Chapitre 6

Implémentation et cas d'application aux données complexes

" D'impossibles à imaginaires, d'imaginaires à complexes. Combien d'idées, de systèmes politiques, de théories, de procédés ont suivi ce chemin pour devenir 'réalité'! "

Denis Guedj, "*Le théorème du perroquet*"

6.1 Introduction

Nous avons mis en œuvre nos propositions théoriques au sein d'une plateforme logicielle appelée MiningCubes. Il s'agit d'une application générale dédiée à l'analyse des données multidimensionnelles et à l'implémentation de nos travaux sur le couplage entre l'analyse en ligne et la fouille de données : la *réorganisation des cubes de données par l'analyse des correspondances multiples*, l'*agrégation par classification ascendante hiérarchique dans les cubes de données* et l'*explication dans les cubes de données par règles d'association.*

Afin d'assurer une interaction efficace avec les utilisateurs à travers une interface conviviale et afin d'impliquer ces derniers le plus possible dans le processus d'analyse, nous avons développé MiningCubes dans un environnement Web offrant ainsi une application ergonomique, facile à utiliser et adaptée au contexte de l'analyse en ligne (voir figure 6.1). Pour l'essentiel, MiningCubes est une application Web qui évolue selon une architecture Client/Serveur lui permettant de se connecter à des cubes de données sur des serveurs OLAP distants.

Avec l'avènement des nouvelles technologies de communication et plus précisément Internet, les entreprises recueillent des masses de données de plus en plus importantes.

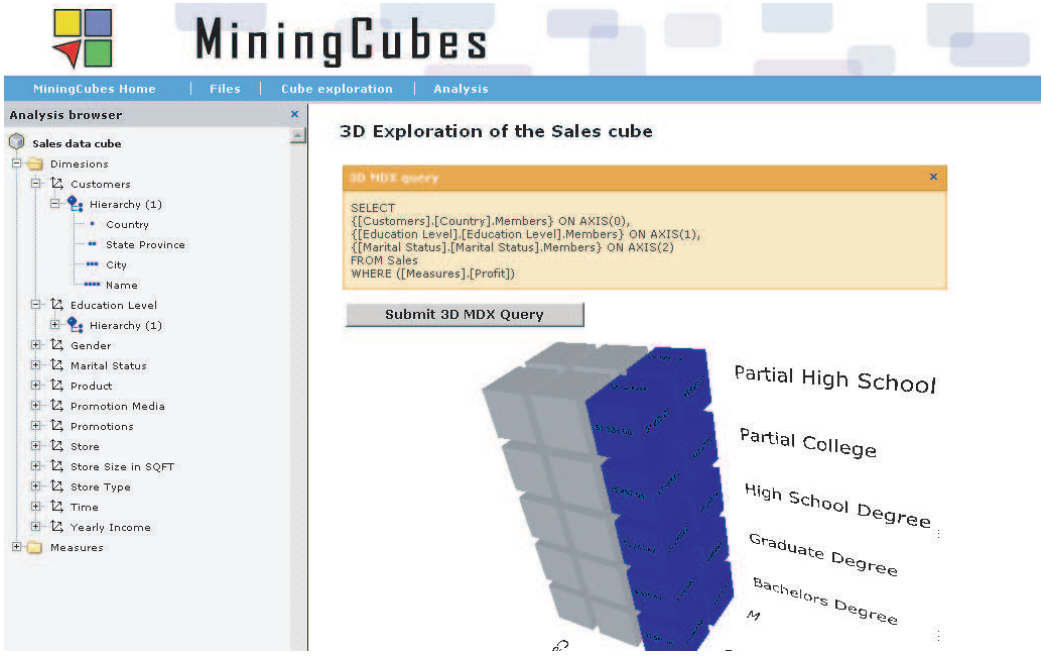


FIG. 6.1 – Environnement général de MiningCubes

Ces données étant généralement hétérogènes car elles proviennent de différentes sources. Elles sont dites complexes car elles sont de formats différents et sont sur des supports différents. Par exemple, dans le domaine médical, le dossier d'un patient contient des informations générales sur le patient (âge, sexe, poids, taille, etc.), ainsi que des images de scanner, des interrogatoires sous forme d'enregistrements sonores ou des compte-rendus manuscrits de médecins.

Pour exploiter de telles données à des fins décisionnelles, notamment dans le cadre de notre plateforme d'analyse MiningCubes, il est nécessaire de les structurer et de les homogénéiser. Le langage XML (*eXtensible Markup Language*) s'avère une solution appropriée à ce travail préparatoire sur les données complexes. XML est une norme de W3C[1] et est considéré comme un standard dans la description et l'échange des données. Il représente les données de façon semi-structurée. Sa capacité d'auto-description et sa structure arborescente donnent à ce formalisme une grande flexibilité et une puissance suffisante pour décrire des données complexes, hétérogènes et provenant de sources éparpillées. Les données sont alors stockées dans des documents XML formés conformément à une grammaire associée, exprimée sous forme de *DTD* (*Document Type Definition*) ou de *Schéma XML*.

[1] `http://www.w3.org/`

En plus du cas des cubes de données simples, tels que les cubes classiques des ventes par exemple, nous avons étendu notre plateforme logicielle aux cas des cubes de données complexes. Pour cela, nous avons défini une démarche méthodologique, appelée X-Warehousing, pour la modélisation et la matérialisation multidimensionnelle des données complexes en se basant sur le formalisme XML [BMCA06a, BMCA06b]. Dans cette démarche, XML est perçu comme un format efficace pour entreposer et préparer des contextes d'analyse multidimensionnelle des données complexes. Ainsi, X-Warehousing permet la création de cubes XML de données complexes que nous exploitons par la suite dans notre plateforme MiningCubes à des fins d'analyse. Par conséquent, avec ce nouveau format des cubes de données, nous sommes capables d'étendre nos propositions d'analyse basées sur le couplage entre l'OLAP et la fouille de données au cas des données complexes.

Dans ce chapitre, nous proposons un cas d'application de notre méthode d'agrégation par classification aux données complexes. Pour cela, nous partons d'un jeu de données médicales relatives au domaine du dépistage du cancer du sein. Dans ce jeu de données, nous disposons d'un ensemble de dossiers de patientes atteintes du cancer du sein où chaque dossier comprend des sources de données hétérogènes et éparpillées sur plusieurs types de supports. À partir de ces dossiers, nous avons défini, dans un premier temps, un contexte d'analyse. Dans un deuxième temps, nous avons employé la méthodologie X-Warehousing afin de concevoir et construire un cube XML de données de mammographies répondant au contexte de l'analyse définie. Enfin, ce cube XML a fait l'objet d'une analyse à l'aide de notre deuxième proposition de couplage entre l'OLAP et la fouille de données, à savoir l'agrégation par classification.

Ce chapitre est organisé de la façon suivante. Dans la section 6.2, nous présentons l'architecture générale de notre plateforme MiningCubes et détaillons ses différents modules, notamment ceux dédiés à nos propositions dans le cadre du couplage entre l'analyse en ligne et la fouille de données. Un exposé sur le jeu des données médicales, utilisées dans le cadre de ce cas d'application, est fourni dans la section 6.3. La section 6.4 présente une vue d'ensemble de notre méthodologie X-Warehousing d'entreposage des données complexes basée sur XML. Nous employons notre méthodologie d'entreposage en vue de concevoir et de construire un cube XML de données complexes de mammographies dans la section 6.5. Dans la section suivante, nous exposons une étude de cas d'agrégation par classification sur ce cube de données complexes. Une série d'expérimentations est fournie, dans la section 6.7, afin d'évaluer la performance de notre implémentation appliquée aux données complexes selon la solution XML. Enfin, la section 6.8 conclue le chapitre et propose des perspectives futures.

6.2 Plateforme MiningCubes

Pour valider nos approches, nous avons développé dans un environnement Web une plateforme générale, baptisée MiningCubes, dans laquelle nous implémentons l'ensemble de nos contributions sur le couplage entre l'analyse en ligne et la fouille de données. MiningCubes est implémentée en programmation Web dynamique avec le langage ASP (*Active Server Pages*). La programmation Web permet la mise en place d'un environnement convivial capable d'interagir en ligne avec un grand nombre d'utilisateurs. Elle permet également l'accès, local ou distant, à diverses sources de données (bases de données, serveurs OLAP, données XML, etc.). De plus, les applications Web sont indépendantes des environnements ou des systèmes d'exploitation utilisés.

Nous détaillons dans la suite l'architecture générale de notre plateforme MiningCubes et présentons ensuite les modules dédiés à nos propositions dans le cadre du couplage entre l'analyse en ligne et la fouille de données.

6.2.1 Architecture générale de la plateforme MiningCubes

La plateforme MiningCubes a une architecture de type Client/Serveur. Comme l'illustre la figure 6.2, elle est composée d'un ensemble de modules complémentaires. Nous distinguons trois types de modules : des *modules de connexion aux données*, des *modules de navigation dans les données multidimensionnelles* et des *modules d'aide à la décision*. La plateforme est également ouverte à l'ajout d'autres modules que nous pourrons intégrer dans la suite de nos travaux.

Modules de connexion aux données

Les modules de connexion aux données permettent d'ouvrir des sources de données à partir du serveur. Un premier module assure la connexion à des cubes de données dans *Analysis Services* de MicroSoft SQL Server 2000. Un second module permet de connecter la plateforme à des *cubes de données XML*. Nous exposerons plus en détail ce module dans la sous-section 6.6.2.

Modules de navigation dans les données

Les modules de navigation dans les données multidimensionnelles permettent, avec des outils visuels, d'explorer un cube de données en se basant sur les opérations OLAP classiques telles que le forage vers le haut (*roll-up*), le forage vers le bas (*drill-down*), la sélection (*slice*) et la projection (*dice*). Nous proposons deux approches de navigation. La première consiste en une navigation 2D selon des *vues bi-dimensionnelles* et la seconde consiste en une navigation 3D selon des vues *tri-dimensionnelles* des données

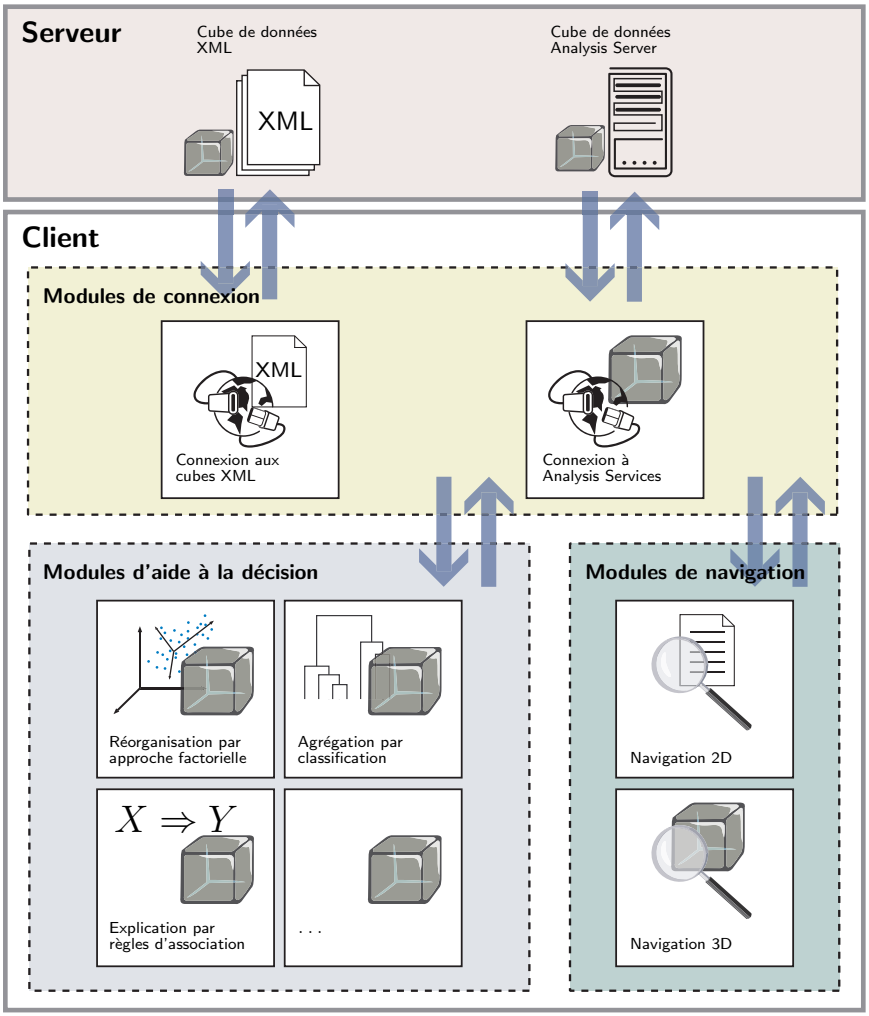


FIG. 6.2 – Architecture générale de MiningCubes

(voir figure 6.1).

Modules d'aide à la décision

Les modules d'aide à la décision comprennent les implémentations de nos méthodes de réorganisation par approche factorielle, d'agrégation par classification et d'explication par règles d'association. Chacune de nos méthodes adopte une approche de couplage entre l'analyse en ligne et la fouille de données. À cet effet, rappelons brièvement que :

1. la réorganisation des cubes de données par approche factorielle associe l'analyse en ligne et l'ACM en se basant sur une phase d'adaptation des données.

