Multimedia medical case retrieval using decision trees.

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

Multimedia medical case retrieval using decision trees

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

Abstract— In this paper, we present a Case Based Reasoning (CBR) system for the retrieval of medical cases made up of a series of 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. As a comparison, the retrieval sensitivity is 52.3% for a standard multimodal case retrieval using a linear combination of heterogeneous distances.

Index Terms— CBR, image indexing, contextual information, incomplete information, decision trees, diabetic retinopathy

I. INTRODUCTION

In medicine, the knowledge of experts is a mixture of textbook knowledge and experience through real life clinical cases. Consequently, there is a growing interest in case-based reasoning (CBR) [1], introduced in the early 1980s, for the development of medical decision support systems. The underlying idea of CBR is the assumption that similar problems have similar solutions, an idea backed up by physicians' experience. In CBR, the basic processes of interpreting a new situation revolve around the retrieval of relevant cases in a case database. This process is followed by the adaptation of the past to the new problem or situation. Relevance, in such systems, is usually modeled via a similarity measure between structured cases.

Doctors make diagnoses using information collected from different sources. For instance, to diagnose Diabetic Retinopathy (DR), physicians analyze multimodal series of images together with structured information like the patient age, sex and medical history. CBR has been widely applied to structured cases, but information like images cannot be processed directly by standard systems. Automatic image indexing using digital content (Content-Based Image Retrieval [2][3]) is a possible solution to define a similarity measure between images and hence extend CBR to cases containing images. This solution is particularly interesting

Gwénolé Quellec, Guy Cazuguel and Christian Roux are with ENST Bretagne, GET-ENST, Brest, F-29200 France gwenole.quellec@enst-bretagne.fr

Mathieu Lamard and Béatrice Cochener are with Univ Bretagne Occidentale, Brest, F-29200 France

Gwénolé Quellec, Mathieu Lamard, Lynda Bekri, Guy Cazuguel, Béatrice Cochener and Christian Roux are with Inserm, U650, Brest, F-29200 France

Lynda Bekri and Béatrice Cochener are with CHU Brest, Service d'Ophtalmologie, Brest, F-29200 France

for its objectivity and reproducibility.

When designing a CBR system to retrieve such cases, several problems arise: we have to aggregate heterogeneous variables (images, nominal and continuous variables), and moreover, we sometimes have to deal with missing information. Decision trees (generally used for classification) are well suited to solve both these problems. So, we propose to adapt decision trees to the indexing context.

The setup of the article is as follows. Section II-A describes the database we used for evaluation. Section II-B shortly describes decision trees and our motivation to use them. Section II-C explains how images are included in a decision tree. Section II-D describes our proposed multimodal indexing model. The calibration procedure is described in section II-E and results are given in section III. We end with a discussion and conclusion in section IV.

II. MATERIAL AND METHODS

A. Diabetic retinopathy database
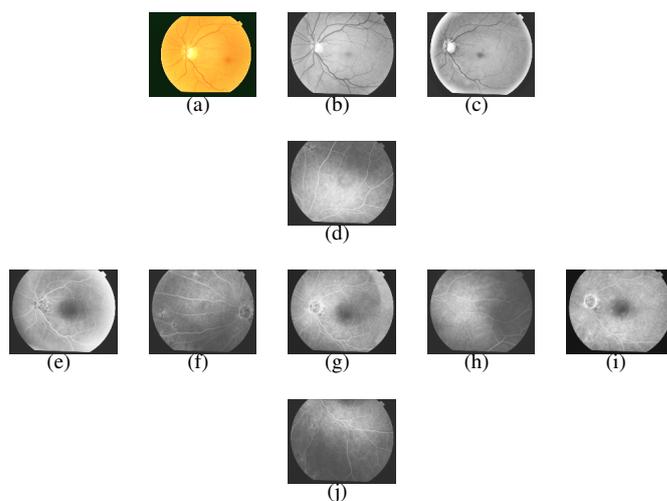
Fig. 1.   Photograph series of a patient eye

Images (a), (b) and (c) are photographs obtained by applying different color filters. Images (d) to (j) form a temporal angiographic series: a contrast product is injected and photographs are taken at different stages (early (e), intermediate (d), (f), (g), (h), (j) and late (i)). For the intermediate stage, photographs from the periphery of the retina are available.

