Incorporating Probabilistic Knowledge in HealthAgents: a Conceptual Graph Approach

Madalina Croitoru, Srinandan Dasmahapatra, Paul Lewis

Electronics and Computer Science,
University of Southampton, SO171BJ, UK

Abstract HealthAgents is a multi-agent, distributed decision support system for brain tumor diagnosis. Knowledge needs to be shared amongst different agents in order to assist clinicians when making diagnosis / prognosis. Existing terminological standards led to the development of a vocabulary to facilitate interoperability. Querying expressivity requirements as well as the need for visual capabilities further led to the development of a Conceptual Graph based description of the data sources: knowledge oriented specification. However, an important part of the medical knowledge is not encoded in this formalism: background knowledge regarding statistical correlations. As a decision support system, HealthAgents should provide the clinician all possible related information about a case. This paper presents a way of encoding and utilising such statistical information. The Simple Conceptual Graphs that describe a given hospital cases will be used to retrieve related information. Logical subsumption will be used for retrieval, while the statistical correlations will be presented to the clinician as part of the decision support system.

1 Introduction

In this paper we address the problem of integrating a set of statistical rules with a first order logic based formalism: Conceptual Graphs. This integration is thought from the perspective of a medical decision support system (DSS). In this context the clinical user of the DSS will be presented with potentially useful information related to a patient case. This new information will help in the selection of appropriate machine learning mechanisms to be used for case classification.

The work described in this paper will present a first step towards the integration of statistical data with Conceptual Graphs. Our choice of Conceptual Graphs is twofold. First, it provides easy integration with the KOS framework described in [4]. Second, the clinician feedback will be done in natural language and Conceptual Graphs will facilitate this translation. While the motivation for the work is obvious: the need of integrating the existing statistical rules with the conceptual graphs formalism; the justification for our approach needs a couple of remarks. First, the decision support system has to provide the clinician with a number of machine learning algorithms for case classification. These algorithms have been trained on a set of data with certain features (age, sex etc.). It is important to select the appropriate classifiers. At the same time the choice of classifiers is not only based on the patient case as such, but also on a set of statistical correlations that the clinician has observed. This rationale calls for the integration of reasoning capabilities for case retrieval (logical subsumption) with existing statistical correlations provided by textbooks or concrete hospital cases. Second, the nature of the system under discussion has to be considered: a decision *support* system. Indeed, our aim is to make best use of the knowledge available by presenting related information to the doctor. We do not want to develop a statistical based reasoning system, but simply to provide the clinician with all potential useful information about a case. Due to this reason, our work is evaluated empirically, looking at the usefulness of the information we provided for clinicians.

In conclusion, the advantages of the proposed approach are two fold: modularization for representation and easy evolution. Indeed, the logic and the statistical aspects are kept separate but exploited in a joined manner. Due to the nature of our representation we can easily integrate new domain knowledge / terminologies / ontologies, as a mapping between the tree representations of the terminologies and the support. In particular, the last point makes our approach very useful for the medical domain in particular, where a number of different names associated to the same object are generally accepted.

2 Motivation and related work

HealthAgents [1] is an agent-based, distributed decision-support system (DSS) that employs clinical information, Magnetic Resonance Imaging (MRI) data, Magnetic Resonance Spectroscopy (MRS) data and genomic DNA profile information. It is important to highlight at this stage that due to the medical nature of our system we are not interested in combining the logical and statistical inference aspects. While this is an interesting directions of work ([6], [5]) we believe that these approaches are unsuitable for our project for the following reasons: (1) The clinical users are reluctant of using a system that performs statistical reasoning for them. The motive is that potentially undiscovered classes of tumors could be discarded as part of the reasoning process; (2) Second, the nature of the domain makes the identification of independent variables difficult; (3) Third, exhaustive scenarios cannot be provided for representational completeness.

We propose a Conceptual Graph based methodology for retrieving relevant information that might help the clinician in the process of classifier selection. The textbook rules and correlations from the literature have been translated into a set of rules with a degree of belief attached. These rules follow the spirit of [2], only with the statistical aspect included. When a new patient case needs to be sent to the appropriate classifiers, the clinical data of the patient is translated into a Conceptual Graph. Subgraphs of this Conceptual Graph will then be projected in order to retrieve relevant information. We detail our methodology further in the next section.

