Dans ces travaux, les variations entre cas sources sont exploités pour apprendre des règles d'adaptation. Tous les couples (cas-source$_i$, cas-source$_j$) de cas sources similaires dans la base de cas sont formés. Puis, pour chacun de ces couples, les variations entre problèmes srce$_i$ et srce$_j$ et solutions Sol(srce$_i$) et Sol(srce$_j$) sont représentés (Δpb et Δsol). Des heuristiques sont ensuite mises en œuvre pour regrouper ces règles d'adaptation et sélectionner la règle à appliquer lors d'une session de RÀPC. Il a été montré expérimentalement que l'utilisation de telles connaissances d'adaptation augmente la performance du système.

CABAMAKA se distingue néanmoins des ces travaux sur plusieurs points :

– Les connaissances d'adaptation obtenues doivent être validées par un expert et des explications doivent y être associées pour qu'elles soient compréhensibles par l'utilisateur. En ce sens, CABAMAKA peut être considéré comme un système d'apprentissage semi-automatique.
– Tous les couples de cas sources distincts de la base de cas sont pris en compte, pas seulement les couples de cas similaires. En conséquence, si n est la taille de la base de cas ($n = |\mathtt{BC}|$) le volume de cas examinés s'élève à $n(n-1)$. Dans notre application, $n \simeq 650$, ce qui amène à examiner un assez grand nombre de couples ($n(n-1) \simeq 5 \cdot 10^5$).

C'est pourquoi des techniques efficaces d'extraction de connaissances (Dunham (2003)) ont été choisies pour ce système.

[1] Dans cet article, nous utilisons le terme acquisition de connaissances au sens large qui inclut l'extraction de connaissances.

2.1 Principes

2.1.1 Principes de l'ECBD

Le but de l'ECBD est d'obtenir des connaissances à partir de données. Le processus d'ECBD se fait sous la supervision d'un analyste, qui est un expert du domaine. Une fois l'acquisition des données réalisée, il se déroule en trois étapes : la préparation des données, la fouille de données et la validation des connaissances extraites.

La préparation des données est une étape de mise en forme et de sélection des données. L'opération de mise en forme met les données dans un format acceptable pour l'algorithme de fouille choisi. La sélection des données permet de concentrer la fouille sur un sous-ensemble pertinent d'objets et/ou d'attributs, et d'éliminer les données bruitées.

La fouille de données extrait des éléments d'information à partir des données. Par exemple, CHARM (Zaki et Hsiao (2002)) est un algorithme de fouille de données qui réalise efficacement l'extraction de *motifs fermés fréquents* (*MFF*). CHARM prend en entrée un ensemble d'objets, chaque *objet* x étant un ensemble de propriétés booléennes. Un *motif* m est un ensemble de propriétés et son *extension* est l'ensemble des objets qui le contiennent. Le support de m, *Supp* (m), est la proportion d'objets x contenant m ($m \subseteq x$). Autrement dit, pour une variable aléatoire X parcourant l'ensemble des objets avec une distribution uniforme de probabilités, on a *Supp* $(m) = P(m \subseteq X)$. m est dit fréquent si son support dépasse un seuil donné $\sigma \in [0;1]$, c.-à-d. si *Supp* $(m) \geqslant \sigma$. Un motif est dit fermé s'il n'existe pas de motif m' contenant strictement m et de même support.

La validation des connaissances extraites se fait avec l'aide de l'analyste, qui interprète les résultats. Cette étape d'interprétation produit des unités de connaissances.

2.1.2 Préparation des données

L'étape de préparation des données génère un ensemble d'objets à partir de la base de cas `BC`, en appliquant successivement deux transformations.

La première transformation Φ formate chaque cas source (`srce`, `Sol(srce)`) en deux ensembles de propriétés booléennes : Φ(`srce`) et Φ(`Sol(srce)`). L'implantation de cette transformation dépend beaucoup du formalisme utilisé pour représenter les cas. Cette transformation entraîne en général une perte d'information, qui doit être minimisée. Le vocabulaire utilisé pour décrire les cas étant celui de l'ontologie du domaine $\mathcal{O}$, si $\Phi(\texttt{srce}) = \{p_1, \ldots, p_n\}$, on ajoute à Φ(`srce`) toute propriété q qui peut se déduire de l'ensemble $\{p_1, \ldots, p_n\}$ en fonction de l'ontologie $\mathcal{O}$.