The diabetic retinopathy (DR) database contains retinal images of diabetic patients, with associated anonymized information on the pathology. Diabetes is a metabolic disorder characterized by sustained inappropriate high blood sugar levels. This progressively affects blood vessels in many

TABLE I

Structured contextual information for diabetic retinopathy patients

attributes	possible values
general clinical context	
familial clinical context	diabetes, glaucoma, blindness, misc.
medical clinical context	arterial hypertension, dyslipidemia, protenuria, renal dialysis, allergy, misc.
surgical clinical context	cardiovascular, pancreas transplant, renal transplant, misc.
ophthalmologic clinical context	cataract, myopia, AMD, glaucoma, unclear medium, cataract surgery, glaucoma surgery, misc.
circumstances, examination and diabetes context	
diabetes type	none, type I, type II
diabetes duration	< 1 year, 1 to 5 years, 5 to 10 years, > 10 years
diabetes stability	good, bad, fast modifications, glycosylated hemoglobin
treatments	insulin injection, insulin pump, anti-diabetic drug + insulin, anti-diabetic drug, pancreas transplant
eye symptoms before the angiography test	
ophthalmologically symptomatic	none, systematic ophthalmologic screening - known diabetes, recently diagnosed diabetes by check-up, diabetic diseases other than ophthalmic ones
ophthalmologically asymptomatic	none, infection, unilateral decreased visual acuity (DVA), bilateral DVA, Neovascular glaucoma, intra-retinal hemorrhage, retinal detachment, misc.
maculopathy	
maculopathy	focal oedema, diffuse oedema, none, ischemic

organs, which may lead to serious renal, cardiovascular, cerebral and also retinal complications. Different lesions appear on the damaged vessels, which may lead to blindness. The database is made up of 63 patient files containing 1045 photographs altogether. Images have a definition of 1280 pixels/line for 1008 lines/image. They are lossless compressed images. Patients have been recruited at Brest University Hospital since June 2003 and images were acquired by experts using a Topcon Retinal Digital Camera (TRC-50IA) connected to a computer. An example of an image series is given in figure 1. The contextual information available is the patients' age and sex and structured medical information (see table I). Thus, at most, patients records are made up of 10 images per eye (see figure 1) and of 13 contextual attributes; 12.1% of these images and 40.5% of these contextual attribute values are missing. The disease severity level, according to ICDRS classification [4], was determined by experts for each patient. The distribution of the disease severity among the above-mentioned 63 patients is given in table II.

TABLE II

Patient disease severity distribution

disease severity	number of patients
no apparent DR	7
mild non-proliferative DR	11
moderate non-proliferative DR	18
severe non-proliferative DR	9
proliferative DR	8
treated/non active DR	10

B. Decision trees

Decision trees (DTs) [5][6] are used to divide a population of cases into homogeneous groups, according to a set of discriminant features; these features are automatically searched for (by a learning process) amongst all the available features

(images and contextual attributes in our case, see section II-A), as explained below. The case population is segmented in a hierarchical way, hence a tree with such structure is built:

- each non-terminal node corresponds to a test on a single feature (e.g. what is the patient sex ?)
- each edge corresponds to a test outcome (e.g. male/female)
- each leaf corresponds to a cluster of cases that provide a similar answer to each test

At the beginning of the learning process, the tree is made up of a single node containing the whole case population. Then, for each leaf L of the developing tree, the most discriminant feature is searched for and the population in L is split into new child nodes, one for each outcome of the test. The discriminant power of a test can be measured by the Shannon entropy gain G [5] (see equation (1)) obtained when dividing the current node into child nodes,

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

where p_c is the percentage of cases with label c ($c = 1..C$) in a node, I^0 is the entropy in the parent node (before dividing it) and I^n ($n = 1..N$) is the entropy in the nth child node. This measure characterizes the purity of the segmentation. The Shannon entropy gain was used in the proposed method. DTs were first designed to segment nominal attribute vectors (each test outcome corresponds to a feature value or group of values). Quinlan [7] extended them to continuous attributes (learning samples are grouped by attribute value ranges). More generally, DTs can process any feature, so long as we provide a way to cluster cases according to that feature. Since each test is performed on a single feature, DTs are well suited to process heterogeneous cases. Moreover, DTs can manage missing information: a simple mechanism is provided by the c4.5 algorithm [5]. Suppose that the value of a feature f, tested at a node v_0, is missing for some case. Then this

case is assigned to each child v_i of v_0 with a weight w_{v_i}, $0 \leq w_{v_i} \leq 1$. w_{v_i} is the percentage of samples, whose value for f is known, assigned to v_i.

C. Images in Decision Trees

The integration of images in a DT was inspired by CBIR. CBIR involves 1) building a signature for images by extracting image features, and 2) defining a distance measure associated with the signature. Thus, measuring the distance between two images comes down to measuring the distance between two signatures. Similarly, we could segment cases according to an "image attribute" by clustering the corresponding signatures, and assign each cluster to a child node. For the DR database, an "image attribute" corresponds to an imaging modality and the photographed part of the retina (the center or the upper, lower, nasal or temporal periphery, see figure 1). By this procedure, images can be easily integrated in a DT.

In previous studies, we proposed to compute a signature for images (i.e. a feature vector summarizing image content) from their wavelet transform (WT) [8]. These signatures model the distribution of the WT coefficients in each subband of the decomposition. The associated distance measures D [8] compute the divergence between these distributions. These signatures and distance measures were used to build the DTs.