Cette phase consiste à transformer, selon un codage binaire, les données multidimensionnelles en un tableau disjonctif complet ;

2. l'agrégation par classification dans les cubes de données est une approche instrumentale qui utilise les opérations OLAP comme outil pour assurer l'association de l'OLAP avec la CAH. Les opérations OLAP sont exploitées pour extraire les données nécessaires pour la construction de la classification ;
3. l'explication dans les cubes de données par règles d'association adapte l'algorithme d'extraction des règles d'association aux données multidimensionnelles. L'algorithme est alors capable d'interroger un cube de données et d'en extraire directement des règles d'association.

6.2.2 Module de réorganisation par approche factorielle

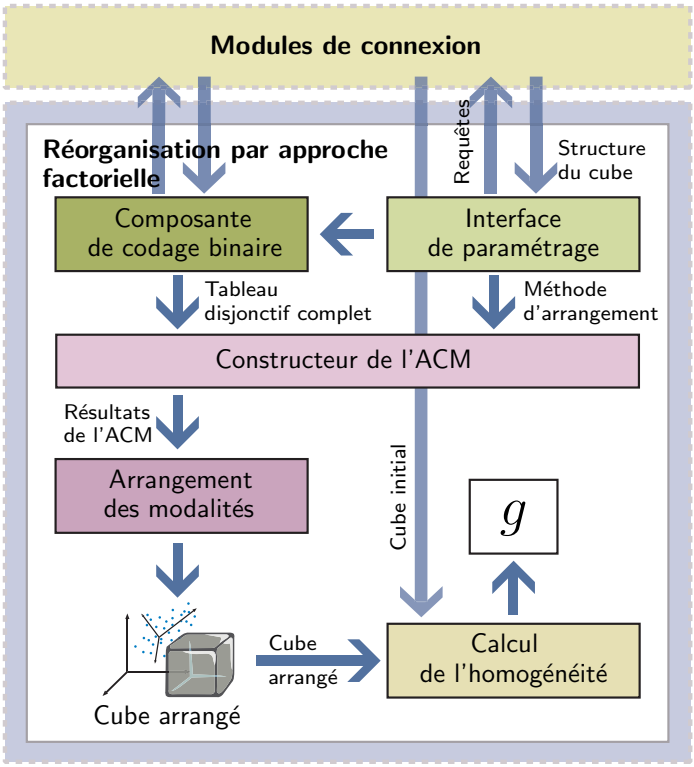


FIG. 6.3 – Architecture du module de réorganisation par approche factorielle

Le module de réorganisation par approche factorielle comporte cinq composantes principales : une *interface de paramétrage*, une composante de *codage binaire*, un *constructeur de l'ACM*, une composante d'*arrangement des modalités* et une

composante pour le *calcul de l'homogénéité*. La figure 6.3 illustre l'architecture de ces cinq composantes dans le module. Dans la suite, nous détaillons le rôle de chacune de ces composantes dans cette architecture.

Interface de paramétrage

L'interface de paramétrage assure la communication entre l'utilisateur et les modules de connexion aux données. Via cette interface, l'utilisateur connecte le module à un cube de données. L'interface importe ainsi la structure du cube en question. À partir de cette structure, l'utilisateur choisit les dimensions selon lesquelles il souhaite étudier et réorganiser les faits du cube. Dans chaque dimension choisie, l'utilisateur est aussi amené à préciser un niveau hiérarchique des modalités à arranger. L'interface de paramétrage demande à l'utilisateur de sélectionner un parmi les deux types d'arrangement que nous avons définis dans le chapitre 3, à savoir l'arrangement des modalités selon leurs projections ou selon leurs valeurs-test.

Ces paramètres étant fixés, la première étape d'exécution de ce module consiste à transformer les données sélectionnées du cube, selon un codage binaire, en un tableau disjonctif complet.

Composante de codage binaire

Cette composante reçoit, de l'interface de paramétrage, les dimensions et leurs niveaux hiérarchiques sélectionnés par l'utilisateur. En fonction de ces paramètres, la composante envoie des requêtes aux modules de connexion afin de récupérer les données du cube à transformer. Concrètement, pour chaque dimension sélectionnée, la composante formule dynamiquement une requête pour récupérer les faits OLAP ayant pris les modalités de cette dimension. Par conséquent la composante se charge du codage binaire de cette dernière en fonction des résultats fournis par la requête. L'ensemble des codages binaires de toutes les dimensions permet ainsi de construire le tableau disjonctif complet. Ce dernier est ensuite transmis au constructeur de l'ACM.

Constructeur de l'ACM

Le constructeur de l'ACM exécute l'algorithme de l'analyse des correspondances multiples et assure la construction des axes factoriels F_α à partir du tableau disjonctif complet. En fonction de la méthode d'arrangement choisie par l'utilisateur, le constructeur de l'ACM calcule les contributions $Cr_\alpha(a_t^i)$ des modalités aux axes factoriels, leurs projections et leurs valeurs-test $V_{\alpha t}^i$. Ces résultats sont ensuite transmis à la composante d'arrangement des modalités.

Composante d'arrangement des modalités

En fonction de la méthode d'arrangement choisie, cette composante arrange les

modalités de chaque dimension du cube de données étudié.

- Dans le cas d'un arrangement selon les projections des modalités, la composante cherche, pour une dimension donnée, l'axe factoriel F_{α^*} qui a été le mieux expliqué par les modalités de cette dimension. Les modalités de cette dimension sont ensuite triées selon l'ordre de leurs projections sur l'axe F_{α^*} (voir la sous-section 3.6.1).
- Dans le cas d'un arrangement selon les valeurs-tests des modalités, la composante calcule les valeurs-test des modalités selon les s premiers axes factoriels choisis par l'utilisateur. Les modalités de chaque dimension sont ensuite triées selon ces valeurs-test (voir la sous-section 3.6.2).

Suite à l'arrangement des modalités de chaque dimension du cube, une nouvelle représentation du cube réorganisé est fournie à l'utilisateur. Afin d'apprécier la qualité de cette nouvelle représentation par rapport à la représentation originale du cube, le module fournit à l'utilisateur le gain en homogénéité réalisé via la composante de calcul de l'homogénéité.

Composante de calcul de l'homogénéité

Cette composante calcule l'indice d'homogénéité du cube initial $IH(\mathcal{C}_{ini})$ et celui du cube réorganisé $IH(\mathcal{C}_{arr})$. Le premier indice est calculé à partir des modules de connexion alors que le second est calculé à partir du cube arrangé obtenu à partir de la composante d'arrangement des modalités selon le formalisme développé dans la section 3.7. Ainsi, le gain d'homogénéité g est calculé et fourni à l'utilisateur.

6.2.3 Module d'agrégation par classification

Le module d'agrégation par classification comprend quatre composantes principales : une *interface de paramétrage*, un *chargeur de données*, un *constructeur de la CAH* et une composante pour l'*évaluation des agrégats*. La figure 6.4 résume l'architecture de ces quatre composantes dont nous détaillons les rôles et les fonctions dans la suite.

Interface de paramétrage

L'interface de paramétrage permet à l'utilisateur de choisir, d'une manière assistée, des dimensions et des mesures qui correspondent aux individus et aux variables de la classification. Cette assistance prend en compte les contraintes logiques et statistiques dans le choix des individus et des variables de la classification que nous avons évoquées dans la section 4.3. L'interface permet aussi à l'utilisateur de choisir les paramètres de l'algorithme de la CAH. En effet, l'utilisateur peut sélectionner une mesure de distance entre individus et entre groupes d'individus parmi celles que nous avons

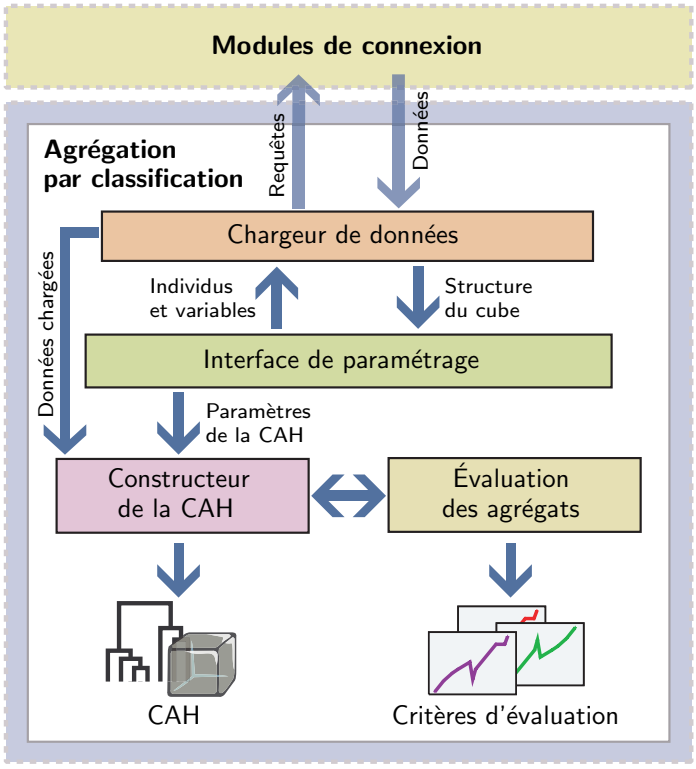


FIG. 6.4 – Architecture du module d'agrégation par classification

présentées dans la section 4.4.

Chargeur de données

Le chargeur des données est la composante qui fait le lien entre les modules de connexion aux sources de données multidimensionnelles et l'interface de paramétrage. Le chargeur des données transmet, à partir d'un module de connexion, la structure d'un cube à étudier. Cette structure comprend les dimensions, les hiérarchies et les mesures du cube en question.

Suite au choix de l'utilisateur des individus et des variables de la classification, le chargeur de données envoie des requêtes au module de connexion afin de récupérer les données relatives à ces individus et variables. Le langage de ces requêtes dépend de la source des données prise en compte dans le module de connexion. Rappelons, que dans notre plateforme MiningCubes les données sources peuvent provenir d'un cube de données sous *Analysis Services* de MicroSoft SQL Server 2000 ou d'un cube de données XML.

Quand le module de connexion répond aux requêtes du chargeur de données, ce

dernier charge en mémoire les données correspondant à l'ensemble de ces requêtes. Il transmet ensuite ces dernières au constructeur de la CAH.

Constructeur de la CAH

Le constructeur de la CAH assure l'exécution de l'algorithme de la classification présenté dans la section 4.4. Il permet aussi de présenter les résultats de cette classification à l'utilisateur sous forme d'un dendrogramme. Le dendrogramme est accompagné par un résumé des données (les dimensions et les mesures, le nombre d'individus, le nombre de variables, etc.), ainsi qu'un rappel des paramètres du modèle de classification (les mesures de distance utilisées). Cependant, l'affichage et l'interprétation d'un dendrogramme deviennent de plus en plus difficiles avec l'augmentation du nombre d'individus. Afin de contourner ce problème et d'assurer une présentation intelligible et interactive de l'information à visualiser, nous avons construit un outil permettant à l'utilisateur de couper le dendrogramme à différents niveaux hiérarchiques. Cet outil permet, également, de réduire et d'agrandir la taille du dendrogramme par une fonction de mise en échelle.

Composante d'évaluation des agrégats

En parallèle à la construction de la CAH, la composante d'évaluation des agrégats calcule les trois critères de qualité des partitions obtenues par la CAH, critères que nous avons définis dans la section 4.5. Pour chaque partition construite par la CAH, cette composante calcule les valeurs des inerties intra et inter-classes ainsi que la valeur du critère de la séparabilité des classes. Quand le constructeur de la CAH passe d'une partition à la partition suivante, la composante d'évaluation des agrégats calcule la variation de l'inertie interne selon le critère de la méthode de *Ward*. À la fin de l'algorithme de classification, des représentations graphiques de chaque critère en fonction du nombre d'agrégats sont fournies à l'utilisateur. Ces graphiques résument le comportement des critères de qualité des partitions et présentent ainsi un support visuel efficace pour l'utilisateur afin de choisir la meilleure partition.

6.2.4 Module d'explication par règles d'association

Le module d'explication par règles d'association est constitué de quatre composantes : une *interface de paramétrage*, une composante pour la *recherche des motifs fréquents* à partir d'un cube de données, une composante pour l'*extraction des règles d'association* à partir des motifs et une composante pour le *codage graphique* des règles d'association. Nous illustrons l'architecture de ce module par la figure 6.5 que nous détaillons dans la suite.

