Recherche de cas médicaux multimodaux à l'aide d'arbres de décision

Gwénolé Quellec, Mathieu Lamard, Lynda Bekri, Guy Cazuguel, Béatrice Cochener, Christian Roux

Recherche de cas médicaux multimodaux à l'aide d'arbres de décision
Multimodal medical case retrieval using decision trees

G. Quellec[a,c,*], M. Lamard[b,c], L. Bekri[b,c,d], G. Cazuguel[a,c], B. Cochener[b,c,d], C. Roux[a,c]

[a] ENST Bretagne, GET-ENST, Brest, F-29200 France

[b] Univ Bretagne Occidentale, Brest, F-29200 France

[c] Inserm, U650, Brest, F-29200 France

[d] CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France

Résumé

Nous proposons dans cet article un système de raisonnement à base de cas pour la recherche de dossiers patients similaires à un cas donné en requête. Nous nous intéressons à des dossiers patients constitués de plusieurs images accompagnées d'informations contextuelles (telles que l'âge, le sexe et les antécédents médicaux du patient). Plusieurs sources d'informations (parfois incomplètes) sont en effet généralement nécessaires pour diagnostiquer une pathologie. Nous proposons un outil de recherche basé sur les arbres de décision, qui sont bien adaptés pour traiter de l'information hétérogène et incomplète. Pour prendre en compte des images dans ce système, nous leur associons une signature définie à partir de leur contenu numérique. La méthode est évaluée sur une base de rétinopathies diabétiques classifiées. Sur cette base de données, les résultats sont intéressants : la précision atteint 79,5% pour une fenêtre de retrouvaille de 5 cas, doublant pratiquement les résultats obtenus en n'utilisant que le contenu numérique d'une image.

Abstract

In this article, we present a Case-based Reasoning system for the retrieval of patient files similar to a case placed as query. We focus on patient files made up of several images with contextual information (such as the patient age, sex and medical history). Indeed, medical experts generally need varied sources of information (which might be incomplete) to diagnose a pathology. Consequently, we derive a retrieval framework from decision trees, which are well suited to process heterogeneous and incomplete information. To be integrated in the system, images are indexed by their digital content. The method is evaluated on a classified diabetic retinopathy database. On this database, results are promising: the retrieval sensitivity reaches 79.5% for a window of 5 cases, which is almost twice as good as the retrieval of single images alone.

Mots clés : arbres de décision ; indexation d'images ; information contextuelle ; information incomplète; rétinopathie diabétique

Keywords: contextual information; decision trees; diabetic retinopathy; image indexing; incomplete information

* Auteur correspondant. Adresse e-mail : gwenole.quellec@enst-bretagne.fr (G. Quellec).

1. Introduction

En médecine, la connaissance des experts résulte des connaissances encyclopédiques et d'expériences à travers des cas cliniques réels. Aussi, le raisonnement à base de cas (RBC ou *Case-based reasoning* - CBR) [1], présenté aux débuts des années 1980, connaît-il un intérêt croissant pour le développement de systèmes d'aide à la décision médicale [2]. L'idée sous-jacente du RBC est que des problèmes proches ont des solutions voisines, une idée confirmée par l'expérience des médecins. Dans le RBC, l'interprétation d'une nouvelle situation passe par la recherche, dans une base de cas, de cas similaires au cas en question. L'interprétation de ces cas connus est ensuite adaptée au cas étudié.

A l'origine, le RBC à été mis au point pour traiter des cas structurés, typiquement des vecteurs de paramètres, textuels ou numériques. Des informations plus complexes telles que des images ne sont pas prises en compte par les systèmes standards. Or les médecins se basent sur des sources d'information très variées pour établir un diagnostic. Ainsi, pour les examens de Rétinopathie Diabétique (RD), les ophtalmologistes analysent des images multimodales, conjointement à de l'information structurée : l'âge du patient, son sexe, ses antécédents médicaux, etc.

Les méthodes d'indexation des images par leur contenu numérique, utilisées pour la recherche d'images en associant des index numériques à des images (*Content-based Image Retrieval* - CBIR [3]), permettent de mesurer la similarité entre deux images et ainsi étendre le RBC à des cas contenant des images. Ces méthodes sont particulièrement intéressantes pour leur objectivité et leur reproductibilité.

La définition d'un système RBC est ici conditionnée à la résolution de plusieurs problèmes, en particulier l'agrégation d'attributs hétérogènes (des images, des attributs nominaux ou continus), et la gestion de l'information manquante. Nous proposons de résoudre simultanément ces différents problèmes en définissant une mesure de similarité entre deux cas cliniques, à partir d'arbres de décision [4-5], généralement utilisés pour des problèmes de classification. Le système proposé est générique et applicable à toute sorte de base de données. Nous l'avons appliqué à la rétinopathie diabétique.

Après une présentation de la base de données (§2), nous préciserons les objectifs du système (§3). Nous décomposerons la présentation du système en trois parties : tout d'abord la présentation des descripteurs de dossiers patients (§4.1), puis l'apprentissage du système (§4.2) et enfin son utilisation (§4.3). Nous terminerons par les résultats et les conclusions (§5).

