MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Case Based Diagnosis in Histopathology of Breast Tumours

Marie-Christine Jaulent, Christel Le Bozec, Eric Zapletal and Patrice Degoulet

Service d'Informatique Médicale - Hôpital Broussais - 96 rue Didot - 75014 Paris - France

Abstract

Relevant knowledge and decision making process in histopathology is mostly included in typical pathological cases encountered by the expert. In this article we address the issue of exploiting this knowledge in the diagnosis process. We present the first steps of a Case-Based-Reasoning (CBR) system that uses the previously resolved pathological cases in order to facilitate decision making and diagnosis formulation of a new case. The work has been performed in two phases. Firstly, an object-oriented model of the domain was developed and 35 pathological cases of breast tumours were represented within this model. Secondly, the functional architecture of the CBR system was designed and the main procedure, the selection of similar cases, was achieved. The selection procedure is based on an original similarity measure that takes into account both semantic and structural resemblances and differences between the cases. A first evaluation of the system was performed on several cases of the data base. The interest of the CBR approach in situations where heuristic rules cannot be clearly defined is discussed.

Keywords:

Case Based Reasoning, Breast Tumour Diagnosis, Decision-support

Introduction

The identification of tumours in histopathology requires the integration of various and complex data. Such data essentially stem from macroscopic pieces and histologic images analysis. The ongoing research on physiopathological process continuously provides descriptions of new entities and remodelling of relations between the different entities. One consequence is a lack of consensus on the semiological description of images in terms of microscopic and cytological features [1]. Moreover, it exists many exceptions and variations in the possible different diagnosis and in the semiological descriptions of images. Finally, there are benign and malignant counterparts which can exhibit similar appearances. In front of the abundance of histologic patterns, the non expert pathologist frequently faces complex decision making problems that require an expert knowledge, based on personal, long-term professional experience. Indeed the expert has the ability to answer questions such as: "Have you already seen a case like mine? What are the cases close to mine? What was the diagnosis process that was performed the previous time? How to adapt the previous diagnosis to the present situation?" [2]. To answer such questions, the expert is likely to use heuristics rather than logically formulated rules and he is particularly good at recognising a new problem as analogous to a certain kind of problem that he/she already knows how to solve.

In this paper, we address the issue of considering that part of knowledge is embedded in all the concrete cases a physician previously solved and that this previous experience is rich enough to help decision making in histopathology.

We present a computer system that uses a Case-Based Reasoning approach to modelise the expert knowledge and the expert diagnosis process in histopathology. The main goal of the system is to assist pathologists in the histologic diagnosis of breast disease. In this study, the domain is limited to the histopathologic cases of breast tumours. The knowledge representation problem and the retrieval of similar cases are emphasised. The system has been tested on a database containing 35 breast tumour pathological cases. The first results are presented and discussed.

Computer systems for histopathologic diagnosis

Diagnosis and prognosis decision support systems in histopathology rely on a wide variety of methods. Some of them are based on a bayesian network [3,4,5]. For instance, the Pathfinder system [3] uses subjective and bayesian probability theory to assist pathologists in the diagnosis of lymph-node pathology. An hypothetico-deductive approach is used to provide a distribution of probability over diseases. One aspect of this system is the construction of the belief network from a set of cases which is *a priori* given. Thus, the introduction of new interesting cases implies a time consuming reconstruction of the network. Generally speaking, these systems require a close collaboration between physicians and knowledge engineers in order to assess probabilities.

Other expert systems, like CancerStage [6], rather use Artificial Intelligence methods and are the usual alternative to the probabilistic approach. The expert knowledge is embedded in IF-THEN-ELSE rules that combine various qualifiers, variables and choices according to different situations. Although the rep-

resentation of knowledge is more natural than in the probabilistic approach, it can be hard to be exhaustive and to handle partially unclear situations as well as to end up with independent rules. As a matter of fact, histopathologic imagery can hardly be modelised in the format of logical knowledge representation. The lack of statistical data and the enormous effort which would be necessary to assert such data impede relationships between morphological features to be expressed in terms of logical conditions. So that, it is natural for the expert to express his/her knowledge using concrete examples rather than decision trees, production rules or probabilistic knowledge. Examples are used to support the decision process as well as to provide explanations and to guide the reasoning process in a new case. The main contention of our work is that the case and the collection of cases include all the necessary and sufficient information to solve new cases and that the solution is based on an analogical reasoning. Moreover, if one assumes that there are variables among the collected data set of a case that affect significantly the diagnosis, one can assert that diagnosis assessment is directly obtained by analogy from the study of a previous resolved similar case. To verify this hypothesis, we have considered the analogical reasoning approach and in particular the CBR technique which is a specialisation of this reasoning approach.