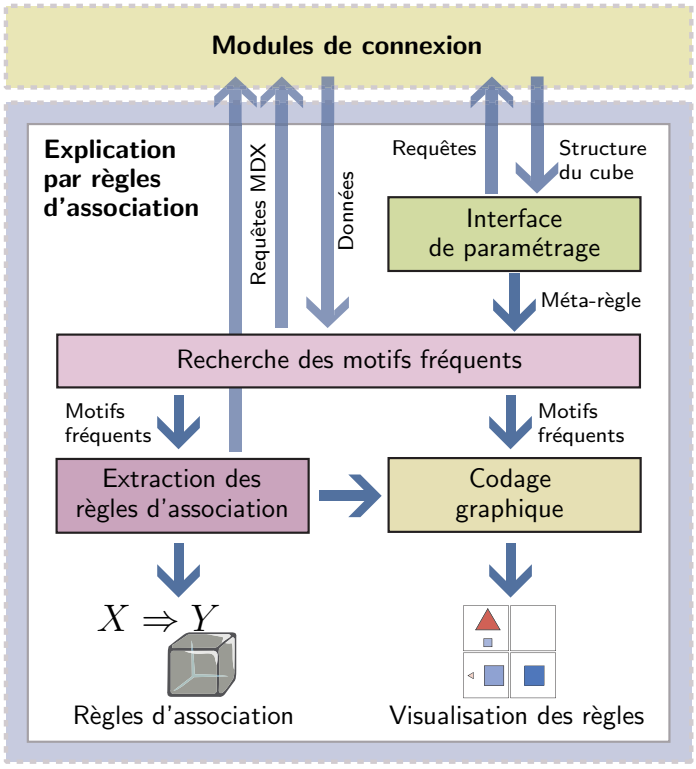


FIG. 6.5 – Architecture du module d'explication par règles d'association

Interface de paramétrage

Cette composante assure le lien entre le module de connexion aux sources de données multidimensionnelles, sous *Analysis Services* de MicroSoft SQL Server 2000, et la composante de recherche des motifs fréquents. À partir du module de connexion, l'interface de paramétrage récupère la structure du cube de données étudié. Cette structure comprend les dimensions, les hiérarchies et les mesures du cube en question. L'utilisateur peut ainsi construire dans l'interface de paramétrage la méta-règle inter-dimensionnelles en définissant : les dimensions d'analyse $\mathcal{D}_{\mathcal{A}}$, les dimensions de contexte $\mathcal{D}_{\mathcal{C}}$, le sous-cube de contexte $(\Theta_1, \ldots, \Theta_p)$, le schéma de la méta-règle $(\alpha_1 \wedge \cdots \wedge \alpha_s) \Rightarrow (\beta_1 \wedge \cdots \wedge \beta_r)$, la mesure M pour le calcul des critères des règles d'association, le support minimum *minsupp* et la confiance minimale *minconf* (voir figure 6.6). Suite à la validation de la méta-règle définie par l'utilisateur, l'ensemble des paramètres est transmis à la composante de recherche des motifs fréquents.

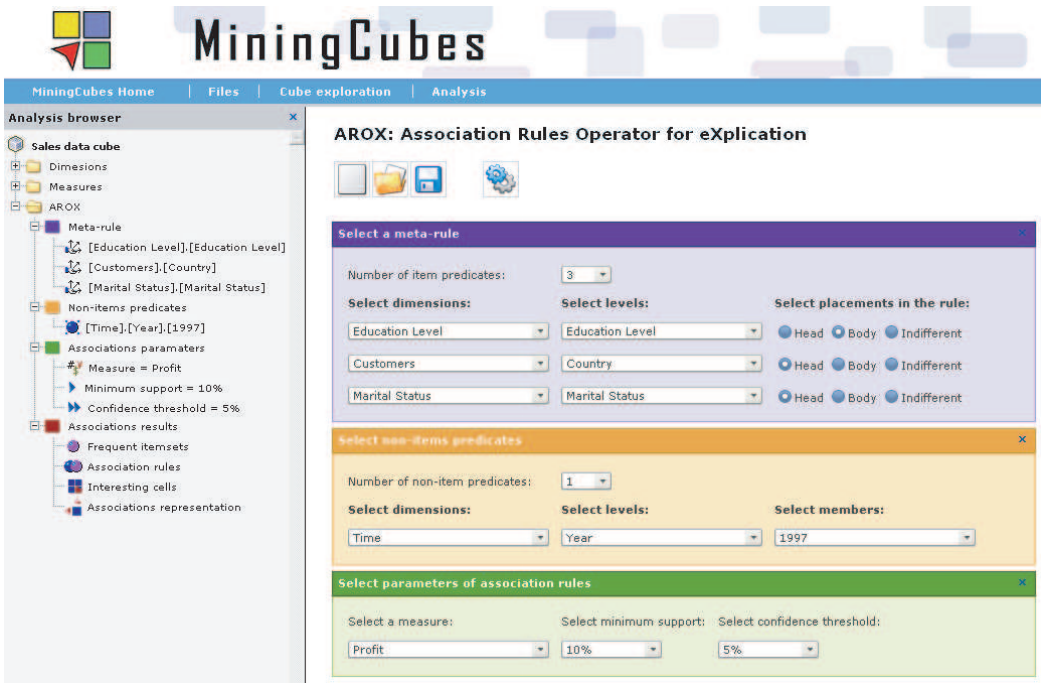


FIG. 6.6 – Interface de paramétrage du module d'explication par règles d'association dans MiningCubes

Composante de recherche des motifs fréquents

Cette composante recherche les motifs fréquents en tenant compte des paramètres de la méta-règle définie par l'utilisateur. Cette recherche se fait selon la stratégie ascendante que nous avons exposée dans la section 5.7. Via un jeu de requêtes MDX, cette composante communique directement avec la source des données multidimensionnelles afin d'en extraire les données nécessaire pour la construction des motifs et les agrégats nécessaires au calcul des supports. Les motifs fréquents trouvés sont ensuite envoyés à la composante d'extraction des règles d'association ainsi qu'à la composante du codage graphique.

Composante d'extraction des règles d'association

À partir des motifs fréquents, cette composante se charge de l'extraction des règles d'association en respectant le schéma de la méta-règle multidimensionnelle défini par l'utilisateur selon l'algorithme que nous avons présenté dans la section 5.7.

L'évaluation des règles découvertes (confiance, *Lift* et indice de *Loevinger*) est assurée par des requêtes MDX qui interrogent directement la source des données multidimensionnelles.

Composante de codage graphique

Cette composante récupère les motifs fréquents et les règles d'association découverts dans les deux composantes précédentes et assure le codage graphique. Ce codage prend en compte les valeurs du support, de la confiance, du *Lift* et de l'indice de *Loevinger* selon l'approche de visualisation que nous avons mis en place dans la section 5.8. La composante se charge également de l'intégration de ces codages graphiques dans l'espace de représentation du cube de données. Ainsi, comme le montre l'exemple de la figure 6.7, une visualisation des motifs fréquents et des règles d'association découvertes est fournie dans l'espace de représentation du cube de données.

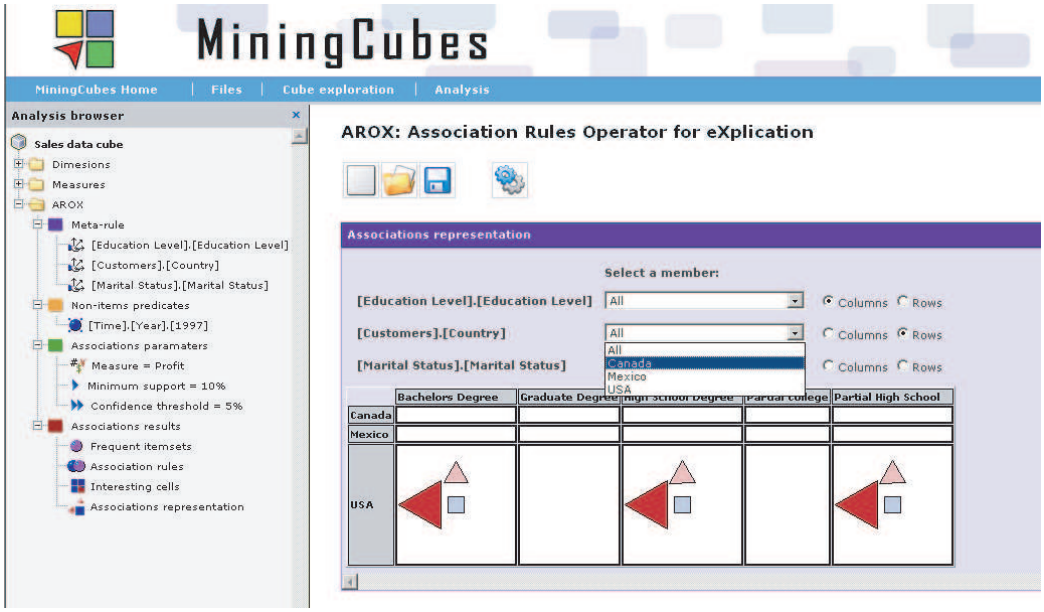


FIG. 6.7 – Visualisation des motifs fréquents et des règles d'association dans MiningCubes

6.3 Jeu de données complexes

Nous présentons, dans cette section, le jeu de données complexes que nous utilisons dans le cas d'application de l'agrégation par classification. Il s'agit d'un jeu de données médicales relatives au domaine du dépistage du cancer du sein. Nous avons extrait ces

données à partir de la base DDSM (*Digital Database for Screening Mammography*) mise en place par l'*université de South Florida* en collaboration avec plusieurs organismes américains dans le cadre d'un projet fédéral de recherche dans le domaine du cancer du sein aux États-Unis d'Amérique[2]. La base DDSM est une ressource libre mise à la disposition des communautés de recherche dans le domaine médical et celui de la fouille de données. L'objectif de ce projet est de promouvoir la recherche et le développement de méthodes d'apprentissage automatique pour la prévention du cancer du sein [HBK+00].

Dans un premier temps, nous récupérons et sauvegardons, à partir de la base DDSM, les données de mammographies selon le format XML. Nous obtenons ainsi un corpus de données semi-structurées qui seront traitées et modélisées par la suite afin de construire un cube de données XML.

6.3.1 Présentation de la base DDSM

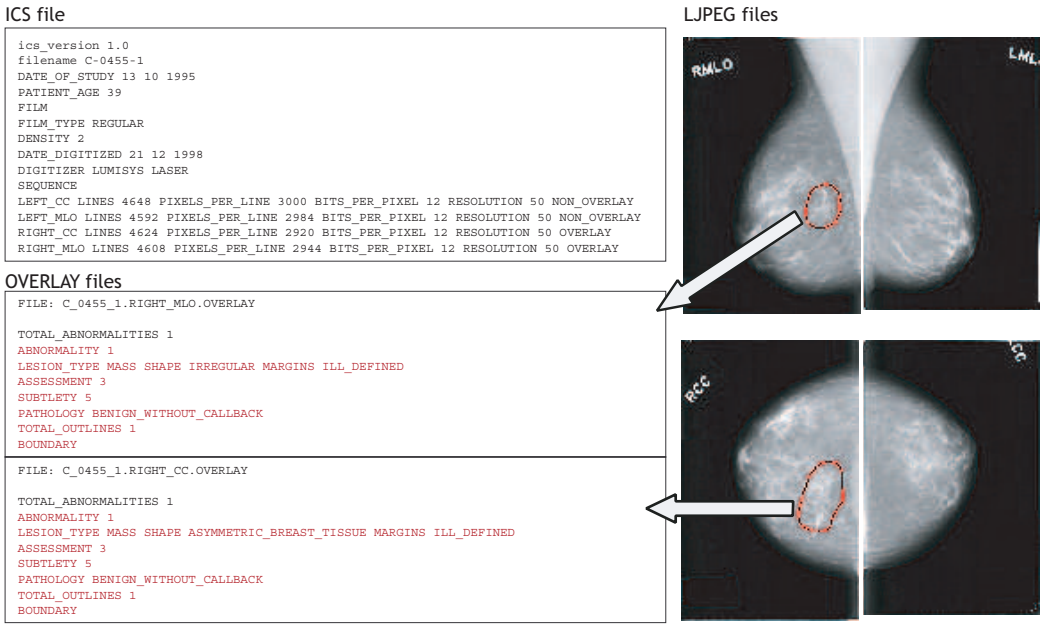


FIG. 6.8 – Exemple d'un dossier de patient pris de la base DDSM

La base DDSM contient 2 604 dossiers numériques où chaque dossier correspond à un cas d'une patiente. Un dossier d'une patiente est constitué par un ensemble

[2]Les ressources de la base DDSM ainsi qu'un exposé complet sur les acteurs de ce projet sont en libre accès sur le Web à l'adresse : `http://marathon.csee.usf.edu/Mammography/Database.html`

d'images et de fichiers textes. Les images représentent des radios numérisées de mammographies suite à la réalisation de scanners lors d'un examen médical. Les fichiers textes contiennent des informations d'ordre général sur la patiente ainsi que des observations médicales annotées par des médecins radiologues en se basant sur les radios des mammographies. Dans la base DDSM, on distingue quatre types de cas de patientes : le cas *normal*, le cas du *cancer bénin sans rappel*, le cas du *cancer bénin* et le cas du *cancer malin*.

1. Le cas *normal* correspond à une mammographie qui, suite à un examen médical, est déclarée saine par les médecins.
2. Le cas du *cancer bénin sans rappel* correspond à une mammographie dans laquelle une ou plusieurs tumeurs bénignes ont été remarquées par les médecins. La patiente n'est pas rappelée pour des examens médicaux complémentaires.
3. Le cas du *cancer bénin* correspond à une mammographie où les médecins ont observé une ou plusieurs régions suspectes dans ses images radios. Une de ces régions peut être à l'origine d'un *cancer malin*. La patiente est rappelée pour des examens médicaux complémentaires.
4. Le cas du *cancer malin* correspond à une mammographie où, après un certain nombre d'examens médicaux, une ou plusieurs tumeurs ont été détectées et confirmées malignes.