2. La base de rétinopathie diabétique

La base de rétinopathies diabétiques, développée spécifiquement pour l'étude, contient des images rétiniennes de patients diabétiques, associées à des informations anonymes sur la pathologie.

Le diabète est un trouble métabolique caractérisé par un excès constant de sucre dans le sang. Cet excès affecte progressivement les vaisseaux sanguins dans plusieurs organes (atteinte des tissus, rétrécissement), ce qui peut entraîner de sérieuses complications rénales, cardiovasculaires, cérébrales et également rétiniennes. Différentes lésions apparaissent sur les vaisseaux endommagés, pouvant entraîner la cécité.

Notre base de données comprend actuellement 63 dossiers patients, contenant au total 1045 photographies. Les patients sont tous diabétiques, ils ont été vus en consultation au centre hospitalier universitaire de Brest. Les dossiers utilisés dans l'étude sont les 63 premiers, par ordre chronologique de la consultation. Les informations cliniques associées aux images sont saisies à partir du dossier patient de l'hôpital.

Les images de la rétine ont été acquises par un ophtalmologiste, à l'aide d'un appareil numérique *Topcon Retinal Digital Camera* (TRC-50IA) connecté à un ordinateur (voir figure 1). Elles ont une définition de 1280 pixels par ligne pour 1008 lignes par image et sont compressées sans perte. Pour mieux identifier les différentes lésions, les images sont acquises dans différentes modalités : les rétinophotos (de simples photographies couleur), des photographies obtenues par application d'un filtre vert (anérythre) et d'un filtre bleu, ainsi que les séries angiographiques. Pour obtenir une série angiographique, un produit de contraste (la fluorescéine) est injecté dans le système sanguin du patient et des photographies de la rétine sont acquises à trois temps différents, afin d'étudier la circulation sanguine dans la rétine. En sus des images du pôle postérieur, quatre photographies sont acquises sur la périphérie de la rétine au temps intermédiaire. Ainsi les dossiers patients sont constitués idéalement de 10 images par œil (voir figure 1).

Pour chaque patient, un expert décrit les lésions présentes dans les images et en déduit le niveau de sévérité ICDRS [6] de la pathologie, sur une échelle de 0 à 5, 0 correspondant à l'absence de lésions. La distribution des niveaux de sévérité parmi les 63 patients de la base de données est fournie dans le tableau I.

3. Objectifs

3.1. Technique d'examen

Afin de déterminer le niveau de sévérité de la RD, l'ophtalmologiste doit acquérir un ensemble de photographies multimodales (voir figure 1) et les analyser de manière différentielle. Le niveau de sévérité est ensuite déterminé en fonction du nombre, du type et de la localisation de ces lésions. L'ophtalmologiste s'aide également du contexte médical du patient pour établir son diagnostic. L'analyse et l'interprétation des images est coûteuse en temps, surtout lorsque la qualité des images est mauvaise. Chaque niveau de sévérité entraîne une procédure spécifique (traitements, fréquence des contrôles, etc.).

3.2. Système d'aide au diagnostic proposé

Afin de diminuer le temps d'interprétation des images, notamment pour les cas difficiles, nous proposons aux médecins un système d'aide au diagnostic de la RD. Plusieurs

études ont été menées pour analyser automatiquement un dossier patient en segmentant les lésions caractéristiques de la RD dans les images et utiliser les résultats de segmentation pour calculer le niveau de sévérité de la pathologie [7]. Cette solution présente cependant des inconvénients : 1) compte tenu du nombre de types de lésions (9) et de modalités d'images (6), il faut paramétrer un nombre important d'algorithmes de détection, 2) ces algorithmes de détection sont généralement sensibles aux variations de qualité des images, ainsi qu'au changement de matériel. Nous proposons une solution alternative : nous caractérisons les images par une méthode générique, basée sur leur contenu numérique, et nous utilisons un algorithme de fouille de données pour apprendre le lien entre le niveau de sévérité et ces images, accompagnées d'informations contextuelles. Nous cherchons et proposons ensuite à l'ophtalmologiste des cas semblables à celui qu'il étudie. Il peut donc s'aider des annotations et du diagnostic effectués par d'autres experts sur des cas similaires pour déterminer lui-même le niveau de sévérité du cas qu'il étudie. Le nombre de dossiers proposés par le système pour une requête est fixé à cinq, à la demande des médecins. Ils jugent en effet ce nombre suffisant, en particulier pour des raisons de temps et au vu des résultats fournis par le système.

4. Le moteur de recherche

4.1. Les descripteurs de cas

Chaque cas est décrit par des informations sémantiques et des descripteurs numériques extraits des images.

Les informations cliniques disponibles sont l'âge et le sexe du patient, plus de l'information spécifique au contexte de la RD (voir tableau II), soit au total 13 descripteurs.