La deuxième transformation produit un objet à partir de chaque couple de cas sources ($\Phi(\texttt{cas-source}_1), \Phi(\texttt{cas-source}_2)$). Suivant le modèle d'adaptation présenté en introduction, x doit encoder les propriétés de Δ`pb` et de Δ`sol`. Δ`pb` encode les similarités et dissimilarités de $\texttt{srce}_1$ et $\texttt{srce}_2$, c.-à-d. :

- Les propriétés communes à $\texttt{srce}_1$ et $\texttt{srce}_2$ (marquées par "="),
- Les propriétés de $\texttt{srce}_1$ que $\texttt{srce}_2$ ne partage pas ("-") et
- Les propriétés de $\texttt{srce}_2$ que $\texttt{srce}_1$ ne partage pas ("+").

Toutes ces propriétés sont reliées à des problèmes et sont marquées par pb. Δsol est calculé de façon similaire et $x = \Delta\text{pb} \cup \Delta\text{sol}$. Par exemple,

$$\text{si} \quad \begin{cases} \Phi(\texttt{srce}_1) = \{a, b, c\} & \Phi(\texttt{Sol}(\texttt{srce}_1)) = \{A, B\} \\ \Phi(\texttt{srce}_2) = \{b, c, d\} & \Phi(\texttt{Sol}(\texttt{srce}_2)) = \{B, C\} \end{cases}$$
$$\text{alors} \quad x = \left\{a^{-}_{\texttt{pb}}, b^{=}_{\texttt{pb}}, c^{=}_{\texttt{pb}}, d^{+}_{\texttt{pb}}, A^{-}_{\texttt{sol}}, B^{=}_{\texttt{sol}}, C^{+}_{\texttt{sol}}\right\} \tag{1}$$

2.1.3 Fouille de données

L'extraction de *MFF* est réalisée par l'application de l'algorithme CHARM sur l'ensemble des objets. Un motif extrait peut être considéré comme une généralisation d'un ensemble d'objets. Par exemple, si $m_{ex} = \left\{a^{-}_{\texttt{pb}}, c^{=}_{\texttt{pb}}, d^{+}_{\texttt{pb}}, A^{-}_{\texttt{sol}}, B^{=}_{\texttt{sol}}, C^{+}_{\texttt{sol}}\right\}$ est un *MFF*, m_{ex} est une généralisation du sous-ensemble d'objets incluant l'objet x. Un objet x encode une règle d'adaptation spécifique $(\texttt{srce}, \texttt{Sol}(\texttt{srce}), \texttt{cible}) \mapsto \texttt{Sol}(\texttt{cible})$. Un motif correspond donc à une règle d'adaptation plus générale.

2.1.4 Interprétation

L'étape d'interprétation est supervisée par un analyste. Le système CABAMAKA fournit à l'analyste les *MFF* extraits et lui permet de naviguer parmi eux. L'analyste peut sélectionner un *MFF*, l'interpréter en règle d'adaptation, puis valider, corriger, voire généraliser la règle.

Chaque motif obtenu par CABAMAKA à la suite de l'étape de fouille peut se lire comme une règle d'adaptation qui exprime une relation entre :

- La présence ou non de certaines propriétés booléennes dans $\Phi(\texttt{srce})$, $\Phi(\texttt{cible})$ et $\Phi(\texttt{Sol}(\texttt{srce}))$,
- La présence ou non de certaines propriétés booléennes dans $\Phi(\texttt{Sol}(\texttt{cible}))$.

Par exemple, le motif m_{ex} correspond à une règle d'adaptation qui peut se lire de la manière suivante :

si *a* est une propriété de srce mais n'est pas une propriété de cible,
c est une propriété à la fois de srce et cible,
d n'est pas une propriété de srce mais est une propriété de cible,
A et *B* sont des propriétés de Sol(srce) et
C n'est pas une propriété de Sol(srce)
alors les propriétés de Sol(cible) sont
$\Phi(\texttt{Sol}(\texttt{cible})) = (\Phi(\texttt{Sol}(\texttt{srce})) \setminus \{A\}) \cup \{C\}$.