3 The HealthAgents System

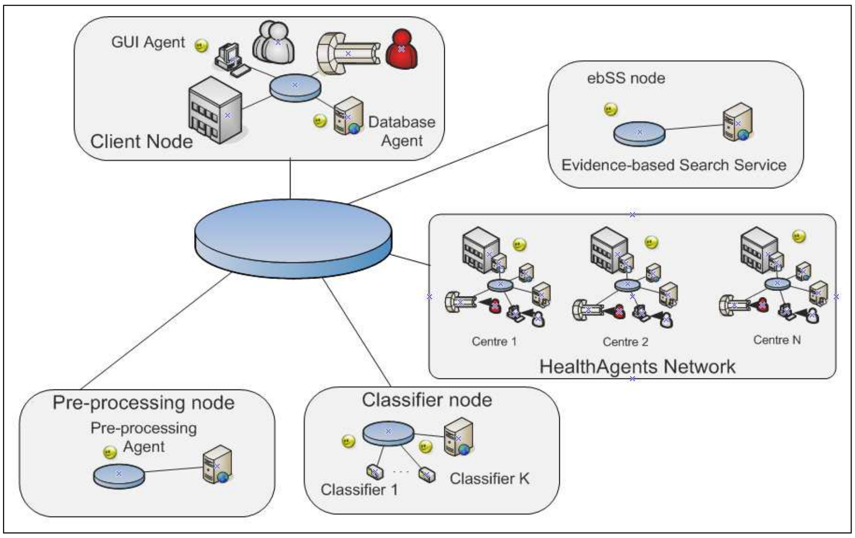

Figure 1. HealthAgents Architecture

The envisaged functionality of HealthAgents (see Figure 1) is to provide better classification accuracy for brain tumors using non invasive procedures: MRI scans, MRS scans, HRMAS and microarray information. The distributed nature of the system (with data located in different geographic areas: Birmingham, Barcelona, Valencia) will ensure a large number of cases available. These cases will be used for training classifiers on particular sets of data (e.g. male vs female, certain age groups, certain types of tumors, brain locations etc.). The classifiers will be invoked when a new patient case is presented to the system. Depending on the clinical data of the patient and the location of the tumor (as available from the MRI scan) the clinician makes the choice for what classifiers to invoke. The classifiers will provide a differentiated diagnosis (discriminating between two or more possible tumor types). Depending on the classifier results and the MRS scan, the clinician makes his decision or invokes another classifier.

Knowledge contained in the data sources is described by the means of Conceptual Graphs. This allows us to build upon the existing HADOM ontology while not overcomplicating the ontology with rules to describe data extraction techniques that employ different parameters which greatly influence the outcome data. An immediate advantage of our Conceptual Graphs choice is their graph based reasoning mechanisms which allow versatile querying algorithms [3]. The Conceptual Graph querying will allow for the clinician to search for a similar case within the cases in the HealthAgents network.

In this paper we would like to provide a functionality that allows to present extra information to the clinician that will allow to make a more informed choice of the classifiers to be invoked. Indeed, all the clinical knowledge relating brain tumor types with age, sex or brain tumor location is not exploited at all in the current version of our prototype. We propose translating such correlation rules (available from textbooks and scientific articles) into Conceptual Graph rules with an associated degree of belief. We will then use projection to select the relevant rules for a given patient case and show them to the doctor in descending order of their belief degree.

4 Using Conceptual Graphs and probabilistic information

In this section we will detail our methodology and provide a concrete example of its functionality.

First, we will describe how textbook rules and statistical correlation have been translated to a Conceptual Graph representation (Section 4.1). This statistical information was made available from books and relevant scientific articles.

Section 4.2 explains how these rules and correlations can be applied on an instance of a patient case (also represented as a Conceptual Graph). As the outcome, the doctor will be presented with a labelled tree where labels reflect the degree of probability of each rule. It is important to highlight that these labels will solely be used for the doctor as a guidance for classifier selection and not for probabilistic inference.