Pour extraire des descripteurs à partir des images, nous nous sommes basés sur des travaux antérieurs [8] : nous proposons de calculer une signature pour chaque image (c'est à dire un vecteur de paramètres synthétisant l'information qu'elles contiennent) à partir de leur transformée en ondelettes. Ces signatures modélisent la distribution des coefficients de la transformée en ondelettes dans chaque sous-bande de la décomposition, fournissant ainsi une description multiéchelle des images. Ces signatures peuvent être calculées à partir des images compressées par la méthode JPEG2000 [9] (sans décompression) ce qui peut s'avérer pratique lorsqu'un nombre important d'images doit être traité. Idéalement, nous calculons une signature pour chacun des 10 types d'images (c'est-à-dire chaque modalité d'acquisition et chaque localisation dans la rétine : centre ou périphéries).

Le nombre de microanévrysmes (un type de lésions de la RD) est également utilisé comme descripteur, car les microanévrysmes sont généralement les premières lésions de la RD à apparaître. Leur détection permet donc de différencier efficacement les patients sains des patients atteints de RD ; des algorithmes de détection robustes sont disponibles [10].

Les vecteurs de descripteurs ainsi définis sont incomplets. En effet, 40,5% des descripteurs cliniques sont manquants (les moins renseignés étant le contexte clinique familial et le contexte clinique chirurgical). De plus, sur les 10 photographies pouvant être acquises par œil, en moyenne 12,1% ne le sont pas. Les images les plus fréquemment absentes sont celles obtenues à l'aide d'un filtre bleu ; ce sont les moins utiles dans le cas de la RD, car les lésions y sont peu visibles. Pour les séries angiographiques, elles sont généralement soit complètes soit totalement absentes. Finalement, le nombre d'images par dossier est très variable, ainsi cinq dossiers patients dans la base ne contiennent qu'une photographie.

4.2. Construction du moteur de recherche

4.2.1. Les arbres de décision

Les arbres de décision [4-5] sont des outils d'aide à la décision. Ils sont constitués d'un ensemble de règles permettant de diviser une population de cas en groupes homogènes. Chaque règle associe une conjonction de tests sur les descripteurs d'un cas à un groupe (par exemple : « si sexe=homme et âge<40 alors le cas appartient au groupe 3 »). Ces règles sont organisées sous la forme d'un arbre dont la structure a la signification suivante :

- chaque nœud non terminal correspond à un test sur un descripteur (par exemple : « sexe = ? »),
- chaque arc correspond à une réponse à un test (par exemple : « homme »),
- chaque feuille correspond à un groupe de cas ayant fourni une réponse identique à tous les tests d'une règle (exemple : « les hommes de moins de 40 ans »)

Un exemple simple d'arbre est présenté sur la figure 2 (a) pour illustrer le principe.

Les arbres de décision ont d'abord été conçus pour traiter des vecteurs de descripteurs nominaux (les cas sont groupés par valeurs ou groupes de valeurs des descripteurs). Quinlan [4] les a étendus à des descripteurs continus (les cas sont groupés par plages de valeur du descripteur). Plus généralement, les arbres de décision peuvent traiter tout type de descripteur, pourvu qu'une méthode soit disponible pour grouper les cas en fonction du descripteur. Etant donné que chaque test est appliqué à un seul descripteur à la fois, les arbres de décision sont bien adaptés pour traiter des cas hétérogènes.

Pour mesurer la similarité entre deux dossiers médicaux, nous proposons de comparer les feuilles auxquelles ces deux dossiers sont affectés. Le calcul de cette mesure de similarité est détaillé au paragraphe 4.3.

4.2.2. Construction d'un arbre de décision

Pour construire un arbre de décision, les cas de la base sont répartis en trois ensembles :

- un ensemble d'apprentissage, noté A , utilisé pour rechercher le test à effectuer en chaque nœud,
- un ensemble de validation, noté V , utilisé pour décider quand arrêter la construction de l'arbre, afin d'éviter le surapprentissage,
- un ensemble de test, noté T , qui n'est pas utilisé pour construire l'arbre (il est conservé pour évaluer les performances du système)

Bien que nous ne cherchons pas à construire un classifieur, nous devons attribuer une classe à chaque cas des trois ensembles A , V et T lors de la constitution de la base, pour l'apprentissage. C'est en effet nécessaire pour caractériser

l'homogénéité des groupes de cas. La classification utilisée est le niveau de sévérité de la RD.

Le mécanisme de construction est décrit ci-dessous et illustré sur la figure 2(b). A l'initialisation de l'apprentissage, l'arbre est constitué simplement d'une feuille, regroupant l'intégralité de l'ensemble d'apprentissage (A). Puis, récursivement, chaque feuille F de l'arbre en construction est divisée. Pour cela, nous recherchons le descripteur d le plus discriminant au sein de la population $P \subset A$ regroupée dans F . P est alors répartie entre de nouveaux nœuds fils, un pour chaque réponse possible au test sur d . Dans la méthode proposée, nous avons mesuré le pouvoir discriminant d'un test par le gain G d'entropie de Shannon. G est le gain d'entropie obtenu lorsqu'un nœud v_0 est divisé en nœuds fils $v_n, n = 1..N$ (algorithme c4.5 [4], voir équation 1).

$$\begin{cases} G = \left(\sum_{n=1}^{N} I^n \right) - I^0 \\ I^n = -\sum_{c=1}^{C} p_c \log p_c, n = 0..N \end{cases} \quad (1)$$

où I^0 est l'entropie calculée dans le nœud parent v_0 (avant la division) et I^n l'entropie calculée dans le $n^{\text{ème}}$ nœud fils v_n , p_c désigne le pourcentage de cas de la classe c ($c = 1..C$) dans un nœud donné. Nous donnons sur la figure 3 le gain d'entropie calculé pour chacun des descripteurs sur l'ensemble de la population. Pour l'exemple de la figure 2(b), le critère qui maximise G sur l'ensemble de la population est le sexe, nous créons donc deux nœuds fils, un pour les hommes et un pour les femmes. Si aucun test n'est suffisamment discriminant ou si la population P est trop faible, alors P n'est pas divisée.