2.2 Application

Le domaine d'application pour lequel cette étude a été réalisée est celui du traitement du cancer du sein. Dans cette application, un problème décrit une classe de patients par un ensemble d'attributs (comme l'âge ou la taille de la tumeur) et de contraintes sur les valeurs prises par ces attributs. Une solution est un ensemble de traitements (radiothérapie, chimiothérapie, etc.) recommandés pour ces patients.

2.3 Résultats

L'application de CABAMAKA au domaine du traitement du cancer du sein a permis d'extraire un ensemble de motifs dont le motif :

$$m = \{(\ \texttt{ge} < 70)^{=}_{\texttt{pb}}, (\texttt{taille-tumeur} \leqslant 4)^{-}_{\texttt{pb}}, (\texttt{taille-tumeur} > 4)^{+}_{\texttt{pb}},$$
$$\texttt{Curage}^{=}_{\texttt{sol}}, \texttt{Mastectomie}^{=}_{\texttt{sol}},$$
$$\texttt{Mastectomie-partielle}^{-}_{\texttt{sol}}, \texttt{Mastectomie-totale}^{+}_{\texttt{sol}}\}$$

Ce motif peut être interprété ainsi : si `srce` et `cible` représentent tous deux des classes de patients de moins de 70 ans, si la différence entre `srce` et `cible` réside dans la taille de la tumeur — moins de 4 cm pour `srce` et plus de 4 cm pour `cible` — et si une mastectomie partielle avec curage axillaire est proposée pour `srce`, alors `Sol(cible)` est obtenue en remplaçant dans `Sol(srce)` la mastectomie partielle par une mastectomie totale.

Cette règle traduit le fait que le type d'intervention chirurgicale proposé dépend de la taille de la tumeur du patient : plus la taille de la tumeur est grande, plus on augmente le geste chirurgical.

L'obtention de telles règles d'adaptation nécessite pour l'instant l'intervention d'un ingénieur de la connaissance qui sélectionne les motifs intéressants parmi les résultats, les interprète comme des règles d'adaptation puis les présente à l'analyste pour validation. Deux obstacles subsistent à un pilotage par l'analyste du processus complet d'extraction de connaissances :

- Les règles obtenues ne sont pas formulées dans un format intelligible par l'analyste : l'analyste n'est pas capable d'interpréter un motif. Pour devenir compréhensibles les règles obtenues doivent être exprimées en langue naturelle.
- Les règles obtenues sont trop nombreuses pour être toutes présentées à l'analyste pour validation. La figure 1 présente les résultats expérimentaux — temps d'exécution de CHARM, implanté dans la plateforme CORON (Szathmary et Napoli (2005)), et nombre de motifs — obtenus pour l'étape de fouille de données sur une base de cas test de 59 cas. Les tests ont été effectués sur une machine ayant un processeur cadencé à 3,2GHz et disposant de 2Go de mémoire vive. On constate que le nombre de règles d'adaptation obtenues devient vite très conséquent lorsqu'on choisit de garder les règles de faible support.

support	temps d'exécution de CHARM (s)	nombre de motifs
10%	2	591
5%	3	3057
1%	44	49651
0%	345	189690

FIG. 1 – *Résultats de l'étape de fouille de données pour une base de cas test de 59 cas.*

A cause du grand nombre de motifs que l'ingénieur de la connaissance doit examiner avant de pouvoir en proposer à l'analyste pour validation, la mise en œuvre expérimentale pour l'évaluation du système prend également du temps. Il est donc nécessaire de doter l'analyste de

moyens de naviguer dans l'ensemble des règles obtenues et de sélectionner les règles intéressantes.

3 Vers la définition d'indices de qualité pour les règles d'adaptation

Un moyen de sélectionner les règles intéressantes parmi l'ensemble des résultats est de les classer selon un indice de qualité. Cet indice pourra être adapté des indices utilisés pour les règles d'association.

On rappelle qu'une règle d'association est la donnée de deux motifs A et B disjoints et dénotée par $A \rightarrow B$. A est appelé l'antécédent de la règle et B le conséquent. Elle traduit, dans un ensemble d'objets, le fait que si le motif A est présent dans un objet, alors il est probable que le motif B soit également présent. Cette probabilité est appelée *confiance* de $A \rightarrow B$ et est la probabilité conditionnelle d'avoir le motif B, sachant qu'on a le motif A :

$$Conf\,(A \rightarrow B) = P(B \subseteq X \mid A \subseteq X) = \frac{Supp\,(A \cup B)}{Supp\,(A)}$$

(X est une variable aléatoire de distribution uniforme sur l'ensemble des objets).