Any clustering algorithm can be used, provided that the distance measure between feature vectors can be specified. We used FCM (*Fuzzy C-Means*) [9], one of the most common algorithms, and replaced the Euclidian distance by D.

In addition to the global image signatures mentioned above, the number of microaneurysms (the most frequent lesion of DR) automatically detected by the algorithm described in [10] is also added to the feature vector. It is processed in DTs as any continuous attribute (such as patients' age).

D. Multimodal Decision Tree based indexing

At the end of the learning step, each supervision example i is assigned to each leaf j ($j = 1..N$) with a weight w_{ij} (w_{ij}=0 or 1 if every tested attribute is known for i, $0 \leq w_{ij} \leq 1$ otherwise, see section II-B). Similarly, when a new case q is presented to the system, we can assign it to each leaf with a weight w_{qj}. To derive a retrieval system from a DT, we apply the following method:

1) The similarity measure S_{qi} between q and each supervision example i is initially set to 0.
2) If q and some example i fall in the same leaf j, their similarity measure are increased according to their assignment weight to j, namely by $w_{qj}.w_{ij}$. In other words, $S_{qi} = \sum_{j=1}^{N} w_{qj}.w_{ij}$
3) Examples i are ordered by decreasing order of S_{qi}.

A similar retrieval system can be derived from several trees simultaneously: the similarity measure S_{qi} is then simply computed over every leaf of the set of trees. Several methods have been proposed in the literature to generate such sets: Random Forests [11] based on CART or randomized c4.5

[12] for instance. Their performances as classifiers are usually better than that of single DTs. In our case, it emerges that a retrieval system based on several trees is also more accurate. To generate DT sets, we randomized the learning algorithm as follows: to select a test for a node, we search for the k most discriminant variables according to the entropy gain (see equation (1)) and pick one of them uniformly.

E. Calibration procedure

Although we do not use DTs as a classification method, we need to define class labels for supervision examples, to evaluate the discrimination ability of each attribute (see equation (1)). The disease severity level was used in that purpose.

To learn DTs, we must divide cases into:

- a learning set (the supervision examples), used to find the most discriminative attributes at each node,
- a validation set, used to determine when we should stop dividing nodes,
- a test set to evaluate the efficiency of the system

We define the efficiency of the system as the mean sensitivity over the test set: the sensitivity is the percentage of retrieved cases whose label is identical to the query case's. We set the number of retrieved cases by query to 5, in accordance with physicians' needs.

To improve the system, only the "best" generated trees are used for retrieval, so trees are first evaluated individually and sorted. We define their score as the mean sensitivity over the validation set.

The following parameters have to be set:

- the number $p^1 = N$ of generated trees
- the random parameter $p^2 = k$ (see section II-D): the number of most discriminant variables among which the testing variable is selected at a node
- the FCM parameter (the fuzzyness coefficient) $p^3 = m$
- the percentage $p^4 = \alpha$ of selected trees among the N generated

A discrete set of values $P^i = \{p_1^i, p_2^i, ..., p_{n_i}^i\}$ is evaluated for each parameter p^i, and the best element of the product space $P^1 \times P^2 \times P^3 \times P^4$ is selected. Elements of the product space are evaluated by a n-fold cross-validation (i.e. the experiment is carried out several times with different learning, validation and test sets, selected at random).

III. RESULTS

The best set of parameters is the following:

- $N = 200$ generated trees
- random parameter $k = 6$
- fuzzyness coefficient $m = 2$
- $\alpha = 20\%$ of trees selected

Thus the size of the DT set is 40 ($\alpha.N$). The mean sensitivity over the test set reaches **79.5%** for these parameters. As a comparison, the mean sensitivity obtained by CBIR (when single images are used as cases) with the same image signatures is 46.1%. To evaluate the contribution of DTs for the retrieval of heterogeneous and incomplete cases, the
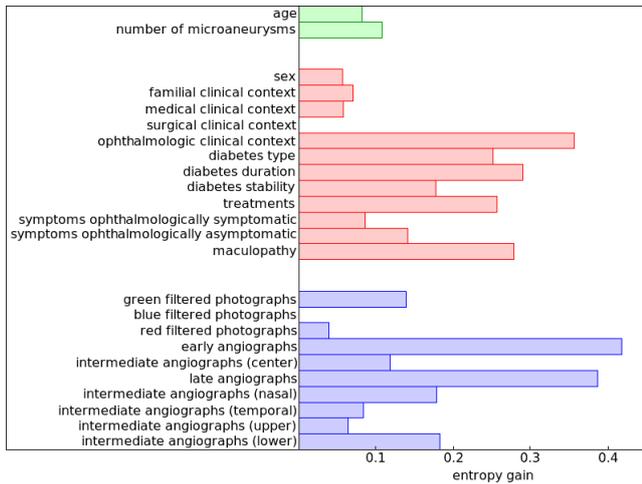
This figure shows the entropy gain using each attribute to split the root node of a tree. These gains are computed over all the database. Continuous attributes are displayed at the top, nominal attributes at the center and vectors at the bottom.