Pendant la construction de l'arbre, il faut éviter le surapprentissage, défini comme l'extraction de règles spécifiques à la base d'apprentissage, mais qui ne sont pas générales, en particulier qui ne s'appliquent pas à la base de validation. Pour ce faire, nous nous inspirons d'une méthode proposée dans l'algorithme CART [5] : tout au long du processus, les exemples de validation sont eux aussi répartis entre les différentes feuilles de l'arbre. Si au sein d'une feuille, la classe majoritaire diffère entre la base d'apprentissage et la base de validation, alors cette feuille ne sera pas divisée, afin d'éviter le surapprentissage.

Notre algorithme peut gérer l'information manquante : un mécanisme simple est fourni par l'algorithme c4.5. Supposons que la valeur du descripteur d , testé au nœud v_0 , soit manquante pour un cas donné. Alors ce cas est affecté à chaque nœud fils v_n avec un poids $p(v_n) \in [0;1]$. $p(v_n)$ est la proportion d'exemples de la base d'apprentissage, dont la valeur pour d est connue, assignée à v_n . Ainsi, à la fin de l'apprentissage, chaque exemple d'apprentissage c_i est affecté à chaque feuille f_j ($j = 1..M$) avec un degré p_{ij} tel que $\sum_{j=1}^{M} p_{ij} = 1$ (p_{ij} = 0 ou 1 si tous les descripteurs testés sont connus pour i , $p_{ij} \in [0;1]$ sinon).

En sortie de l'algorithme, nous obtenons donc un graphe dont la structure est un arbre et dont chaque nœud v est pondéré par $p(v)$. En moyenne, les arbres construits par cette procédure sont constitués d'une trentaine de nœuds.

4.2.3. Intégration d'images dans un arbre de décision

L'intégration d'images dans un arbre de décision est inspirée de la recherche d'images par leur contenu numérique [3], elle implique de : 1) construire une signature numérique de chaque image, et 2) définir une mesure de distance D entre deux signatures. Ainsi, mesurer la distance entre deux images revient à mesurer la distance entre deux signatures. Les signatures utilisées dans ce travail ont été rappelées dans le paragraphe 4.1.

Nous utilisons un principe analogue pour répartir un ensemble d'images en groupes d'images similaires dans un arbre de décision : la notion de groupe d'images est définie au sens de la mesure de distance D entre signatures. Cette mesure est paramétrée par un ensemble de poids. Nous ajustons ces poids de telle sorte que D soit mieux corrélée avec le niveau de sévérité [8]. Pour construire ces groupes, de manière non supervisée, nous utilisons l'algorithme FCM (*Fuzzy C-Means*) [11] ; nous remplaçons simplement dans cet algorithme la distance Euclidienne par D .

Ainsi, lors de l'apprentissage, pour tester le pouvoir discriminant d'un type d'images au sein d'une population de cas P , nous appliquons l'algorithme FCM. En sortie de cet algorithme, P est divisée en plusieurs groupes. Nous calculons alors le gain d'entropie entraîné par cette division. Si le gain est supérieur au gain des autres descripteurs, alors nous divisons la feuille associée à P en créant un nœud fils pour chacun de ces groupes. Ainsi, sur l'exemple de la figure 2 (b), les signatures d'images angiographiques tardives de la population des femmes ont été séparées par FCM en deux groupes, G1 et G2, et un nœud fils a été créé pour chacun d'entre eux.

4.2.4. Amélioration de la robustesse du système

Un arbre de décision représente un sous-ensemble des règles pertinentes associant les descripteurs de cas et la classe (le niveau de sévérité de la RD). Ainsi, un arbre de décision fournit une division de la population en groupes homogènes qui n'est pas unique. Or nous nous basons sur les groupes auxquels deux cas sont affectés pour définir leur similarité. Nous proposons donc de construire plusieurs arbres de décision afin de comparer plus finement les groupes auxquels ces cas sont affectés, et ainsi avoir une mesure de similarité plus précise.

Quelques méthodes ont été proposées dans la littérature pour générer des ensembles d'arbres : *Random Forests* [12] ou *Randomized c4.5* [13] notamment. Leurs performances en tant que classifieurs sont généralement meilleures que celles d'arbres de décision seuls. Pour générer des ensembles d'arbres, nous avons rendu aléatoire l'algorithme d'apprentissage présenté au paragraphe 4.2.2. Nous l'avons

modifié de la manière suivante : pour choisir le test à appliquer en un nœud, nous trions les descripteurs du plus discriminant au moins discriminant en fonction du gain d'entropie, puis nous en sélectionnons un aléatoirement parmi les k premiers. Un ensemble de N arbres est construit par cette procédure.