Methods

The case based reasoning approach derives from analogical research [7]. It is used by the Artificial Intelligence community to design problem-solving computer systems as expert systems or more generally knowledge based systems. The originality of the approach stems from the nature of the knowledge. Indeed, it uses the specific knowledge of previously experienced concrete situations (cases). The knowledge base is defined by the collection of these cases. A new problem is solved by finding a similar case and reusing it in the new problem situation. From a theoretical point of view, CBR can be considered as a form of intra-domain analogy [8].

The usual life cycle of a CBR system (figure 1) is composed of four processes [8].

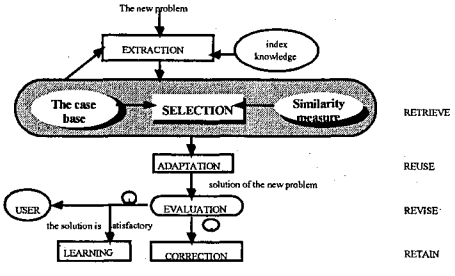

Figure 1 - - The life cycle of a CBR system

The first process includes two phases. Firstly, an *extraction* of interesting cases out of the knowledge base, based on an indexation mechanism for example, allows to restrict the searching area. Secondly, a *selection* allows to retrieve the most similar cases. The retrieval capacity stems from the definition of a similarity measure between a case and a new problem that encompasses the notion of distance. As much as possible, it is convenient to use a similarity measure derived from existing models like the "contrast model" of Tversky [9]. However, in many situations, cases are complex entities and similarities of different natures have to be considered like for instance syntactical and semantic similarities. Gentner in [10] discusses a general classification of similarity measures depending on whether the similarity is based on structural information (architectural aspects of the case) or surface information (semantic description).

The second process has in charge the reuse of the information and knowledge present in the most similar case to solve the new problem. This phase is often called the *adaptation* step. It realizes a knowledge transfer between the old and the new case that can be complex or limited to an identity transfer [11].

The third process consists in comparing the solution provided by the system to the reality. In some case, the evaluation can be performed by the system itself, if it exists a theory of the domain. Otherwise, the evaluation is finally done directly by the expert. Finally, the parts of this experience likely to be useful for future problem solving must be retained. In particular, if there is a failure (for instance no similar cases was found), it is important to explicate and manage this failure in a learning phase.

A CBR system in breast tumour histopathology

In the context of our application, we have focused on the first process, that is the retrieval in the base of the most similar cases. The two main aspects that are presented concern:

- the representation of a case in the domain
- the definition of the resemblance between two cases.

The knowledge representation

The usual description of a case is written in natural language by the physician in a more or less standardised report [1]. From the analysis of several reports, we have developed an object oriented model for our restricted histopathological domain (figure 2). Within this model, a case is described as a collection of macroscopic areas, each of them associated to a collection of histologic areas. An histologic area can contain several histologic areas as well as a cytological description.

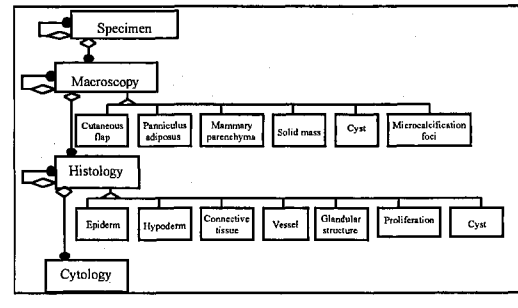

Figure 2 - The hierarchical model of the restricted histopathological domain

Practically, according to the model, the standard reports (figure 3), representative of interesting cases of the domain, have been translated in tree structures. In a structure, nodes correspond to macroscopic or histologic areas. Each type of macroscopic area, each type of histologic area and each type of cell is defined by a set of specific features. A feature is represented by a couple (Attribute Value) where the Attribute refers to a specific property like for instance, the colour of a macroscopic specimen, the type of an histologic area and so on. The value is a linguistic label which is specific of the case. For instance, (CELL-SIZE BIG) or (ARCHITECTURE PAPILLARY) are two features, among others, of an histologic area named "proliferation". Examples of values are BIG for the attribute CELL-SIZE or PAPILLARY for the attribute ARCHITECTURE. The domain of the possible values has been defined for each attribute. For instance, the possible values for the attribute "CONSISTENCY" are "SOFT", "SUPPLE", "HARD" and "VERY-HARD".