Pour l'essentiel, comme le montre la figure 6.8, dans le dossier d'une patiente on trouve un fichier texte (format ASCII) contenant des descriptions textuelles générales (des méta-données) concernant la date de l'examen médical, l'âge de la patiente, la date de numérisation des radios, le type de numérisation et la liste des fichiers images contenues dans le dossier.

Le dossier contient aussi quatre fichiers images LJPEG (*Image compressed with lossless JPEG encoding*) représentant des radios numérisées. Chacune de ces images correspond à une radio numérisée présentant un angle de vue du sein. Nous trouvons ainsi deux radios *Crânio-Caudales* pour le sein droit et le sein gauche (Left_CC et Right_CC) et deux radios *Médio-Latérales Obliques* du sein droit et du sein gauche (Left_MLO et Right_MLO).

De plus, chaque radio présentant une ou plusieurs régions suspectées est associée à un fichier texte (format ASCII) de recouvrement (*OVERLAY*) qui décrit ces dernières. Ainsi, un dossier d'une patiente peut contenir, au maximum, quatre fichiers de recouvrement. Seuls les dossiers des patientes présentant des cas normaux ne contiennent pas de fichiers de recouvrement.

Un fichier de recouvrement contient les annotations des médecins portant sur les propriétés médicales des régions suspectées : le type de la lésion, la pathologie, un indice d'évaluation et un indice de subtilité. Il contient aussi une description spatiale des régions suspectées. Cette description se base sur un codage numérique de

la frontière d'une région suspectée, délimitée et marquée sur la radio par un médecin radiologue.

6.3.2 Corpus XML des données de mammographies

À l'image des données complexes, les données de la base DDSM sont hétérogènes car elles proviennent de différentes sources. Ces données sont de format différents et se présentent sur plusieurs types de supports. En effet, nous avons vu qu'un dossier d'une patiente peut contenir un nombre variable de supports de données. En plus, ces supports n'ont pas tous le même format, présentent des données non structurées et font abstraction des niveaux de granularité de l'information.

Pour exploiter ces données à des fins décisionnelles, il est nécessaire de les structurer et de les homogénéiser. Le langage XML s'avère une solution appropriée à ce travail préparatoire sur ces données médicales et sur les données complexes d'une manière générale. En effet, la représentation semi-structuré des données, la capacité d'auto-description et la structure arborescente donnent au formalisme XML une grande flexibilité et une puissance suffisante pour décrire des données complexes, hétérogènes et provenant de sources éparpillées.

De plus, un document XML peut-être formé conformément à une grammaire associée, exprimée sous forme de DTD ou de Schéma XML. Cette grammaire assure la définition et la validation d'une structure commune à un ensemble de documents homogènes et respectant les mêmes types de données. Néanmoins, si les DTDs n'offrent qu'un seul type de données (chaînes de caractères), les schémas XML utilisent davantage de types de données et permettent également de définir des types complexes.

Ainsi, dans le cadre de notre cas d'application, nous stockons, dans un premier temps, les données de la base DDSM dans des documents XML. Nous construisons ainsi un *corpus XML de données de mammographies*[3]. Dans ce corpus, tous les documents XML respectent la même grammaire définie par un schéma XML commun et chaque document XML retranscrit les données d'un, et d'un seul, dossier médical. La figure 6.9 montre un exemple d'un documents XML de ce corpus.

Le langage XML permet certes de stocker et d'homogénéiser les ressources d'une base de données complexes en leur donnant une représentation semi-structurée. Cependant, à ce stade, XML ne permet pas encore d'exploiter ces données dans un processus décisionnel. Nous pensons qu'il est possible d'utiliser davantage XML, non seulement comme un langage de stockage des données, mais aussi comme un langage de modélisation multidimensionnelle des entrepôts de données. Nous proposons dans la suite d'utiliser XML à un niveau d'abstraction supérieur, dépassant

[3]Nous avons mis le corpus XML des données de mammographies en libre accès sur le Web à l'adresse : `http://eric.univ-lyon2.fr/~rbenmessaoud/?page=donnees`

```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
  <!--  Edited by Riadh Ben Messaoud from Digital Database for Screening Mammography (DDSM)  -->
- <case>
    <case_id>996</case_id>
    <case_type>cancer</case_type>
    <ics_version>1.0</ics_version>
    <ics_file_name>C-0025-1.ics</ics_file_name>
    <ics_file_url>ftp://figment.csee.usf.edu/DDSM/C-0025-1.ics</ics_file_url>
    <date_of_study>1993-08-16</date_of_study>
    <patient_age>49</patient_age>
    <film_type>regular</film_type>
    <density>1</density>
    <date_digitized>1997-10-09</date_digitized>
    <digitizer>lumisys laser</digitizer>
    <sequence>sequence</sequence>
  - <left_cc_scanner>
     <scanner_file_name>C_0025_1.LEFT_CC.LJPEG</scanner_file_name>
     <scanner_file_url>ftp://figment.csee.usf.edu/DDSM/C_0025_1.LEFT_CC.LJPEG</scanner_file_url>
     <scanner_lines>5872</scanner_lines>
     <scanner_pixels_per_line>3696</scanner_pixels_per_line>
     <scanner_bits_per_pixel>12</scanner_bits_per_pixel>
     <scanner_resolution>50</scanner_resolution>
     <scanner_overlay_presence>true</scanner_overlay_presence>
   - <overlay>
      <overlay_file_name>C_0025_1.LEFT_CC.OVERLAY</overlay_file_name>
      <overlay_file_url>ftp://figment.csee.usf.edu/DDSM/C_0025_1.LEFT_CC.OVERLAY</overlay_file_url>
      <abnormalities_number>1</abnormalities_number>
    - <abnormality>
       <abnormality_id>1</abnormality_id>
       <lesion_type>mass shape oval margins spiculated</lesion_type>
       <assessment>5</assessment>
       <subtlety>5</subtlety>
       <pathology>malignant</pathology>
       <boundaries_number>1</boundaries_number>
     - <boundary>
        <boundary_id>1</boundary_id>
        <starting_column>2000</starting_column>
        <starting_row>2480</starting_row>
        <chain_code>7 7 7 7 6 7 7 7 7 7 7 6 6 ... 7 7 6 7 6 7 6 7 6 7 7 7 6 6 7 7 7 </chain_code>
       </boundary>
      </abnormality>
     </overlay>
    </left_cc_scanner>
  + <left_mlo_scanner>
  + <right_cc_scanner>
  + <right_mlo_scanner>
  </case>
```

FIG. 6.9 – Un document XML du corpus des données de mammographies

le simple stockage des données complexes. Nous employons XML dans le cadre d'une démarche conceptuelle, logique et physique pour l'entreposage des données complexes, notamment ceux du corpus XML des données de mammographies.

En marge de notre problématique de base du couplage entre l'analyse en ligne et la fouille de données, nous avons développé en parallèle des travaux sur une méthodologie, appelée X-Warehousing, entièrement basée sur XML pour la modélisation et l'entreposage des données complexes [BMCA06a, BMCA06b]. Dans la section suivante, nous présentons brièvement notre méthodologie X-Warehousing qui va nous permettre par la suite de construire un cube de données XML à partir du corpus XML que nous avons construit.

6.4 Méthodologie d'entreposage des données complexes

Comme nous l'avons montré avec les données des mammographies, le document XML est perçu, incontestablement, comme un moyen efficace pour représenter et stocker des données complexes. Cependant, dans le cadre des applications décisionnelles, des efforts sont nécessaires, plus précisément d'un point de vue méthodologique, pour construire des solutions d'entreposage de documents XML. En effet, l'organisation multidimensionnelle des entrepôts de données diffère de celle semi-structurée des documents XML. Un entrepôt de données a une architecture intégrée, centralisée, orientée sujets et nécessite des rafraîchissements périodiques des données pour garantir leur historisation [Kim96, Inm96]. À partir de l'entrepôt, il est possible de construire des cubes de données représentant des contextes d'analyse multidimensionnels. Ainsi, la difficulté est comment peut-on concevoir un modèle multidimensionnel à l'aide d'un formalisme semi-structuré tel que XML ?

6.4.1 Entrepôts de données XML

L'entreposage des données XML est un problème relativement récent dans la communauté des bases de données. À notre connaissance Krill était le premier, vers la fin des années 90, à souligner l'importance de l'emploi du formalisme XML dans les systèmes de gestion des bases de données [Kri98]. Il affirme que XML assure une interopérabilité entre les entrepôts, les sources de données et les outils d'entreposage. L'auteur prévoit même que les éditeurs, tels que Microsoft, IBM et Oracle, emploient largement XML dans les années à venir dans leurs systèmes de gestion de bases de données. Dès lors, nous avons assisté, ces dernières années, à l'émergence de plusieurs efforts de recherche qui ont porté un intérêt particulier aux entrepôts de données XML.

Néanmoins, nous distinguons deux grandes approches qui, d'ailleurs, traduisent une nuance souvent rencontrée dans la communauté impliquée dans ce sujet de

recherche. La première est une approche d'*entreposage des données XML*. Elle considère les entrepôts de données XML comme un support de stockage physique des documents XML où les données sont exprimées selon le formalisme XML. Par exemple, les travaux de Golfarelli *et al.* [GRV01a, GRV01b], Trujillo *et al.* [TLMS04], Nassis *et al.* [NRDR04, NRDR05, NRDR06] et Zhang *et al.* [ZWLZ05] abordent le problème de l'alimentation des entrepôts par les données XML. Le formalisme XML est ainsi considéré comme une technologie efficace, supportant des données faiblement structurées, adaptée à l'interopérabilité et à l'échange des données. La seconde approche est une approche de *modélisation multidimensionnelle en XML*. Par exemple, les travaux de Baril et Bellahsène [BB00, BB03], Pkorný [Pok01, Pok02], Vrdoljak *et al.* [VBR03], Hümmer *et al.* [HBH03], Rusu *et al.* [RRT04, RRT05], Rajugan *et al.* [RCDF03, RCD05a, RCD05b, RCDF05] et Park *et al.* [PHS05] tentent de modéliser les entrepôts de données avec le formalisme XML et considèrent XML comme un support pour la modélisation des schémas des entrepôts de données. Le formalisme XML est alors utilisé pour concevoir des entrepôts, ou des magasins de données, selon les modèles multidimensionnels classiques tels que les schémas en étoile ou en flocons de neige.

Entièrement basée sur XML, notre méthodologie X-Warehousing combine les deux approches précédentes. En effet, elle permet en même temps, de concevoir des entrepôts avec des schémas XML et de les alimenter avec des documents XML valides. Dans notre proposition, XML est ainsi perçu, à la fois, comme un langage de modélisation logique et comme un format de stockage physique des données. Notre méthodologie est également pilotée par les besoins d'analyse. Elle part des besoins d'analyse des utilisateurs afin de fournir le modèle multidimensionnel adéquat et qui répond le mieux à ces besoins.

Notons aussi que notre finalité n'est pas seulement d'entreposer les données XML. Nous considérons plutôt XML comme un moyen intermédiaire pour l'entreposage des données complexes. En effet, il est possible de représenter avec XML des données complexes comme une information entière plutôt qu'un ensemble d'informations séparées comme c'est le cas dans un dossier de patient par exemple. Une collection de documents XML – où chaque document est une entité informationnelle – peut alors être considérée comme une base décisionnelle représentant un entrepôt (ou un magasin) de données complexes. Pour l'essentiel, dans X-Warehousing, nous utilisons XML comme formalisme pour la modélisation et l'alimentation des entrepôts de données complexes. Le choix de construire des structures multidimensionnelles en XML est dicté par le fait de la prolifération des documents XML d'une part, et par le fait que les données complexes peuvent être décrites sous forme de documents XML d'autre part.

Nous présentons, dans la suite, une vue d'ensemble sur notre méthodologie et exposons ensuite le cas de l'entreposage des données complexes du corpus des mammographies.

6.4.2 Vue d'ensemble de la méthodologie X-Warehousing

À partir d'un corpus XML de données complexes, notre méthodologie permet de générer un cube XML exprimant un contexte d'analyse. Ce dernier est composé d'une collection de documents XML. Chaque document correspond alors à un fait OLAP qui doit satisfaire certaines contraintes, comme respecter une information minimale pour que le fait à observer soit consistant. Pour cela, nous proposons de valider les documents par un schéma XML qui représente le modèle conceptuel du cube qui généralement consiste en un schéma en étoile ou en flocons de neige.