Pour améliorer les performances du système, nous sélectionnons un sous-ensemble des arbres générés : un pourcentage prédéfini α de ces arbres est conservé. Pour cela, les N arbres sont d'abord évalués individuellement : le score individuel de chaque arbre est défini comme la précision moyenne sur la base de validation ; la précision étant le pourcentage de cas retournés par le système ayant la même classe que le cas placé en requête. Nous conservons ensuite les arbres qui obtiennent les meilleurs scores individuels.

L'efficacité de l'ensemble d'arbres est quant à elle définie comme la précision moyenne sur la base de test. Les paramètres suivants sont calibrés afin de maximiser ce critère :

- le nombre $p^{(1)} = N$ d'arbres générés
- le pourcentage $p^{(2)} = \alpha$ d'arbres sélectionnés
- le paramètre d'aléa $p^{(3)} = k$ de la génération d'arbres
- le paramètre de l'algorithme FCM $p^{(4)} = m$ (le coefficient de flou [11])

Un ensemble discret de valeurs $p^{(i)} = \left\{p_1^{(i)}, p_2^{(i)}, ...\right\}$ est évalué pour chaque paramètre $p^{(i)}$, et le meilleur élément de l'espace produit $P^{(1)} \times P^{(2)} \times P^{(3)} \times P^{(4)}$ est sélectionné. Ces éléments sont sélectionnés par validation croisée : l'expérience est répétée plusieurs fois avec différentes bases d'apprentissage, de validation et de test, sélectionnées aléatoirement. La proportion d'exemples affectés à chaque base est la suivante : 70% pour A, 20% pour V, 10% pour T.

4.2.5. Système obtenu

Le meilleur ensemble de paramètres trouvé pour le système est le suivant :

- $N = 200$ arbres générés
- $\alpha = 20\,\%$ d'arbres sélectionnés
- paramètre d'aléa $k = 6$
- coefficient de flou $m = 2$.

En conclusion, à la fin de l'apprentissage, nous avons construit un ensemble de $N' = \alpha N = 40$ arbres qui constituent la structure de notre moteur de recherche. Son utilisation est présentée ci-dessous.

4.3. Utilisation du moteur de recherche

Soit un cas c_r présenté en requête au système. Nous souhaitons trouver les cinq cas de la base de données les plus proches de c_r. Le principe de la recherche est d'abord illustré sur un système constitué d'un seul arbre (voir figure 4), puis étendu à un système à N' arbres.

4.3.1. Moteur de recherche basé sur un arbre

Pour mesurer la similarité entre c_r et un cas c_i de la base, nous comparons leur degré d'affectation aux différentes feuilles f_j de l'arbre de décision, $j = 1..M$, respectivement p_{ij} et p_{rj} (voir §4.2.2). Le poids $p(v)$ de chaque nœud v de l'arbre est utilisé pour calculer ces degrés d'affectation. Pour déterminer $\left(p_{rj}\right)_{j=1..M}$, nous avons besoin de calculer la signature des images dont le type est testé dans l'arbre, si elles sont disponibles. Ainsi, dans l'exemple de la figure 4, seule la signature de l'image angiographique tardive a besoin d'être calculée.

Nous avons défini une mesure de similarité S_{ab} entre deux cas c_a et c_b à partir de leur vecteur de degrés d'affectation, $\left(p_{aj}\right)_{j=1..M}$ et $\left(p_{bj}\right)_{j=1..M}$ respectivement (équation 2).

$$S_{ab} = \sum_{j=1}^{M} p_{aj}\, p_{bj} \qquad (2)$$

Cette mesure, le produit scalaire des deux vecteurs, est à valeurs dans $[0;1]$. Elle vaut 1 quand les deux cas sont entièrement affectés au même groupe. Elle est nulle s'il n'existe pas de groupe auquel les cas sont tous deux partiellement affectés.

Cette mesure de similarité peut être calculée rapidement entre la requête c_r et chaque cas c_i de la base, il n'est pas nécessaire de parcourir toute la base de données :

- Pour chaque feuille f_j de l'arbre, nous construisons préalablement la liste L_j des cas c_i tels que $p_{ij} \neq 0$. Cette liste peut être construite pendant l'apprentissage de l'arbre et mise à jour lorsqu'un nouveau cas est ajouté à la base
- La mesure de similarité S_{ri} est initialisée à 0, pour tout cas c_i.
- Pour chaque feuille f_j telle que $p_{rj} \neq 0$, nous parcourons la liste L_j : pour chaque cas c_i dans L_j, S_{ri} est incrémenté de $p_{ri}\, p_{ij}$.