Inside the tree structure of a case, the internal representation of a node has two facets:

- some features are descriptive. They are called semiological features and are involved in the resemblance definition.
- some features are conclusive. They are called goal features and represent the solution. It can be for instance partial diagnosis.

There is a dependence relation between descriptive and conclusive features that expresses to what extent the goal features are consequences of the relevant semiological features. Since there is a lack of formal definition, it is most of the time assumed and expressed by an hypothetical relation [11]. In our context, a possible dependence relation could be: "the diagnosis depends on the architecture of the lesion, the size and shape of the constitutive cells". In order to express this dependence relation, a degree of importance is assigned to each semiological feature. In the example of figure 3, we see that the goal feature (DIAGNOSIS ADENOSIS) depends more on the (ARCHITECTURE PAPILLARY) feature of the lesion (1) rather than the presence of the (STROMA FIBROUS) in the histologic area (0.25).The set of degrees is heuristically given by the expert.

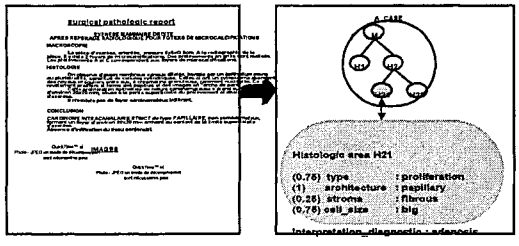

Figure 3 - The tree structure of a case

Similar cases retrieval

The restriction of the domain to breast tumour determines the nature of the cases present in the database and the context in which a new problem occurs. Then, the first phase of the first process (the extraction) is not necessary and the selection of the most similar cases is done over the whole base. The second phase is the retrieval step and it is based on a resemblance evaluation between a case in the database and a new problem. It is important to note that the representation of the new problem is the same as the representation of a case, except that the descriptive features can be incomplete and that the conclusive features are not available.

The resemblance is expressed through similarity measures. The entities to compare (a case and a new problem) are composite and their descriptions include structural and semantic characteristics. The semantic characteristics are given by the set of couples (Attribute Value) concerning the descriptive features. The structural characteristics correspond to the tree structure of the global entity. The global measure takes these two aspects into account. The similarity algorithm is a matching procedure between the tree structure of the case (C) and the tree structure of the new problem (N). The global similarity $S(N,C)$ is made of the composition of a "surface" similarity and a "structure" similarity [12].

The structure similarity

The *Structure similarity* is based on a tree matching algorithm. The procedure consists in finding the best matching between the tree structure of the new problem and the tree structure of the known case [12]. The best matching correspondsto the best global surface similarity relying on the similarities established at each level of the tree.

The surface similarity

The surface similarity, $S_{surface}$, expresses the semantic resemblance between two macroscopic areas or two histologic areas of the same type.

An attribute similarity is computed for each descriptive feature and expresses a proximity concept between the different values of the feature. A similarity table is defined for each feature for the comparison of the values two by two. These tables are different according to the representational space of the values.

- if the representation space is symbolic, the attribute similarity corresponds to the strict equality of the symbol values. In that case, the similarity table has 1 on the diagonal and 0 elsewhere.
- for other attributes, we have integrated a symbolic/ numeric approach that allows a more flexible similarity (table 1). In that case, it is possible to place the different values (the linguistic labels) on a scale.

The value "s(i,j)" of similarity can be directly given by the expert, otherwise, a default value is given by equation 1. let p be the number of possible values for a given feature,

$$(\forall i,j \in [1,p])\left(sim = s(i,j) = max\left(0, k \cdot \frac{|i-j|}{p}\right)\right) \qquad (1)$$

where $k = \frac{3}{2}$.

For instance, in the table 1, we have placed the 5 labels of the attribute consistency: soft, supple, hard, very hard and heterogeneous.