À l'issue de l'étape de fouille de CABAMAKA, un motif m peut être décomposé en deux sous-motifs $\Delta\texttt{pb}$ et $\Delta\texttt{sol}$ et lu comme la règle d'association $\Delta\texttt{pb} \rightarrow \Delta\texttt{sol}$: $\Delta\texttt{pb}$ est l'ensemble des propriétés de m indicées par pb et $\Delta\texttt{sol}$ est $m \backslash \Delta\texttt{pb}$. Dans ce cas, la confiance de la règle $\Delta\texttt{pb} \rightarrow \Delta\texttt{sol}$ vaut :

$$Conf\,(\Delta\texttt{pb} \rightarrow \Delta\texttt{sol}) = P(\Delta\texttt{sol} \subseteq x \mid \Delta\texttt{pb} \subseteq x)$$

Néanmoins, la règle produite par une telle décomposition du motif m ne correspond pas à une règle d'adaptation. En effet, la règle $\Delta\texttt{pb} \rightarrow \Delta\texttt{sol}$ exprime ce que doit être le couple ($\texttt{Sol(srce)}$, $\texttt{Sol(cible)}$) étant donné un certain couple ($\texttt{srce}$, $\texttt{cible}$). Une règle d'adaptation exprime quant à elle ce que doit être $\texttt{Sol(cible)}$ étant donné un problème d'adaptation ($\texttt{srce}$, $\texttt{Sol(srce)}$, $\texttt{cible}$). Le même motif doit donc se lire :

$$(\Phi(\texttt{srce}), \Phi(\texttt{Sol(srce)}), \Phi(\texttt{cible})) \mapsto \Phi(\texttt{Sol(cible)})$$

Or une telle règle ne correspond pas à une décomposition d'un motif en deux sous-motifs, donc les indices définis pour les règles d'association ne s'appliquent pas directement.

Pour définir une mesure de confiance qui soit plus adaptée aux règles d'adaptation, il convient de considérer non seulement l'ensemble des objets constituant l'extension d'un motif, mais également les couples de cas sources que ces objets représentent. Prenons par exemple un objet faisant partie de l'extension d'un motif m et le couple ($\texttt{cas-source}_1$, $\texttt{cas-source}_2$) de cas sources représenté par cet objet. Si le motif contient une propriété $p^{=}_{pb}$, alors le couple de cas sources est tel que les deux problèmes sources partagent la propriété p, soit $p \in \Phi(\texttt{srce}_1) \cap \Phi(\texttt{srce}_2)$. Les propriétés constitutives du motif et leur marquage expriment ainsi un ensemble de conditions portant sur la présence ou non de certaines propriétés dans les ensembles $\Phi(\texttt{srce})$, $\Phi(\texttt{Sol(srce)})$, $\Phi(\texttt{cible})$ et $\Phi(\texttt{Sol(cible)})$ d'un couple de cas. La

	=	−	+
pb	$\Phi(\texttt{srce}_1) \cap \Phi(\texttt{srce}_2)$	$\Phi(\texttt{srce}_1) \backslash \Phi(\texttt{srce}_2)$	$\Phi(\texttt{srce}_2) \backslash \Phi(\texttt{srce}_1)$
sol	$\Phi(\texttt{Sol}(\texttt{srce}_1)) \cap \Phi(\texttt{Sol}(\texttt{srce}_2))$	$\Phi(\texttt{Sol}(\texttt{srce}_1)) \backslash \Phi(\texttt{Sol}(\texttt{srce}_2))$	$\Phi(\texttt{Sol}(\texttt{srce}_2)) \backslash \Phi(\texttt{Sol}(\texttt{srce}_1))$

FIG. 2 – *Appartenance des propriétés booléennes aux cas sources selon leur marquage dans un motif.*

figure 2 résume à quel ensemble de propriétés d'un couple de cas sources doit appartenir une propriété suivant la façon dont elle est marquée dans un motif.