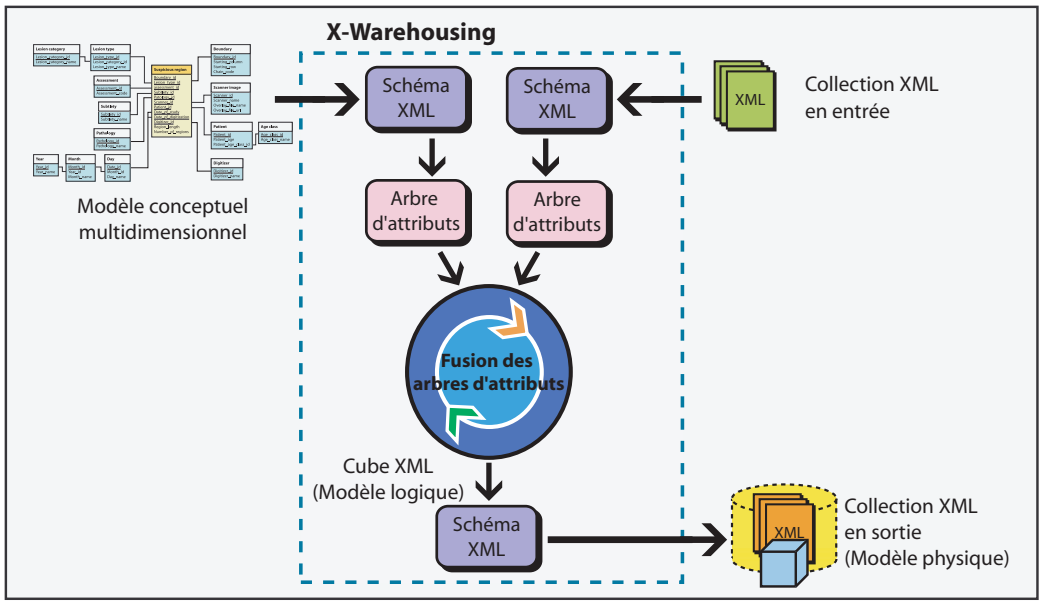


FIG. 6.10 – Les étapes de l'approche X-Warehousing

La figure 6.10 résume les différentes étapes de X-Warehousing. Il s'agit d'une démarche méthodologique pilotée par les besoins d'analyse. Au départ, l'utilisateur déclare ses objectifs d'analyse sous la forme d'un modèle conceptuel multidimensionnel (MCM). L'utilisateur soumet aussi un ensemble de données complexes structurées dans des documents XML. Il est à noter qu'il n'est pas utile de chercher au départ une structure commune multidimensionnelle dans les documents XML sources, car elle n'y existe pas *a priori*. Ainsi, la méthodologie de X-Warehousing se focalise plutôt sur les besoins d'analyse de l'utilisateur qui sont représentés par un schéma en étoile. Ce dernier exprime comment représenter conceptuellement les données pour les orienter vers l'analyse et cela indépendamment de la manière de les organiser aux niveaux logiques ou physiques.

Le MCM et les documents XML sources sont initialement exprimés à l'aide

de schémas XML pour être appariés. Cependant pour les rendre comparables, ils sont alors traduits sous forme d'arbre d'attributs [GMR98, GR99]. Les deux arbres d'attributs représentant respectivement le MCM et les documents XML sources qui seront fusionnés selon une certaine stratégie à l'aide d'opérations de fusion par élagage (*pruning*) et/ou par greffe (*grafting*) [GRV01a, GRV01b].

Selon cet appariement, nous n'extrayons dans les documents XML sources que les données utiles pour les besoins d'analyse, pour alimenter le cube à construire. Cependant, afin d'obtenir des faits OLAP consistants, les documents XML sources doivent aussi satisfaire une information minimale requise par les objectifs d'analyse de l'utilisateur. Ainsi, au moment de la fusion de deux arbres d'attributs, deux cas de figure sont possibles :

1. si le document en entrée contient l'information minimale requise dans le MCM, il est alors accepté. Une instance de ce document sera créée et validée en accord avec le schéma XML du cube XML que génère l'application X-Warehousing ;
2. si le document soumis en entrée ne contient pas une information suffisante pour un fait OLAP, il est alors rejeté et ne fera donc pas partie du cube XML.

La démarche de X-Warehousing permet d'obtenir une collection homogène de documents XML avec des contraintes strictes sur leurs contenus. Cette collection entreposés correspond au modèle physique d'un cube OLAP que nous désignons par *cube XML*. Ainsi, le cube XML est composé de documents XML valides et conformes le plus possible au MCM de départ, c'est-à-dire, aux objectifs d'analyse de l'utilisateur. Chaque document XML de ce cube représente un fait OLAP constitué d'un ou de plusieurs indicateurs (mesures) à observer à travers des axes d'analyse (dimensions et hiérarchies de dimensions). Pour un exposé plus complet sur cette démarche méthodologique, nous renvoyons le lecteur à [BMCA06a, BMCA06b].

6.5 Construction du cube XML des données de mammographies

Pour préparer les données médicales de la base DDSM à nos fins d'analyse, nous avons mis en œuvre notre méthodologie d'entreposage des données complexes pour générer un cube XML à partir du corpus XML des données de mammographies. Nous considérons alors les documents XML du corpus des mammographies comme la collection des données sources en entrée de X-Warehousing.

Comme le veut la méthodologie X-Warehousing, le cube XML correspond à un contexte OLAP piloté par les objectifs d'analyse de l'utilisateur qui sont exprimés par ce dernier via un modèle conceptuel multidimensionnel (MCM). Ainsi, en plus de la collection des documents XML en entrée, nous définissons un contexte d'analyse en ligne des données de mammographies.

6.5.1 Contexte d'analyse dans les données de mammographies

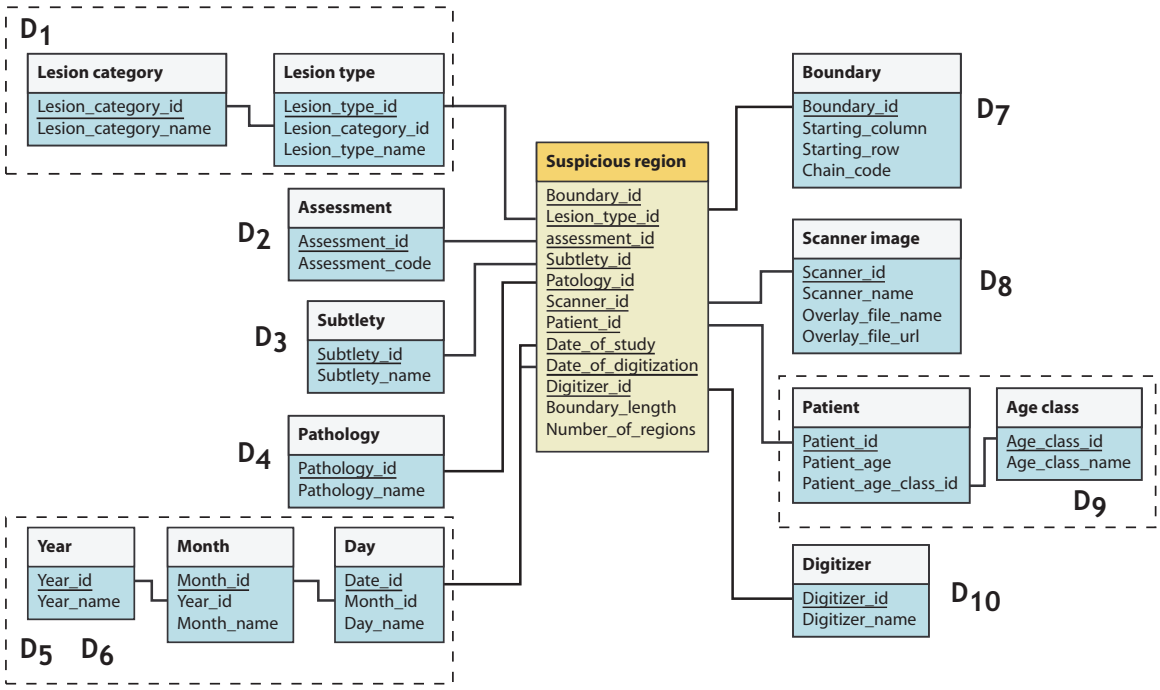


FIG. 6.11 – Modèle conceptuel du cube de données des *régions suspectes*

Dans le cas d'une application médicale dont l'objectif est de prévenir le cancer du sein, nous pouvons imaginer qu'un expert cherche à étudier la population des régions suspectes dans les radios de mammographies. Ces régions peuvent correspondre à des cas de tumeurs cancérigènes.

Dans le cas de la base des mammographies DDSM, nous avons vu, dans la section 6.3, qu'une région suspecte peut correspondre à un cas de cancer bénin sans rappel, un cas de cancer bénin ou un cas de cancer malin. De plus, une région suspecte est décrite par un ensemble de descripteurs catégoriels qui renseignent sur ses propriétés médicales. Une région suspecte est aussi décrite par ses propriétés spatiales qui déterminent sa position et sa frontière sur une radio.

Ainsi, comme le montre le modèle conceptuel de la figure 6.11, nous proposons un contexte d'analyse en ligne où les régions suspectes représentent les faits OLAP à étudier. Selon ce contexte, l'objectif est d'analyser les régions suspectes dans la base des mammographies selon les dimensions suivantes :

- D_1 : le type de la lésion (*Lesion_type*) ;
- D_2 : l'indice d'évaluation de la lésion (*Assessment*) ;

- D_3 : l'indice de subtilité de la lésion (*Subtlety*) ;
- D_4 : la pathologie (*Pathology*) ;
- D_5 : la date d'examen de la patiente qui correspond à la région suspecte (*Date_of_study*) ;
- D_6 : la date de numérisation de la radio de la région suspecte (*Date_of_digitization*) ;
- D_7 : la frontière de la région suspecte marquée par les médecins radiologues sur la radio (*Boundary*) ;
- D_8 : l'image radio numérisée de la région suspecte (*Scanner_image*) ;
- D_9 : la patiente qui correspond à la région suspecte (*Patient*) ;
- D_{10} : la machine utilisée pour la numérisation de l'image radio (*Digitizer*).

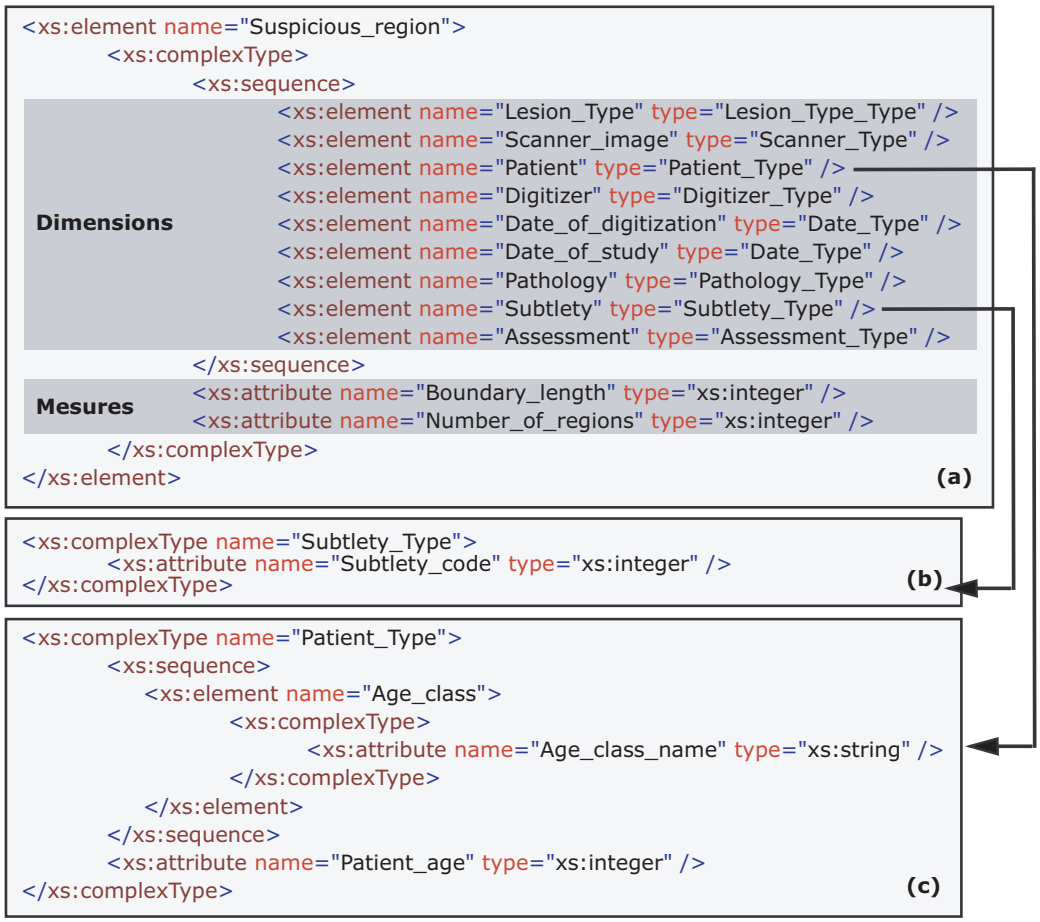
Selon ce modèle, une région suspecte est observée selon la mesure du *périmètre* de sa frontière (*Boundary_length*). Cette mesure indique l'importance de la région suspecte. En effet, plus le périmètre d'une région est grand, plus la surface de cette région est importante. Une région à grande surface peut représenter un risque d'une tumeur maligne. En plus du périmètre, pour une région suspecte, nous proposons aussi d'utiliser une deuxième mesure dérivée. Elle renseigne sur le *nombre des régions* suspectes dans le dossier de la patiente concernée (*Number_of_regions*).

6.5.2 Cubes XML des données de mammographies

En tenant compte du contexte d'analyse précédent, la mise en œuvre de X-Warehousing sur le corpus XML des données de mammographies nous a fourni un cube XML dont le modèle logique est décrit par un schéma XML. La figure 6.12 représente une vue fragmentaire de ce schéma XML.