4.3.2. Moteur de recherche basé sur plusieurs arbres

Considérons maintenant le système entier, constitué de N' arbres. La mesure de similarité globale S'_{ab} entre deux cas c_a et c_b est simplement la somme des similarités calculées séparément pour chacun des arbres par la procédure rapide

présentée ci-dessus. Soit M_n le nombre de feuilles du $n^{ième}$ arbre et p_{anj} (resp. p_{bnj}) le poids d'affectation du cas c_a (resp. c_b) à la feuille f_i de l'arbre n. S'_{ab} est donc définie par l'équation 3.

$$S'_{ab} = \sum_{n=1}^{N'} \sum_{j=1}^{M_n} p_{anj}\, p_{bnj} \qquad (3)$$

Finalement, le système propose les cinq cas $c_i, i = 1..5$, qui maximisent S'_{ri}.

4.3.3. Résultats

La précision moyenne obtenue par le système sur la base de test atteint **79,5%**. Concrètement, cela veut dire que quatre dossiers sur les cinq trouvés par le système présentent le même niveau de sévérité de la RD que le dossier passé en requête. Pour estimer la contribution de l'information numérique (signatures d'images) d'une part, et de l'information sémantique d'autre part, à la précision moyenne, des ensembles d'arbres sont construits en utilisant uniquement l'un des deux types d'information. Les scores de précision sont donnés sur la figure 5. Pour évaluer l'intérêt des arbres de décision pour la recherche de cas hétérogènes et incomplets, la méthode proposée a été comparée à une combinaison linéaire de mesures de distances hétérogènes, avec gestion des valeurs manquantes [14]. Nous avons choisi cette méthode comme référence car c'est la généralisation naturelle du raisonnement à base de cas (RBC) à notre problématique. Nous l'avons étendue à des cas contenant des images en utilisant la mesure de distance entre signatures (voir §4.2.3). Une précision moyenne de 52,3% a été atteinte par cette méthode.

5. Conclusion et discussion

Nous avons présenté dans cet article un système de raisonnement à base de cas, qui s'appuie sur les arbres de décision, pour la recherche de cas cliniques similaires à un cas donné en requête. Au lieu de les utiliser pour classifier les cas, nous définissons une mesure de similitude entre les dossiers patients et le dossier requête, en fonction de leur situation dans les arbres de décision. Nous avons introduit une méthode pour intégrer des images et leurs signatures numériques, avec de l'information contextuelle, dans les arbres de décision. Le système exploite la capacité des arbres de décision à combiner de l'information hétérogène, en particulier des signatures d'images. L'algorithme de recherche exploite également la capacité des arbres de décision à gérer les valeurs manquantes et à éviter le surapprentissage. Cette dernière propriété garantit que le système de recherche est adapté aussi bien à des bases de grande taille qu'à des bases de petite taille, comme celle étudiée. L'algorithme utilisé pour la construction des arbres est basé sur c4.5. Cet algorithme a été préféré à l'algorithme classique CART [5], car il permet la gestion des valeurs manquantes. De plus, le critère de sélection des tests proposé par c4.5 (le gain d'entropie) s'est avéré plus intéressant que celui proposé dans CART (le critère de Gini), du point de vue de la précision du moteur de recherche construit. En revanche, le mécanisme de gestion du surapprentissage proposé dans CART (l'utilisation d'une base de validation) s'est avéré plus intéressant que celui proposé dans c4.5 (l'élagage).

La précision obtenue par le moteur de recherche sur la base de données de RD (79,5%) est très satisfaisante, compte tenu notamment du faible nombre d'exemples à disposition pour l'apprentissage, du fort taux de valeurs manquantes et du nombre important de classes (6) à considérer. Par comparaison, l'algorithme de CBIR basé sur la recherche d'images seules [8], d'où proviennent les signatures d'images utilisées, fournit une précision de 46,1%. Comme le montre la figure 5, en utilisant plusieurs images, au lieu d'un image seule, la précision augmente d'un facteur 1,45 (66,7%/46,1%). L'ajout d'information clinique apporte également une augmentation importante de la précision (d'un facteur 1,19 : 79,5%/66,7%). Les résultats obtenus sont logiques d'un point de vue clinique : une image seule n'est généralement pas suffisante pour qu'un médecin puisse correctement diagnostiquer le niveau de sévérité de la RD chez un patient ; la figure 3 confirme néanmoins que les images sont des descripteurs discriminants, notamment les images obtenues à l'aide d'un filtre vert anérythre (voir figure 2 (b)) et les images angiographiques, qui sont les images les plus utiles pour les médecins, dans le cadre de la RD. Ils confirment que toutes les sources d'informations sont utiles pour correctement interpréter un dossier patient. Le choix du système utilisé pour fusionner ces informations joue cependant un rôle important : la différence de précision entre le système proposé et une simple combinaison linéaire de distances hétérogènes en atteste (79,5% dans un cas, 52,3% dans l'autre).

La méthode proposée est en outre intéressante pour sa généricité : toute base multimédia peut être traitée, pourvu qu'une procédure pour regrouper des cas similaires soit fournie pour chaque nouvelle modalité (son, vidéo, ...).

Enfin, cet article ouvre des perspectives encourageantes pour associer de l'information numérique (non structurée) et sémantique (structurée) dans des problématiques de fouille de données.

Références

[1] Aamodt A. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 1994;7(1):39-59.

[2] Bichindaritz I, Marling C. Case-based reasoning in the health sciences: what's next?, Artificial Intelligence in Medecine 2006;36(2):127-135.