Une règle d'adaptation décrit alors les propriétés que doit ou non contenir $\Phi(\texttt{Sol}(\texttt{cible}))$ compte tenu de la présence ou non de certaines propriétés dans $\Phi(\texttt{srce})$, $\Phi(\texttt{Sol}(\texttt{srce}))$ et $\Phi(\texttt{cible})$. La règle d'adaptation correspondant à un motif m peut être lue de la façon suivante :

si pour $p^{=}_{pb} \in m$, $p \in \Phi(\texttt{srce}) \cap \Phi(\texttt{cible})$,
pour $p^{-}_{pb} \in m$, $p \in \Phi(\texttt{srce}) \backslash \Phi(\texttt{cible})$,
pour $p^{+}_{pb} \in m$, $p \in \Phi(\texttt{cible}) \backslash \Phi(\texttt{srce})$,
pour $p^{=}_{sol} \in m$, $p \in \Phi(\texttt{Sol}(\texttt{srce}))$,
pour $p^{-}_{sol} \in m$, $p \in \Phi(\texttt{Sol}(\texttt{srce}))$,
pour $p^{+}_{sol} \in m$, $p \notin \Phi(\texttt{Sol}(\texttt{srce}))$,
alors les propriétés de $\texttt{Sol}(\texttt{cible})$ sont
$\Phi(\texttt{Sol}(\texttt{cible})) = (\Phi(\texttt{Sol}(\texttt{srce})) \setminus \{p \mid p^{-}_{sol} \in m\}) \cup \{p \mid p^{+}_{sol} \in m\}$.

On peut dès lors définir une mesure de confiance pour les règles d'adaptation, en considérant une variable aléatoire non plus sur l'univers des objets, mais sur l'univers des couples $(\Phi(\texttt{cas-source}_1), \Phi(\texttt{cas-source}_2))$ d'images par Φ de cas sources distincts (avec une distribution uniforme sur cet ensemble). La confiance d'une règle d'adaptation mesure alors la probabilité conditionnelle que le conséquent d'une règle d'adaptation soit vérifié, sachant que l'antécédent l'est :

$$Conf'(m) = P\left(\begin{array}{l|l} \text{pour } p^{-}_{sol} \in m, p \notin \Phi(\texttt{Sol}(\texttt{srce}_2)) & \text{pour } p^{=}_{pb} \in m, p \in \Phi(\texttt{srce}_1) \cap \Phi(\texttt{srce}_2) \\ \text{pour } p^{+}_{sol} \in m, p \in \Phi(\texttt{Sol}(\texttt{srce}_2)) & \text{pour } p^{-}_{pb} \in m, p \in \Phi(\texttt{srce}_1) \backslash \Phi(\texttt{srce}_2) \\ \text{pour } p^{=}_{sol} \in m, p \in \Phi(\texttt{Sol}(\texttt{srce}_2)) & \text{pour } p^{+}_{pb} \in m, p \in \Phi(\texttt{srce}_2) \backslash \Phi(\texttt{srce}_1) \\ & \text{pour } p^{=}_{sol} \in m, p \in \Phi(\texttt{Sol}(\texttt{srce}_1)) \\ & \text{pour } p^{=}_{sol} \in m, p \in \Phi(\texttt{Sol}(\texttt{srce}_1)) \\ & \text{pour } p^{=}_{sol} \in m, p \notin \Phi(\texttt{Sol}(\texttt{srce}_1)) \end{array}\right)$$

Le classement des règles selon cet indice n'a pour l'instant pas été mis en place. Des travaux sont en cours pour tenter de ramener le calcul de ce nouvel indice à des calculs de supports, ce qui faciliterait son opérationnalisation.

En dehors de la confiance, d'autres indices de qualité d'une règle d'association ont été définis (voir Lallich et al. (2006) ou Vaillant et al. (2005) pour une présentation des principaux indices et de méthodes d'évaluation formelles et expérimentales de ces indices). De la même façon que nous avons défini un indice de confiance pour les règles d'adaptation en nous appuyant sur l'indice de confiance des règles d'association, il devrait être possible de définir,

en s'appuyant sur ces indices, d'autres indices d'une règle d'adaptation. Ceci constitue une perspective de recherches. Dans le même ordre d'idée, les travaux en analyse statistique implicative (voir, p.ex. Gras et al. (2001)), pourraient être réutilisées pour définir des indices de qualité des règles d'adaptation.