Selon un formalisme XML de modélisation multidimensionnelle propre à la méthodologie X-Warehousing [BMCA06a, BMCA06b], un document XML valide par rapport à ce modèle logique représente un fait à analyser. D'une manière synthétique, dans un tel document, un fait est représenté en XML en respectant l'ensemble des règles suivantes :

- le nom du fait correspond au nom de l'élément racine du document XML ;
- le nom d'une mesure correspond au nom d'un attribut dans l'élément racine du document XML ;
- la valeur d'une mesure correspond à la valeur que prend l'attribut associé à cette mesure dans le document XML ;
- le nom d'une dimension correspond au nom d'un sous-élément de l'élément racine du document XML ;
- le nom d'un niveau hiérarchique d'une dimension correspond au nom d'un sous-élément de l'élément associé à cette dimension dans le document XML ;
- la modalité que prend un fait dans un niveau hiérarchique d'une dimension correspond à la valeur que prend un attribut dans l'élément associé à ce niveau hiérarchique dans le document XML.




```
<xs:element name="Suspicious_region">
    <xs:complexType>
        <xs:sequence>
            <xs:element name="Lesion_Type" type="Lesion_Type_Type" />
            <xs:element name="Scanner_image" type="Scanner_Type" />
            <xs:element name="Patient" type="Patient_Type" />
            <xs:element name="Digitizer" type="Digitizer_Type" />
            <xs:element name="Date_of_digitization" type="Date_Type" />
            <xs:element name="Date_of_study" type="Date_Type" />
            <xs:element name="Pathology" type="Pathology_Type" />
            <xs:element name="Subtlety" type="Subtlety_Type" />
            <xs:element name="Assessment" type="Assessment_Type" />
        </xs:sequence>
        <xs:attribute name="Boundary_length" type="xs:integer" />
        <xs:attribute name="Number_of_regions" type="xs:integer" />
    </xs:complexType>
</xs:element>
```

```
<xs:complexType name="Subtlety_Type">
    <xs:attribute name="Subtlety_code" type="xs:integer" />
</xs:complexType>
```

```
<xs:complexType name="Patient_Type">
    <xs:sequence>
        <xs:element name="Age_class">
            <xs:complexType>
                <xs:attribute name="Age_class_name" type="xs:string" />
            </xs:complexType>
        </xs:element>
    </xs:sequence>
    <xs:attribute name="Patient_age" type="xs:integer" />
</xs:complexType>
```

FIG. 6.12 – Fragments du modèle logique du cube XML des *régions suspectes*

Par exemple, d'après la figure 6.12 (a), le nom du fait *région suspecte* (*Suspicious_region*) correspond au nom de l'élément racine déclaré par le schéma XML. Les mesures de ce fait, *périmètre de la frontière* (*Boundary_length*) et *nombre de régions* (*Number_of_regions*), sont déclarées en tant qu'attributs, de type entier, dans l'élément racine.

Les dimensions d'une région suspecte sont représentées par des sous-éléments de l'élément racine. Chacun de ces sous-éléments est associé à un *type complexe* (*complexType*) qui définit la hiérarchie d'une dimension. Par exemple, la figure 6.12 (b) représente le type complexe qui définit la hiérarchie de la dimension de l'indice de *subtilité* (*Subtlety*). Cette dernière comporte un seul niveau hiérarchique. La figure 6.12 (c) représente un autre type complexe associé à la dimension de la *patiente* (*Patient*) qui définie deux niveaux hiérarchiques : l'*âge* (*Patient_age*) et la *classe d'âge*

```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
  <!--  Edited automatically by X-Warehousing software  -->
- <Suspicious_region Boundary_length="3325" Number_of_regions="4">
 - <Patient Patient_age="51">
    <Age_class Age_class_name="Between 50 and 59 years old" />
   </Patient>
 - <Lesion_type Lesion_type_name="mass shape round margins n/a">
    <Lesion_category Lesion_category_name="mass shape round" />
   </Lesion_type>
   <Assessment Assessment_code="2" />
   <Subtlety Subtlety_code="3" />
   <Pathology Pathology_name="benign_without_callback" />
 - <Date_of_study Date="1998-06-09">
  - <Day Day_name="June 9, 1998">
   - <Month Month_name="June, 1998">
      <Year Year_name="1998" />
     </Month>
    </Day>
   </Date_of_study>
 - <Date_of_digitization Date="1998-07-14">
  - <Day Day_name="July 14, 1998">
   - <Month Month_name="July, 1998">
      <Year Year_name="1998" />
     </Month>
    </Day>
   </Date_of_digitization>
   <Digitizer Digitizer_name="lumisys laser" />
   <Scanner_image Scanner_name="right_cc" Overlay_file_name="B_3160_1.RIGHT_CC.LJPEG"
   Overlay_file_url="ftp://figment.csee.usf.edu/DDSM/B_3160_1.RIGHT_CC.LJPEG" />
   <Boundary Starting_column="2568" Starting_row="2800" Chain_code="6 6 6 6 6 6 ... 0 0 0 0 0 0" />
  </Suspicious_region>
```

FIG. 6.13 – Un fait du cube XML des données de mammographies

de la patiente (*Age_class*).

La collection des instances du modèle logique constitue le modèle physique du cube XML des données de mammographies. À partir du corpus XML initial, la méthodologie X-Warehousing a permis de construire un cube XML dont le modèle physique comporte 4 686 documents XML. Chacun de ces documents est valide par rapport au modèle logique précédent et représente un fait relatif à une région suspecte dans les dossiers des patientes de la base DDSM. La figure 6.13 montre un exemple de l'un de ces documents XML.

Enfin, en utilisant XML comme formalisme de modélisation multidimensionnelle et comme format de stockage, nous avons réussi à construire un cube de données complexes relatives au domaine médical. Dans la suite, nous utiliserons notre cube XML des données de mammographies dans le cadre d'un cas d'application d'analyse. En utilisant notre plateforme MiningCubes, nous proposons de mettre en œuvre notre approche d'agrégation par classification sur ces données complexes du cube XML [MBR06a].

6.6 Agrégation des données complexes par classification

6.6.1 Objectifs de l'agrégation des données de mammographies par classification

Nous avons montré dans la section 4.2, à travers un exemple d'un cube de ventes, que notre approche d'agrégation par classification établit une nouvelle agrégation sémantique des modalités d'une dimension du cube de données. À la différence de l'agrégation OLAP classique, qui prend en compte l'ordre d'appartenance logique des modalités sur différents niveaux hiérarchiques de la dimension, notre agrégation par classification tient plutôt compte des faits réels contenus dans le cube de données.

Dans le cas du cube XML des données des mammographies, selon une agrégation OLAP classique dans la dimension *patiente* (D_9), nous pouvons construire des agrégats de régions suspectes comme ceux représentés dans l'exemple de la figure 6.14. Dans cet exemple, nous pouvons remarquer que, dans chaque agrégat, les régions suspectes ont des formes et des tailles différentes. En plus, ces régions n'ont pas forcément les mêmes propriétés médicales. En effet, selon les annotations des médecins, les régions suspectes (c), (e) et (g) dans l'agrégat "*patientes entre 40 et 49 ans*" représentent des pathologies malignes. Les autres régions suspectes dans ce dernier agrégat représentent plutôt des pathologies bénignes. Dans le second agrégat "*patientes entre 50 et 59 ans*", seules les régions (b) et (c) sont malignes, alors que les autres régions sont bénignes.

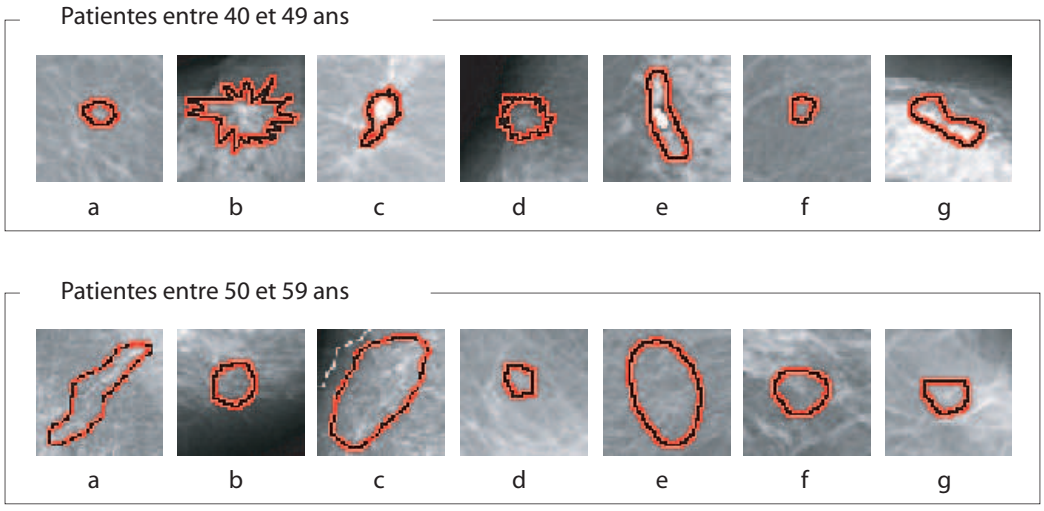


FIG. 6.14 – Exemple d'une agrégation OLAP dans dans le cube XML des données de mammographies

Cette agrégation classique, établie dans la phase conceptuelle du contexte d'analyse, ne fournit pas à l'expert des connaissances pertinentes sur les ressemblances des régions suspectes selon leur propriétés médicales. Dans le domaine du dépistage du cancer du sein, nous pouvons imaginer que les experts ont besoin d'extraire à partir de la population des régions suspectes des agrégats de régions ayant des propriétés médicales qui se ressemblent. Ces agrégats peuvent ainsi constituer une base d'apprentissage et une source de connaissances intéressantes pour les experts. Les experts peuvent éventuellement établir des associations sémantiques entre les propriétés médicales communes d'un agrégat et le type de pathologie le plus dominant dans cet agrégat.

De plus, dans une démarche préventive du cancer du sein, ce type de connaissance peut être d'une grande utilité. En effet, lors d'un premier examen clinique d'une nouvelle patiente, en exploitant les connaissances acquises dans la phase d'apprentissage, les propriétés médicales d'une nouvelle région suspecte peuvent donner une première idée sur le type de pathologie.

6.6.2 Module de connexion aux cubes XML

Comme nous l'avons déjà évoqué dans la section 6.2, nous avons inclus dans notre plateforme MiningCubes un module de connexion aux cubes de données XML. Ce module a pour objectif d'assurer la prise en compte des données complexes dans notre plateforme d'analyse.

Sur un plan conceptuel, le module de connexion aux cubes XML remplit les mêmes fonctionnalités que celles du module de connexion au serveur OLAP *Analysis Services* de MicroSoft SQL Server 2000. Cependant, sur le plan des traitements physiques ce module emploie plutôt le parseur DOM (*Document Object Model*) MSXML afin de lire le schéma XML qui représente le modèle logique du cube XML à étudier. Le parseur DOM est aussi employé pour charger les données à partir du schéma physique du cube de données, c'est-à-dire, à partir des documents XML du cube. Comme le montre la figure 6.15, dans notre plateforme MiningCubes, un formulaire est prévu pour charger le schéma XML du cube ainsi que ses documents XML.

6.6.3 Agrégation par classification dans le cube XML des données de mammographies

Nous proposons maintenant d'appliquer le module d'agrégation par classification de notre plateforme d'analyse MiningCubes sur le cube XML des données de mammographies. Supposons qu'un utilisateur cherche à agréger les radios des mammographies selon leurs ressemblances dans le cube. Les *images radios* (*Scanner_image*) des mammographies correspondent à l'ensemble des modalités $\mathcal{A}_{81}$ du premier niveau hiérarchique H_1^8 de la dimension D_8 du cube XML. Comme défini par la formule 4.3.1,

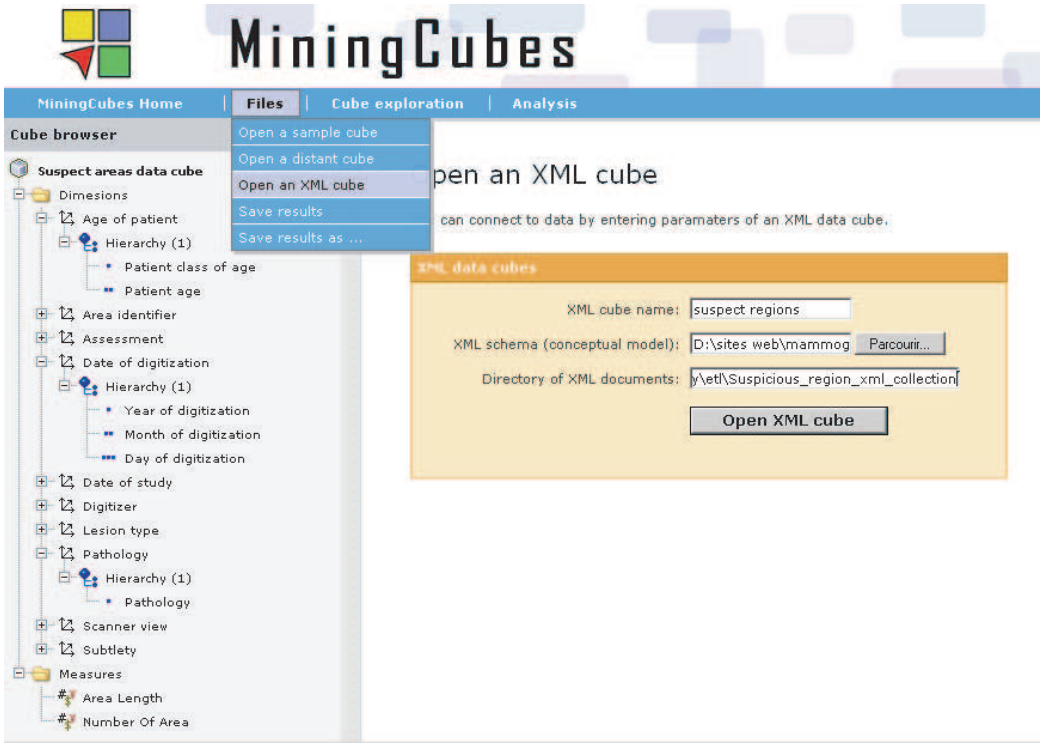


FIG. 6.15 – Chargement du cube XML des mammographies dans MiningCubes

les individus Ω de l'agrégation par classification peuvent correspondre à un sous-ensemble de modalités prises par l'utilisateur de l'ensemble $\mathcal{A}_{81}$. Supposons, dans notre cas d'application, qu'un utilisateur souhaite agréger 36 images radios de mammographies que nous résumons dans la figure 6.16.