Table 4: Example of a similarity table

consistency	soft	supple	hard	very hard	heterogeneous
soft	1	s(1,2)	s(1,3)	s(1,4)	s(1,5) = 0.5
supple	s(2,1)	1	s(2,3)	s(2,4)	s(2,5) = 0.5
hard	s(3,1)	s(3,2)	1	s(3,4)	s(3,5) = 0.5
very hard	s(4,1)	s(4,2)	s(4,3)	1	s(4,5) = 0.5
heterogeneous	s(5,1) = 0.5	s(5,2) = 0.5	s(5,3) = 0.5	s(5,4) = 0.5	1

In that example, the concept "heterogeneous" is equidistant from the other concepts, that are equally distributed.

All the attribute similarities are aggregated to obtain the surface similarity between the two areas n and c. The aggregation is based on the weighted minimum operator [12]. Let n and c be described as an ordered set of features: $n = (n_1, n_2, \ldots, n_i, \ldots, n_T)$ and $c = (c_1, c_2, \ldots, c_i, \ldots, c_T)$. Let a_i be the importance given to the i^{th} descriptive feature, the *surface similarity* is then given by :

$$S_{surface}(n, c) = min_{i \in \{1 \ldots T\}} \{max(sim(n_i, c_i), 1 - a_i)\} \quad (2)$$

where

a_i = degree of importance for the feature i ,

$n = \{n_i\}, i \in \{1 \ldots T\}$ $c = \{c_i\}, i \in \{1 \ldots T\}$

Results

Thirty-five pathological specimen were selected to constitute the case base. These cases correspond to confirmed reference cases in an histopathology department. They include five disjoint diagnostic categories A, B, C, D and E. A first prototype has been implemented in C++ under UNIX and Apple Macintosh environments. Once the most similar case has been selected, the system presents to the user the semiology, the associated images and the diagnosis for this closest case. It gives also the value of the computed similarity as well as the computed partial similarity values at each level of areas hierarchy.

In the current state of the system, the prototype has been tested on the cases of the base itself, that is, each case is considered in turn to be the new problem to solve. The idea was at first to verify that a case is most similar to the cases having the same diagnosis than to those having other diagnosis. The results are registered in the table 2. A number in table 2 is a global degree $S_{Inter-Class}$, ranging from 0 to1 and defined as follows.

Let n be the number of cases c_i in class A, for instance. The degree $S_{Inter-Class}$ is expressed by the equation 3 and corresponds to the mean of the global similarity obtained for the cases inside the same diagnostic class taken two by two. This first evaluation step yields the fact that there is a better similarity between cases in a same diagnostic class than between cases in different classes. This is true except for the class C. For this class, it appears that the diagnosis may correspond to very different semiologies so that it is necessary to distinguish sub-diagnostic classes to get better results.

$$S_{Inter-Class} = \frac{\sum_{i,j \in (1,n), i \neq j} S(N,C)}{n} \quad (3)$$

Table 2 : The mean similarities for the diagnostic categories

	A	B	C	D	E
A	0.91	0.63	0.81	0.69	0.65
B	0.63	0.88	0.76	0.63	0.88
C	0.81	0.76	0.83	0.85	0.81
D	0.69	0.63	0.85	0.95	0.73
E	0.65	0.88	0.81	0.73	0.91

The figure 5 provides a synoptic view of the results showing that an element of a class resembles globally more to elements inside the class than to elements to other classes.

Discussion and conclusion

In this paper we presented a case-based reasoning approach to exploit the experience on resolved cases for diagnostic decision making in histopathology. The main advantage of this method is the fact that the knowledge is expressed in a natural way [8]. The expert does not need to express his/her knowledge through production rules or decision trees. In a first step, an object-oriented model of the domain has been realised and the relevant cases of the domain have been described within this model. Globally, a case is a collection of macroscopic areas (described by several features), associated to a collection of histologic areas (described by other features). histologic areas contain other histologic areas and can also contain cytological descriptions. Thirty-five cases are actually in the case base.

In a second step, we have implemented the retrieval process of the CBR system. The retrieval is based on a similarity algorithm (matching) between two tree structures. The global similarity measure integrates a surface similarity between areas and a structural similarity between the hierarchy of areas. The originality of the measure can be expressed in two points: - the use of similarity tables to compare qualitative features, - the weighted aggregation for the surface similarity.