[3] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(12):1349-1380.

[4] Quinlan JR, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers ; 1993.

[5] Breiman L, Friedman JH, Olshen RA, Stone CJ, Classification and Regression Trees. Chapman & Hall/CRC ; 1984.

[6] Wilkinson CP, Ferris FL, Klein RE, Lee PP, Agardh CD, Davis M et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales, Ophthalmology 2003;110(9):1677-1682.

[7] Teng T, Lefley M, Claremont D. Progress towards automated diabetic ocular screening: A review of image analysis and intelligent systems for diabetic retinopathy, Medical and Biological Engineering and Computing 2001; 40(1):2-13
[8] Lamard M, Daccache W, Cazuguel G, Roux C, Cochener B. Use of JPEG-2000 Wavelet Compression Scheme for Content-Based Ophthalmologic Retinal Retrieval. Proceedings of the 27th annual international conference of IEEE engineering in medicine and biology society. 2005 September 4010-4013.
[9] Taubman D, Marcellin M. JPEG2000: Image Compression Fundamentals, Standards and Practice. Kluwer Academic Publishers ; 2001.
[10] Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. Detection of lesions in retina photographs based on the wavelet transform. Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2006 August 2618-2621.
[11] Bezdek JC. Fuzzy Mathemathics in Pattern Classification [Thèse]. Ithaca : Applied Math. Center, Cornell University ; 1973.
[12] Breiman L. Random Forests, Machine Learning 2001;45(1):5-32.
[13] Dietterich T. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning 2000;40(2):139-157.
[14] Wilson DR, Martinez TR. Improved Heterogeneous Distance Functions, Journal of Artificial Intelligence Research 1997;6:1-34.

Tableau I : Distribution des niveaux de sévérité

Niveau de sévérité	Nombre de patients
Pas de rétinopathie apparente	7
Rétinopathie non proliférante minime	11
Rétinopathie non proliférante modérée	18
Rétinopathie non proliférante sévère	9
Rétinopathie proliférante	8
Rétinopathie traitée non active	10

Distribution des niveaux de sévérité de la rétinopathie diabétique parmi les 63 patients de la base de données.

Tableau II : Information contextuelle structurée à propos des patients

attributs	valeurs
contexte clinique général	
contexte clinique familial	diabète, glaucome, cécité, autre
contexte clinique médical	HTA, dyslipidémie, proténurie, dialyse rénale, allergie, autre
contexte clinique chirurgical	cardiovasculaire, greffe pancréas, greffe rénale, autre
contexte clinique ophtalmologique	cataracte, myopie, DMLA, troubles des milieux, glaucome, chirurgie de la cataracte, chirurgie du glaucome, autre
circonstances, réalisation de l'examen, cadre du diabète	
type du diabète	DID (I), DNID (II), néant
ancienneté du diabète	< 1 an, 1 à 5 ans, 5 à 10 ans, > 10 ans
équilibre du diabète	bon, mauvais, modifications rapides, hémoglobine glycosylée
traitements	insuline injection, insuline pompe, ADO + insuline, anti diabétiques oraux, greffe du pancréas
symptômes ophtalmologiques avant réalisation angiographie	
asymptomatique sur le plan ophtalmologie	néant, dépistage ophtalmologique systématique, diabète connu, attente diabétique extra, ophtalmologie, diabète de découverte récente par bilan
symptomatique sur le plan ophtalmologie	infection, autre, néant, BAV unilatérale, BAV bilatérale, glaucome néovasculaire, hémorragie intrarétinienne, décollement de rétine
maculopathie	
maculopathie	œdémateuse focale, œdémateuse diffuse, néant, ischémique

Les informations contextuelles éventuellement disponibles à propos des patients, groupées par catégories, sont énumérées sur la colonne de gauche. La colonne de droite indique la valeur que peut prendre chacun de ces attributs.

Figure 1 : Photographies de l'œil d'un patient

Les images (a), (b) et (c) sont des photographies obtenues par application de différents filtres de couleur. Les images (d) à (j) forment une série d'images angiographiques prises à différents intervalles de temps (précoce (e), intermédiaire (d), (f), (g), (h), (j) et tardif (i)).