4 Exhiber des dépendances qualitatives entre variables

Lors de l'étape de préparation des données, les cas sources sont décrits par des ensembles de propriétés booléennes et un objet encode les variations de présence et d'absence de ces propriétés lorsqu'on passe d'un cas source à un autre. Lorsque ces propriétés booléennes représentent différentes modalités d'une même variable, comme par exemple l'âge, la taille ou le sexe, il peut être intéressant d'encoder également dans les objets les variations de modalité subies par ces variables. Les motifs obtenus expriment alors des dépendances qualitatives entre variables.

Prenons par exemple les deux propriétés booléennes `ge = 30` et `ge = 45`, qui correspondent à deux modalités de la variable âge, l'une caractérisant les problèmes pour lesquels la valeur de l'âge est 30 ans et l'autre les problèmes pour lesquels la valeur de l'âge est 45 ans. Soient deux cas sources $\texttt{cas-source}_1$ et $\texttt{cas-source}_2$, tels qu'à l'issue de la première transformation $\Phi(\texttt{srce}_1)$ contient la propriété `ge = 30` et $\Phi(\texttt{srce}_2)$ contient la propriété `ge = 45`. L'objet produit lors de la deuxième transformation pour ces deux cas sources contient donc le motif $\{(\texttt{ge = 30})^{-}_{\texttt{pb}}\ (\texttt{ge = 45})^{+}_{\texttt{pb}}\}$. Ce motif correspond à une augmentation de la valeur de la variable âge lorsqu'on passe du cas source $\texttt{cas-source}_1$ au cas source $\texttt{cas-source}_2$. Pour encoder également dans cet objet la variation qualitative que subit la variable âge, on peut enrichir l'objet d'une nouvelle propriété `ge:varie`, qui représente une variation de la variable âge, et d'une propriété `ge:augmente`, qui représente son sens de variation.

Si le même objet contient la propriété `chimioth rapie:varie`, qui représente une variation de dose prescrite pour la chimiothérapie, et la propriété `chimioth rapie:diminue`, qui représente une diminution de cette dose, un motif obtenu à l'issu de l'étape de fouille peut être le motif { `ge:varie`, `chimioth rapie:varie`}. Ce motif exprime une dépendance fonctionnelle entre les variables `ge` et `chimioth rapie`. De la même façon, le motif plus spécifique { `ge:augmente`, `chimioth rapie:diminue`} peut être obtenu. Ce motif exprime la dépendance qualitative $\texttt{ge} \xrightarrow{-} \texttt{chimioth rapie}$, selon laquelle la dose de chimiothérapie diminue avec l'âge du patient[2].

De telles règles abstraites ont l'avantage d'être plus facilement interprétables par l'analyste que les règles d'adaptation. De plus, chacune d'elle étant partagée par plusieurs motifs, elles constituent un moyen efficace de les regrouper hiérarchiquement.

[2]Cette connaissance qualitative peut être associée à deux explications complémentaires. D'une part, avec l'âge, la virulence du cancer diminue et une dose de chimiothérapie moins importante est nécessaire. D'autre part, l'espérance de vie des personnes âgées étant moindre, le gain en bénéfice thérapeutique d'une chimiothérapie à dose élevée est faible vis-à-vis de l'augmentation des effets indésirables de ce traitement.

5 Conclusion

Dans cet article, nous avons présenté CABAMAKA, un système qui met en œuvre une technique d'ECBD, l'extraction de motifs fermés fréquents, pour extraire des connaissances d'adaptation à partir des variations qui existent au sein d'une base de cas. Ce système est assez unique en son genre car il extrait des connaissances à partir de connaissances.

Nous avons montré en quoi les indices de qualité existant pour les règles d'association sont inadaptés pour mesurer la qualité d'une règle d'adaptation obtenue par ce système et comment l'on peut s'en inspirer pour créer des indices plus appropriés. Des travaux sont actuellement en cours sur la définition de tels indices.

Nous avons également proposé une amélioration du système de façon à découvrir des règles plus abstraites qui expriment des dépendances qualitatives entre les variables entrant en jeu dans la description des cas. La validation de telles dépendances doit être plus facile car ces règles sont moins nombreuses et plus intelligibles par l'analyste. Cette méthode est actuellement en cours d'implantation.