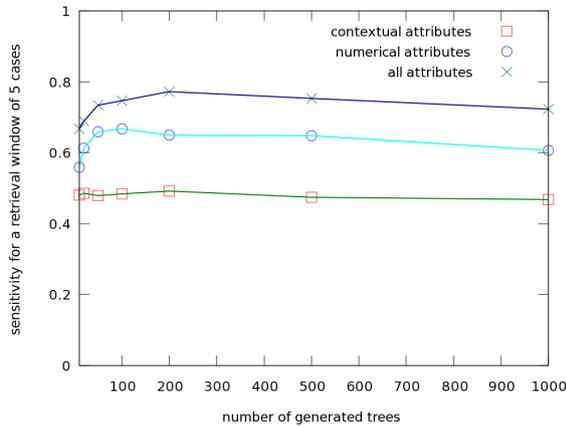Fig. 2.   Discrimination ability of each attribute

Fig. 3.   Influence of numerical and contextual attributes on the diabetic retinopathy database

proposed method was compared to a linear combination of heterogeneous distance functions, managing missing values [13]. This method was used as a reference since it is the natural generalization of CBR. Its extension to vectors containing images is based on the distance between image signatures (see section II-C). A mean sensitivity of 52.3% was achieved by this method on the DR database.

To bring out the discrimination ability of single attributes, we give in figure 2 the entropy gain (see equation (1)) when the root node of any tree is split, according to each attribute. More generally, to estimate the contribution of numerical (image series signatures) and contextual information, DT sets are learnt using numerical or contextual information alone. Sensitivity results are reported in figure 3.

IV. DISCUSSION AND CONCLUSION

In this article, we introduce a method to include image series and their numerical signatures, with contextual infor-

mation, in CBR systems. It exploits the abilities of decision tree based hierarchical clustering to combine heterogeneous information. In particular, a way to include image signatures in a DT was proposed. This retrieval method takes advantage of the ability of DTs to handle missing values and to avoid over learning. The latter property makes this method well suited to process databases with few cases such as the DR database. On this database, the method outperforms our first CBIR algorithm by a factor of 172.5% ($\frac{79.5\%}{46.1\%}$) in sensitivity. This stands to reason since an image alone is generally not sufficient for experts to correctly diagnose the disease severity level of a patient. Indeed, as figure 3 shows, using images series without contextual information, instead of single images, increases by itself the sensitivity by a factor of 144.7% ($\frac{66.7\%}{46.1\%}$). Besides, this non-linear retrieval method is 152.0% ($\frac{79.5\%}{52.3\%}$) more sensitive than a simple linear combination of heterogeneous distances on the DR database. The proposed framework is also interesting for being generic: any multimedia database may be processed so long as a procedure to cluster cases is provided for each new modality (sound, video, etc).

This article gives promising results about the use of data mining techniques to combine numerical and contextual information in a retrieval framework, so we are now focusing on more elaborate data mining algorithms.

REFERENCES

[1] A. Aamodt, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, March 1994.

[2] C. Nastar, "Indexation d'images par le contenu: un etat de l'art," in *CORESA'97*, March 1997.

[3] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, December 2000.

[4] C. Wilkinson, F. Ferris, R. Klein, and al., "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.

[5] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[6] L. Breiman, J. Friedman, R. Olshen, , and C. Stone, "Classication and regression trees," Wadsworth, Belmont, Ca., 1984.

[7] J. Quinlan, "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*, 1992, pp. 343–348. [Online]. Available: citeseer.ist.psu.edu/quinlan92learning.html

[8] M. Lamard, W. Daccache, G. Cazuguel, C. Roux, and B. Cochener, "Use of jpeg-2000 wavelet compression scheme for content-based ophtalmologic retinal retrieval," in *Proceedings of the 27th annual international conference of IEEE engineering in medecine and biology society*, september 2005.

[9] J. Bezdek, "Fuzzy mathematics in pattern classification," Ph.D. dissertation, Applied Math. Center, Cornell University, Ithaca, 1973.

[10] G. Quellec, M. Lamard, P. Josselin, G. Cazuguel, B. Cochener, and C. Roux, "Detection of lesions in retina photographs based on the wavelet transform," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, august 2006, pp. 2618–2621.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: citeseer.ist.psu.edu/breiman01random.html

[12] T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[13] D. Wilson and T. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, pp. 1–34, 1997. [Online]. Available: citeseer.ist.psu.edu/wilson97improved.html