Supposons aussi que l'utilisateur choisit d'agréger ces images radios selon leurs ressemblances en fonction des *catégories de lésions* des régions suspectes qu'elles contiennent et du *périmètre* de ces régions. Ainsi, l'utilisateur devra choisir des modalités du deuxième niveau hiérarchique H_2^1 de la dimension D_1 *type de la lésion* (*Lesion_type*) et la mesure M_1 correspondant au *périmètre* d'une région suspecte. Dans ce cas, selon la formule 4.3.4, l'ensemble des variables de la classification est :

$$\Sigma \subset \left\{ \begin{array}{c} V_t = M_1(\ a_i\ ,\ All\ , \ldots\ ,\ All\ ,\ \omega\ ,\ All\ ,\ All\ ) \\ \text{où } a_t \in \mathcal{A}_{12} \text{ et } \omega \in \Omega \end{array} \right\} \tag{6.6.1}$$

De manière plus explicite, selon les catégories de lésions que nous trouvons dans le cube XML des données de mammographies, les variables de l'ensemble Σ de la

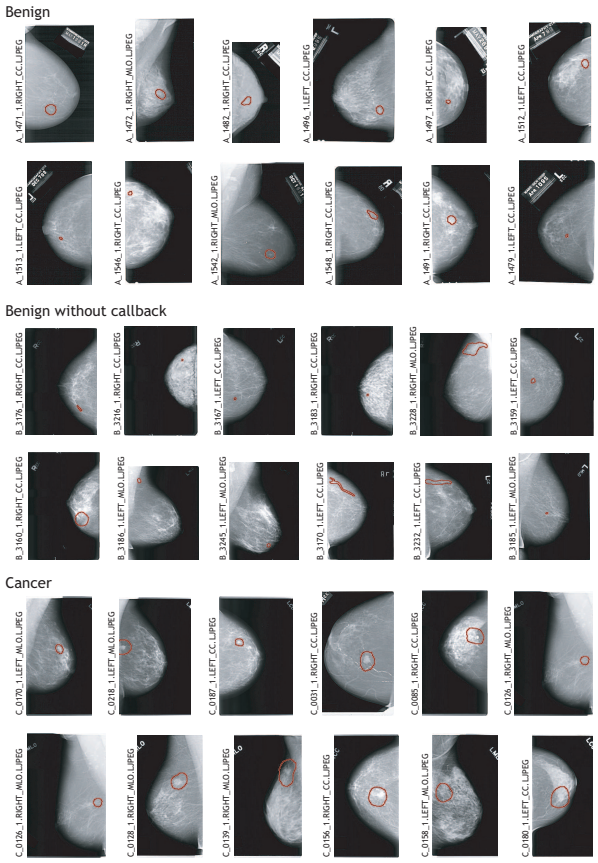


FIG. 6.16 – L'ensemble Ω des images radios de mammographies

classification peuvent être choisies par l'utilisateur parmi les 11 variables possibles suivantes :

- V_1 : *périmètre* des régions suspectes avec *calcification amorphe* ;
- V_2 : *périmètre* des régions suspectes avec *calcification de lucent* ;
- V_3 : *périmètre* des régions suspectes avec *calcification pléomorphe* ;
- V_4 : *périmètre* des régions suspectes avec *calcification de type punctate* ;
- V_5 : *périmètre* des régions suspectes avec *calcification de type peau* ;
- V_6 : *périmètre* des régions suspectes avec *calcification vasculaire* ;
- V_7 : *périmètre* des régions suspectes avec *forme de masse asymétrique* ;
- V_8 : *périmètre* des régions suspectes avec *forme de masse irrégulière* ;
- V_9 : *périmètre* des régions suspectes avec *forme de masse lobulé* ;
- V_{10} : *périmètre* des régions suspectes avec *forme de masse ovale* ;
- V_{11} : *périmètre* des régions suspectes avec *forme de masse ronde.*

En sélectionnant l'ensemble des 11 variables précédentes et en choisissant la *distance euclidienne* comme mesure de similarité entre individus et le critère de *Ward* comme distance entre groupes d'individus, nous obtenons une agrégation des 36 images radios de mammographies selon le dendrogramme de la figure 6.17.

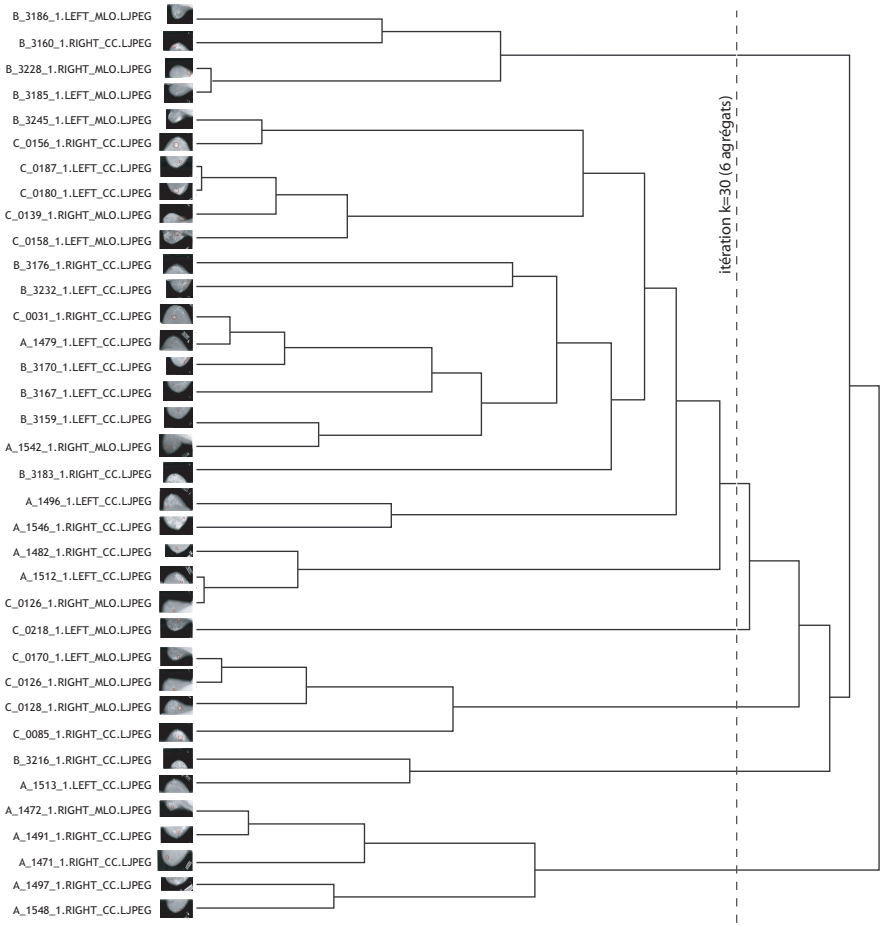


FIG. 6.17 – Dendrogramme de l'agrégation par classification dans le cube XML des données de mammographies

Notons que le dendrogramme obtenu n'est pas simple à interpréter. Un utilisateur est incapable, à partir de ce résultat, de décider du nombre d'agrégats qui répond au mieux à ses objectifs d'analyse. Il est nécessaire d'avoir recours à une aide supplémentaire afin de le guider l'utilisateur et de l'assister dans son choix. Pour cela, on met en œuvre les critères d'évaluation des agrégats que nous avons présentés dans la section 4.5.

6.6.4 Évaluation des partitions des agrégats

Le module d'agrégation par classification de MiningCubes fournit des représentations graphiques de la figure 6.18 pour les critères d'évaluation des partitions de la CAH.

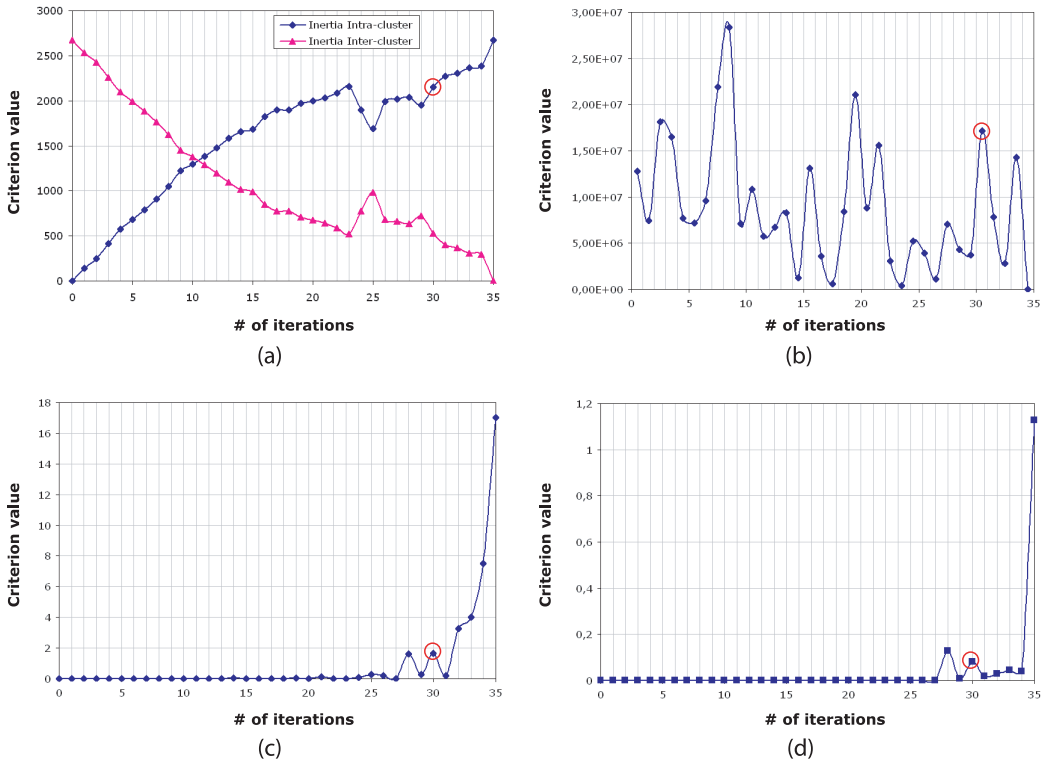


FIG. 6.18 – Représentations graphiques de (a) l'inertie intra et inter-classes, (b) la méthode de *Ward* (c) séparabilité des classes avec arêtes non pondérées et (d) séparabilité des classes avec arêtes pondérées.

À partir de ces représentations graphiques, nous notons que les trois critères réalisent des sauts remarquables pour les petits nombres d'agrégats. D'une manière générale, le choix d'une partition avec un très petit ou un très grand nombre d'agrégats n'apporte pas d'intérêt particulier à l'analyse de l'utilisateur. En effet, plus on se rapproche de la partition grossière ayant une seule classe, plus le nombre des agrégats est petit. Ces derniers ne reflètent pas de bonnes propriétés de séparabilité ou d'homogénéités internes. D'un autre côté, en se rapprochant de la partition la plus fine de la CAH, où chaque individu représente lui-même un agrégat, on risque d'avoir un très grand nombre d'agrégats qui peut nuire à l'interprétation.

Rappelons que, en plus des résultats des critères d'évaluation des agrégats, le choix

de la meilleure partition, parmi celles fournies par la CAH, dépend aussi des objectifs de l'analyse menée par l'utilisateur. Dans le cas des résultats obtenus à partir du cube XML des données de mammographies, supposons qu'un expert du cancer du sein cherche à trouver des agrégats de régions suspectes où chaque agrégat caractérise une pathologie particulière.

Rappelons que, selon le formalisme que nous avons développé au chapitre 4, l'itération k de la CAH correspond à une partition à $(n - k)$ agrégats. En tenant compte de ses objectifs d'analyse, un expert peut choisir l'itération $k = 30$, qui correspond à un nombre d'agrégat égal à 6. Ce choix peut aussi se justifier par les résultats des critères d'évaluation.