Par ailleurs, les connaissances d'adaptation obtenues par CABAMAKA ont vocation à venir alimenter un portail sémantique (d'Aquin (2005)) développé dans le cadre du projet KASIMIR (Lieber et al. (2002)), dont l'objet est la gestion de connaissances et l'aide à la décision en cancérologie. En particulier, les travaux actuels portent sur l'intégration de ces connaissances au moteur de raisonnement à partir de cas de KASIMIR.

Remerciements

Les auteurs remercient les relecteurs de cet article pour leurs remarques et suggestions.

Références

Carbonell, J. (1983). Learning by analogy : Formulating and generalizing plans from past experience. In R. Michalski, J. Carbonell, et T. Mitchell (Eds.), *Machine Learning : An Artificial Intelligence Approach*, pp. 137–162. Cambridge, MA : Tioga.

Craw, S., N. Wiratunga, et R. C. Rowe (2006). Learning adaptation knowledge to improve case-based reasoning. *Artificial Intelligence 170*, 1175–1192.

d'Aquin, M. (2005). *Un portail sémantique pour la gestion des connaissances en cancérologie*. Ph. D. thesis, Université Henri Poincaré - Nancy 1.

d'Aquin, M., J. Lieber, et A. Napoli (2006). Adaptation knowledge acquisition : A case study for decision support in oncology. (à paraître dans le journal Computational Intelligence).

Dunham, M. H. (2003). *Data Mining : Introductory and Advanced Topics*. Prentice-Hall.

Gras, R., P. Kuntz, et H. Briant (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et sciences humaines* (154), 9–29.

Hanney, K. et M. Keane (1996). Learning Adaptation Rules from Cases. In I. Smith et B. Falting (Eds.), *Proc. of the 3rd European Workshop on Case-Based Reasoning, EWCBR-96*, Volume 1168 of *LNAI*. Springer.

Lallich, S., O. Teytaud, et E. Prudhomme (2006). *Association rules interestingness : measure and validation*. Quality Measures in Data Mining. Heidelberg, Germany : Springer.

Lieber, J., M. d'Aquin, P. Bey, B. Bresson, O. Croissant, P. Falzon, A. Lesur, J. Lévêque, V. Mollo, A. Napoli, M. Rios, et C. Sauvagnac (2002). The kasimir project : knowledge management in cancerology. In *Proceedings of the Fourth International Workshop on Enterprise Networking and Computing in Health Care Industry, HealthCom 2002*, pp. 125–12.

McSherry, D. (1998). An adaptation heuristic for case-based estimation. In *EWCBR '98 : Proceedings of the 4th European Workshop on Advances in Case-Based Reasoning*, London, UK, pp. 184–195. Springer-Verlag.

Riesbeck, C. K. et R. C. Schank (1989). *Inside Case-Based Reasoning*. Mahwah, NJ, USA : Lawrence Erlbaum Associates, Inc.

Szathmary, L. et A. Napoli (2005). CORON : A Framework for Levelwise Itemset Mining Algorithms. In *Proc. of the Third International Conference on Formal Concept Analysis, ICFCA '05*, pp. 110–113.

Vaillant, B., P. Meyer, E. Prudhomme, S. Lallich, P. Lenca, et S. Bigaret (2005). Mesurer l'intérêt des règles d'association. In *Atelier Qualité des Données et des Connaissances (DQK 05), EGC 05, Paris*, pp. 69–78.

Zaki, M. J. et C.-J. Hsiao (2002). CHARM : An efficient algorithm for closed itemset mining. In R. L. Grossman, J. Han, V. Kumar, H. Mannila, et R. Motwani (Eds.), *Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, USA, April 11-13, 2002*. SIAM.

Summary

In case-based reasoning, the purpose of the adaptation step is to modify a retrieved source case in order to solve a target problem. This task appears to be at the same time crucial and difficult to implement, mainly because of the need for adaptation knowledge, which is usually domain-dependant and thus hard to acquire. This is what motivates the current research on adaptation knowledge acquisition (AKA). In this paper, we propose an original approach of AKA that makes use of knowledge discovery in databases (KDD) principles and techniques. This approach is implemented in CABAMAKA, a system that analyses the variations that exist within the case base in order to elicit adaptation knowledge units, using a frequent closed itemset extraction algorithm. The KDD process is described in details, and some ways of organising the results to facilitate the validation step are studied.