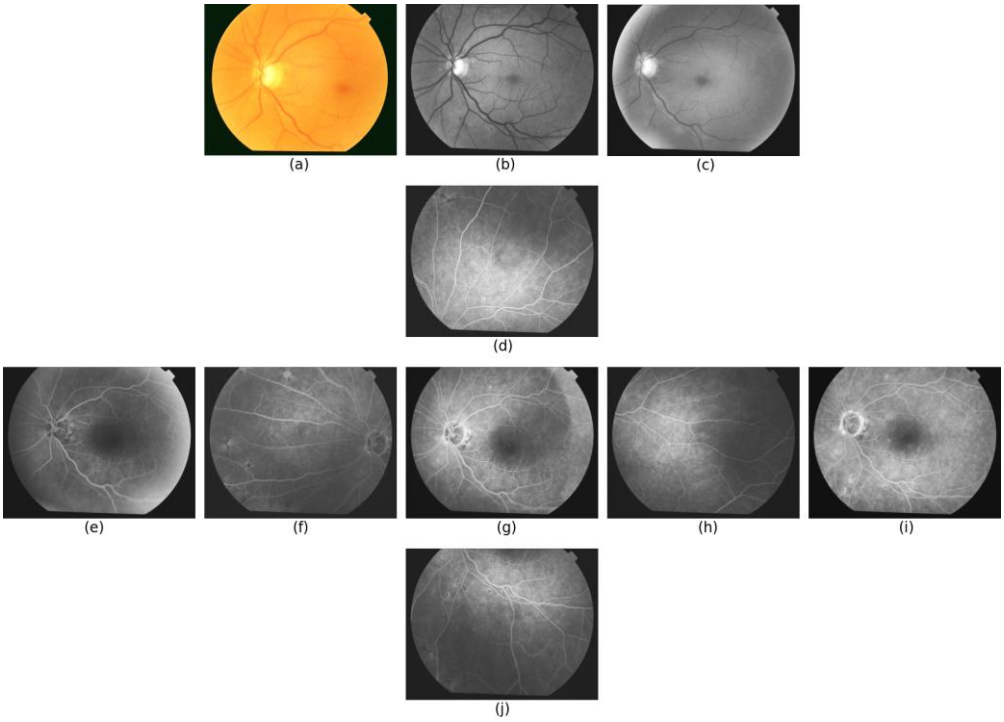

Figure 2 : Structure et apprentissage d'un arbre de décision

Le schéma (a) représente un exemple d'arbre de décision. Il est constitué de quatre règles permettant de diviser la population en quatre groupes : « si sexe=femme et image angiographique tardive ∈ G1 alors groupe 1 », « si sexe=femme et image angiographique tardive ∈ G2 alors groupe 2 », « si sexe=homme et âge<40 alors groupe 3 » et « si sexe=homme et âge≥40 alors groupe 4 ». Le schéma (b) décrit le processus d'apprentissage. 1) Dans cet exemple, le descripteur le plus discriminant sur la population entière est le sexe, donc on divise la population totale entre celle des hommes et celle des femmes. 2) Le descripteur le plus discriminant sur la population des femmes est l'image angiographique tardive. Nous divisons donc cette population en fonction du groupe d'images auquel appartient celle de chaque patient. 3) Le descripteur le plus discriminant sur la population des hommes est l'âge ; cette population est donc divisée en fonction de l'âge de chaque patient.

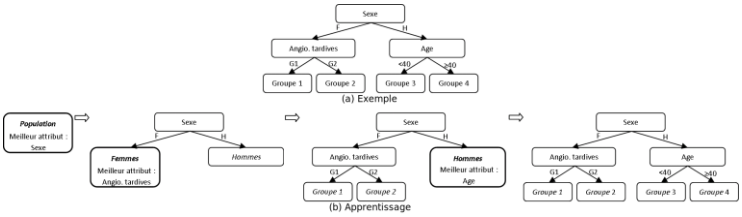

Figure 3 : Pouvoir discriminant de chaque descripteur

Cette figure montre le gain d'entropie obtenu pour chaque descripteur en divisant l'ensemble des cas de la base. Les descripteurs continus sont affichés à gauche de la figure, les descripteurs nominaux au centre et les images à droite. Les descripteurs les plus discriminants sont ceux qui maximisent le gain d'entropie.

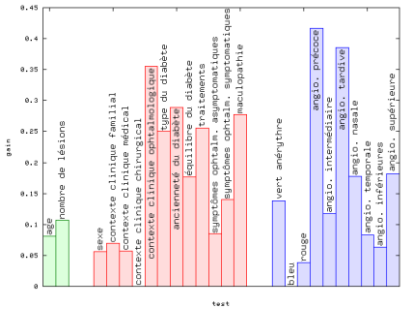

Figure 4 : Principe du moteur de recherche basé sur un arbre de décision

Le schéma (a) représente l'arbre de décision : le poids p(v) de chaque nœud v de l'arbre est indiqué dans le rectangle représentatif du nœud. Les schémas (b) et (c) représentent le degré d'affectation de deux cas c_1 et c_2 aux nœuds de l'arbre (a). D'après le poids d'affectation des deux cas aux feuilles de l'arbre, (0,1 0,1 0,8 0) et (0,2 0 0,8 0) respectivement, on en déduit leur mesure de similarité : elle vaut (0,1 0,1 0,8 0).(0,2 0 0,8 0)^t = 0,66.

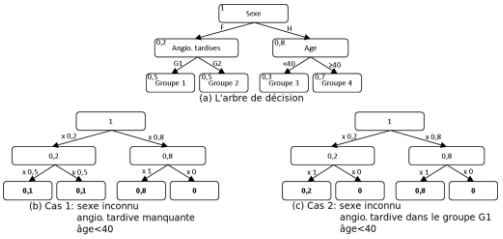

Figure 5 : Influence des descripteurs numériques et contextuels sur la précision du système

La figure représente la précision de la recherche pour une fenêtre de cinq cas, selon que le système utilise les descripteurs numériques, les descripteurs contextuels ou l'ensemble des descripteurs.