En effet, dans la courbe de la figure 6.18 (a), l'inertie intra-classe augmente d'une manière remarquable en passant de l'itération $k = 29$ à l'itération $k = 30$. Dans la figure 6.18 (b), le critère de la méthode de *Ward* enregistre aussi un pic quand la CAH passe de l'itération $k = 30$ à l'itération $k = 31$. Rappelons que le critère de la méthode de *Ward* consiste à minimiser l'inertie interne des agrégats formés par la CAH. Ainsi, l'expert est amené à préférer la partition de l'itération $k = 30$ à celle de l'itération $k = 31$. Quand aux deux critères basés sur la séparabilité des classes, représentés par les figures 6.18 (c) et 6.18 (d), nous détectons clairement des pics au niveau de leurs courbes à l'itération $k = 30$.

Rappelons aussi que, dans le contexte de l'analyse des régions suspectes dans les mammographies, selon la dimension D_4 du cube XML, les experts distinguent trois types de pathologies possibles : une pathologie *bénigne*, une pathologie *bénigne sans rappel* et une pathologie *maligne*. Dans les six agrégats obtenus par classification, un expert cherche également à faire émerger dans chaque agrégat une pathologie dominante en vue de déterminer les causes relatives à cette pathologie à partir des propriétés médicales. Dans notre cas d'application, les propriétés médicales correspondent aux variables choisies pour la classification, à savoir les catégories de lésions.

Dans le tableau 6.1, nous résumons l'ensemble des six agrégats des images radios obtenus et montrons le bilan des types de pathologies dans chaque agrégat. À partir de ce tableau, nous remarquons l'existence d'agrégats ayant des régions suspectes avec des pathologies homogènes. C'est le cas de l'*agrégat 1* et l'*agrégat 6* qui correspondent à 100% de pathologies bénignes. C'est aussi le cas de l'*agrégat 3* et l'*agrégat 4* qui correspondent à 100% de pathologies malignes. Ces résultats véhiculent des connaissances intéressantes et traduisent des similarités sémantiques entre les objets à l'intérieur de chaque agrégat. Évidement, le choix des variables de la classification est aussi important que le choix du nombre d'agrégats. En effet, plus ces variables sont corrélées avec les types de pathologies, plus nous avons de chances d'obtenir des agrégats avec des pathologies dominantes. Encore une fois, les connaissances de l'expert jouent un rôle important dans le choix des variables de la classification.

Agrégat 1 – 5 régions suspectes – 100% Bénignes
A_1548_1.RIGHT_CC.LJPEG, A_1497_1.RIGHT_CC.LJPEG, A_1471_1.RIGHT_CC.LJPEG, A_1491_1.RIGHT_CC.LJPEG, A_1472_1.RIGHT_MLO.LJPEG
Agrégat 2 – 2 régions suspectes – 50% Bénignes – 50% Bénignes sans rappel
A_1513_1.LEFT_CC.LJPEG, B_3216_1.RIGHT_CC.LJPEG
Agrégat 3 – 4 régions suspectes – 100% Malignes
C_0085_1.RIGHT_CC.LJPEG, C_0128_1.RIGHT_MLO.LJPEG, C_0126_1.RIGHT_MLO.LJPEG, C_0170_1.LEFT_MLO.LJPEG
Agrégat 4 – 1 régions suspectes – 100% Malignes
C_0218_1.LEFT_MLO.LJPEG
Agrégat 5 – 20 régions suspectes – 30% Bénignes – 35% Bénignes sans rappel – 35% Malignes
C_0126_1.RIGHT_MLO.LJPEG, A_1512_1.LEFT_CC.LJPEG, A_1482_1.RIGHT_CC.LJPEG, A_1546_1.RIGHT_CC.LJPEG, A_1496_1.LEFT_CC.LJPEG, B_3183_1.RIGHT_CC.LJPEG, A_1542_1.RIGHT_MLO.LJPEG, B_3159_1.LEFT_CC.LJPEG, B_3167_1.LEFT_CC.LJPEG, B_3170_1.LEFT_CC.LJPEG, A_1479_1.LEFT_CC.LJPEG, C_0031_1.RIGHT_CC.LJPEG, B_3232_1.LEFT_CC.LJPEG, B_3176_1.RIGHT_CC.LJPEG, C_0158_1.LEFT_MLO.LJPEG, C_0139_1.RIGHT_MLO.LJPEG, C_0180_1.LEFT_CC.LJPEG, C_0187_1.LEFT_CC.LJPEG, C_0156_1.RIGHT_CC.LJPEG, B_3245_1.LEFT_MLO.LJPEG
Agrégat 6 – 4 régions suspectes – 100% Bénignes
B_3185_1.LEFT_MLO.LJPEG, B_3228_1.RIGHT_MLO.LJPEG, B_3160_1.RIGHT_CC.LJPEG, B_3186_1.LEFT_MLO.LJPEG

TAB. 6.1 – Agrégats des images radios obtenus à l'itération $k = 30$

6.7 Expérimentations et performances

Nous avons mener des tests empiriques afin d'évaluer la performance du notre méthode d'agrégation par classification dans les cubes XML. Nous avons tiré plusieurs échantillons aléatoires à partir de la population des 4 686 documents XML du cube des données de mammographies. Rappelons que chaque document correspond à un fait à analyser dans le schéma physique du cube XML. Ainsi, en faisant varier le nombre de documents XML, on fait aussi varier la taille du cube. Ces expérimentations ont été réalisées dans le cadre de notre plateforme MiningCubes sous un environnement Windows XP sur une machine de 480MB de mémoire vive, un processeur Intel Pentium 4 avec une fréquence de 1,60GHz. La taille totale de la collection des 4 686 document XML est égale à 17,7Mo. Le but de ces expérimentations est de mesurer les temps de connexion et de chargement des cubes XML et le temps d'exécution de la CAH en fonction de la taille du cube XML étudié.

Ainsi, avec les différents échantillons de documents XML, nous avons tester les temps de réponse du parseur DOM du module de connexion aux cubes XML. La figure 6.19 montre les résultats obtenus suite à ces tests. La courbe de cette figure

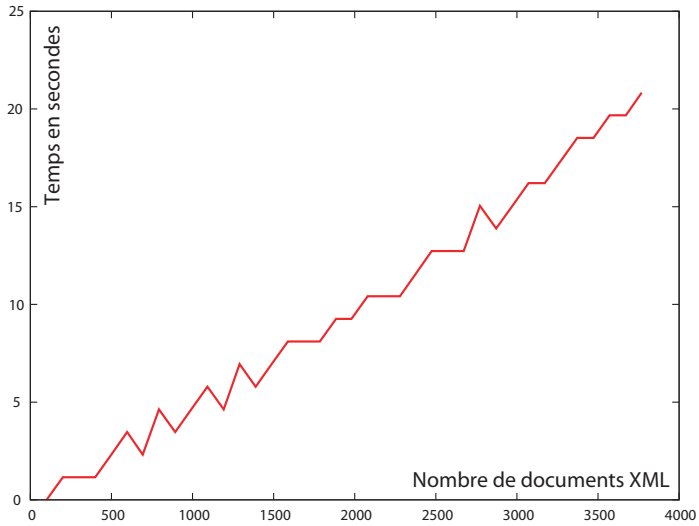


FIG. 6.19 – Temps de lecture du cube XML en fonction du nombre de documents XML

décrit un comportement quasi-linéaire du temps de réponse du parseur en fonction du nombre de documents XML. Ainsi, la performance du parseur DOM évolue d'une manière proportionnelle à la taille du cube XML. Notons que, étant donné que les tests ont été réalisés sur un serveur local (*localhost*), dans le cas d'une connexion à un cube XML distant, il faut aussi rajouter à ces temps de réponse les temps d'envoi des documents XML à travers le réseau.

Nous avons aussi tester les temps de réponse du module d'agrégation par classification en fonctions des échantillons des documents XML. Nous résumons les temps obtenus par le graphique de la figure 6.20. Nous remarquons que le temps d'exécution de la CAH décrit un comportement polynomial en fonction du nombre de documents XML. Rappelons que, dans notre cas d'application, les documents XML correspondent aux individus de la classification. Soit donc n le nombre d'individus à classifier. Comme nous l'avons évoqué dans la section 4.2, lors de la première itération de la CAH, il faut calculer toutes les agrégations possibles de deux individus parmi n. Ceci correspond à $n(n-1)/2$ possibilités.

Rappelons aussi que, dans notre méthode d'agrégation par classification, nous regroupons les modalités d'un niveau hiérarchique d'une dimension. Habituellement, dans un contexte OLAP on fait souvent affaire à des cubes de données avec un nombre de modalités relativement petit par dimension. Ainsi, la complexité polynomiale de la CAH ne pose pas de problèmes particuliers au niveau du temps de réponse de notre méthode.

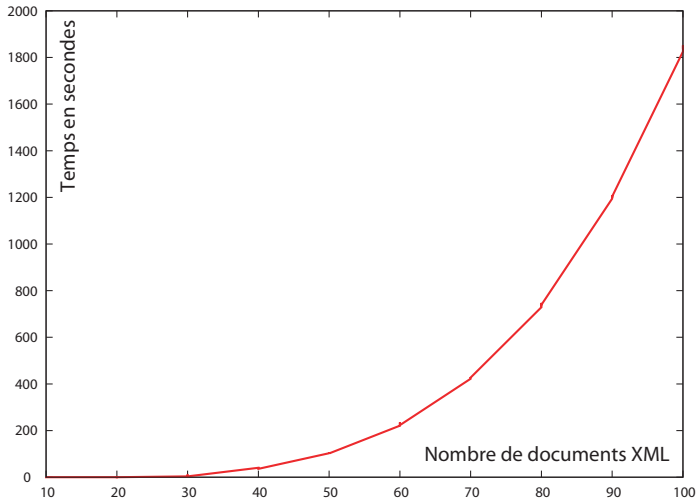


FIG. 6.20 – Temps d'exécution de la CAH en fonction du nombre de documents XML

6.8 Conclusion et perspectives

Dans ce chapitre, nous avons présenté notre application MiningCubes que nous avons développée afin de concrétiser nos propositions théoriques dans le cadre du couplage entre l'analyse en ligne et la fouille de données. Il s'agit d'une plateforme logicielle ouverte dédiée à l'analyse des données multidimensionnelles sous un environnement Web. Elle offre une interface conviviale et ergonomique assurant une interaction adéquate avec les utilisateurs.

En plus des cubes de données simples, nous avons également proposé d'étendre notre plateforme afin de prendre en compte les données complexes. Une donnée complexe est souvent caractérisée par des sources hétérogènes, des supports de différents types et des représentations diverses. Ainsi, nous avons fait le choix d'utiliser XML comme un formalisme pour structurer, homogénéiser et entreposer les données complexes. Pour cela, nous avons mis en place l'approche X-Warehousing. Il s'agit d'une démarche méthodologique, entièrement basée sur XML, capable de concevoir et de construire des cubes XML qui représentent des contextes d'analyse des données complexes. D'un autre côté, nous avons aussi mis en place dans notre plateforme d'analyse un module de connexion aux cubes XML. Ce module permet, par conséquent, d'exploiter les données complexes. dans un processus décisionnel.

Afin de valider notre démarche, nous avons proposé un cas d'étude sur un jeu de données complexes concernant le domaine du dépistage du cancer du sein. À partir de ce jeu de données, nous avons conçu et construit un cube XML de données de mammographies sur lequel nous avons ensuite appliqué notre méthode d'agrégation

par classification. Ce cas d'application nous a permis d'apprécier l'intérêt et la pertinence des résultats obtenus par rapport aux objectifs d'analyse. Une série de tests empiriques nous a permis aussi d'évaluer les performances de notre plateforme d'analyse appliquée aux données complexes.

Suite à ce travail, plusieurs perspectives d'ordre technique sont envisagées. Tout d'abord, nous pensons qu'il est nécessaire d'étendre nos deux autres propositions, à savoir la réorganisation des cubes de données par l'analyse des correspondances multiples et l'explication des cubes de données par règles d'association, au cas des données complexes.

Nous souhaitons également étendre notre plateforme afin de couvrir, en plus de l'analyse, un processus d'entreposage général basé sur XML. D'après notre méthodologie X-Warehousing, nous avons vu que le formalisme XML est une solution efficace pour l'entreposage des données complexes. XML reste aussi une solution valable pour les données simples. D'une manière générale, nous pensons que dans un avenir proche, le formalisme XML va devenir un standard à part entière dans le domaine des entrepôts de données.

Enfin, à l'heure actuelle, notre plateforme MiningCubes évolue sur un serveur local dans un cadre expérimental. Nous projetons la mise en ligne de la plateforme. Nous espérons faire de MiningCubes un fournisseur de services Web en matière d'entreposage et d'analyse en ligne associée à la fouille de données. En plus, nous projetons de faire de MiningCubes un projet ouvert dans lequel il est possible d'accéder aux codes sources et de rajouter de nouveaux modules d'analyse en ligne et/ou de fouille de données.

Vers un cadre formel général

Résumé

Dans ce chapitre, nous engageons des réflexions pour la mise en place d'un cadre formel général dédié au couplage de l'analyse en ligne et de la fouille de données. À cet effet, nous formulons un modèle multidimensionnel et une algèbre OLAP. Notre algèbre repose sur un noyau minimal fermé incluant des opérations de structuration et de navigation.

En se basant sur nos précédentes contributions, nous fournissons une première proposition pour étendre le noyau minimal de notre algèbre à des opérateurs de fouille de données en ligne. Nous souhaitons généraliser cette proposition afin d'établir une base théorique générale servant à la formalisation de toute sorte d'approches d'analyse en ligne couplée avec la fouille de données.

Sommaire