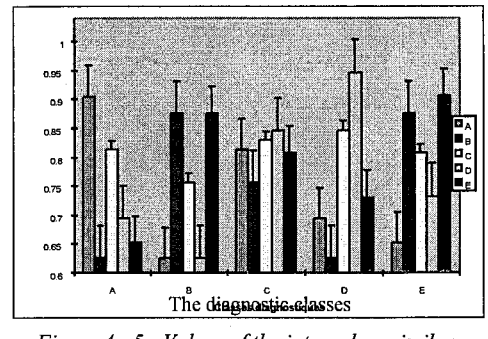

Figure 4 - 5 - Values of the inter_class similar-

The first prototype of the system demonstrates clearly the interest of this method. It is satisfactory in the sense that we obtain a meaningful resemblance rate between cases. Presently, the system is not fully evaluated. However, we had in mind its validity all along the realisation of the prototype. Indeed, in the restricted chosen domain, we looked after the exhaustiveness of the case base and we took into account the variability of the case description in the definition of the similarity measure. Moreover, the fact to provide partial degrees of similarities allows the expert to analyse the causes of a potential failure. It is then possible to modify the different parameters of the measure in order to improve the performances.

In its preliminary version, the system yields a pre-adaptation phase with regards to classical CBR systems. However, from this preliminary study, it appears that this approach is adapted and robust to missing information and provides important clues to the pathologist.

The perspectives of this work are both methodological and practical. At the methodological level, an important point concerns the representation of the imprecision inherent of the information contained in the case through the use of the fuzzy set theory. Another point concerns the representation of the case, in particular the necessity to improve the description of images by introducing morphometric attributes. At the practical level, the perspectives are broad. For instance, the domain can be extended to the whole breast pathology. Furthermore, the possibility to access, through a network, large multi-experts bases of anonymous cases could be useful in the daily practice for both medical and educational purposes.

Acknowledgments

The authors thanks the Dr. G. Contesso, MD and the Dr. J-M. Guinebretiere, MD of the Gustave Roussy Institute(IGR) rue Camille Desmoulins - 94800 Villejuif - France. They have been very helpful for the constitution of the image database.

References

[1] Rosai J. Standardized Reporting of Surgical Pathology Diagnoses for the Major Tumor Types. A proposal. *Am J Clin Pathol* 1993;100: 240-255.

[2] Chute C. Clinical data retrieval and analysis. I've seen a case like that before. *Ann N Y Acad Sci.* 1992;17: 139-140.

[3] Heckerman DE, Natathwani BN. An evaluation of the diagnostic accuracy of Pathfinder. 1992. *Comput Biomed Res.*1992;25(1): 56-74.

[4] Bartels PH, Thompson D, Bibbo M, Weber JE. Bayesian belief networks in quantitative histopathology. *Analyt Quant Cytol Histol* 1992;14:459-473.

[5] Montironi R, Bartels PH, Thompson D, Scapelli M, Hamilton P. Prostatic Intraepithelial Neoplasia, development of a Baysian beliéf network for diagnosis and grading. *Analyt Quant Cytol Histol* 1994;16:101-112.

[6] Marchevsky AN, Coons G.Expert systems as Aid for the Pathologist's role of clinical consultant: Cancer-Stage. *Modern Pathology* 1993;6(3):265-269.

[7] Riesbeck C, Schank R. Inside Case-Based Reasoning. 1989. Lawrence Erlbaum Associates. Hillsdale, New Jersey.

[8] Aamodt A. Case-Based Reasoning: Foundational Issues, Mathodological Variations, and System Approaches. *AICOM* 1994; 7:39-59.

[9] Tversky A. Features of Similarity. Psychological review 1977;84: 327-352.

[10] Gentner D. Analogical inference and analogical access. Analogica : Proceedings of the first workshop on analogical reasoning 1987;

[11] Py M. Un modèle conceptuel de raisonnement par analogie. *Revue d'Intelligence Artificielle* 1994;8: 63-99

[12] Jaulent MC, Le Bozec C, Zapletal E and Degoulet P. A case-based reasoning method for computer-assisted diagnosis in histopathology. *Lecture Notes in Artificial Intelligence*, 1997; 1211. pp 239-242.

Address for correspondence :

Marie-Christine Jaulent (jaulent@hbroussais.fr)
tel : +33 1 43 95 82 90
fax : +33 1 43 95 92 09