Each section we will first present an intuitive overview of the proposed methodology, followed by the formal description of our work. At the end of each section a concrete example is provided. However, a few definitions are needed to ensure consistency of the formalism presented throughout the paper. These definitions are provided below.

Let $G = (V_C, V_R; E_G)$ be a bipartite graph. If, for each $v_R \in V_R$, there is a linear order $e_1 = \{v_R, v_1\}, \ldots, e_k = \{v_R, v_k\}$ on the set of edges incident to v_R ($k = d_G(v)$ is the degree of v_R), then G is called an *ordered bipartite graph*. Given a node $v \in V_C \cup V_R$, $N_G(v)$ denotes the *neighbours set* of this node, i.e. $N_G(v) = \{w \in V_C \cup V_R | \{v, w\} \in E_G\}$. Similarly, if $A \subseteq V_R \cup V_C$, its *neighbours set* is denoted as $N_G(A) = \cup_{v \in A} N_G(v) - A$. We also denote the *i-th neighbour* of $v_R \in V_R$ by $N_G^i(v_R)$, meaning that $e_i = (v_R, N_G^i(v_R)) \in E_G$. If $G = (V_C^G, V_R^G; E)$ is an ordered bipartite graph and $A \subseteq V_R^G$, then the *subgraph spanned by A in G* is the graph $[A]_G = (N_G(A), A, E')$, where $N_G(A)$ is the neighbor set of A in G.

A conceptual graph support consists of a concept type hierarchy, a relation type hierarchy, a set of individual markers that refer to specific concepts and a generic marker, denoted by *, which refers to an unspecified concept. More precisely, a *support* is a 4-tuple $S = (T_C, T_R, \mathcal{I}, *)$ where:
- T_C is a finite, partially ordered set (poset) of *concept types* $(T_C, \leq)$ that defines a type hierarchy where $\forall x, y \in T_C$, $x \leq y$ means that x is a subtype of y; the top element of this hierarchy is the universal type $\top_C$;

- T_R is a finite set of *relation types* partitioned into k posets $(T_R^i, \leq)_{i=1,k}$ of relation types of arity i $(1 \leq i \leq k)$, where k is the maximum arity of a relation type in T_R; each relation type of arity i, namely $r \in T_R^i$, has an associated *signature* $\sigma(r) \in \underbrace{T_C \times \ldots \times T_C}_{i \text{ times}}$, which specifies the maximum concept type of each of its arguments; this means that if we use $r(x_1, \ldots, x_i)$, then x_j is a concept of $\mathit{type}(x_j) \leq \sigma(r)_j$ $(1 \leq j \leq i)$; the partial orders on relation types of the same arity must be *signature-compatible*, i.e. $\forall r_1, r_2 \in T_R^i\ r_1 \leq r_2 \Rightarrow \sigma(r_1) \leq \sigma(r_2)$;
- $\mathcal{I}$ is a countable set of *individual markers* that refer to specific concepts;
- $*$ is the *generic marker* that refers to an unspecified concept (however, this concept has a specified type);
- The sets T_C, T_R, $\mathcal{I}$ and $\{*\}$ are mutually disjoint;
- $\mathcal{I} \cup \{*\}$ is partially ordered by $x \leq y$ if and only if $x = y$ or $y = *$.

A (Simple) Conceptual Graph (SCG) is a 3-tuple $SG = [S, G, \lambda]$, where:
- $S = (T_C, T_R, \mathcal{I}, *)$ is a support;
- $G = (V_C, V_R; E_G, l)$ is an ordered bipartite graph;
- λ is a labelling of the nodes of G with elements from the support S: $\forall r \in V_R,\ \lambda(r) \in T_R^{d_G(r)};\quad \forall c \in V_C,\ \lambda(c) \in T_C \times \big(\mathcal{I} \cup \{*\}\big)$ such that if $c = N_G^i(r)$, $\lambda(r) = t_r$ and $\lambda(c) = (t_c, ref_c)$ then $t_c \leq \sigma_i(r)$.

When the support is fixed, we use the notation $SG = (G, \lambda)$, or we refer to the CG G and its labelling function λ_G.
If (G, λ_G) and (H, λ_H) are two CGs (defined on the same support S) then $G \geq H$ (G subsumes H) if there is a *projection* from G to H. A projection is a mapping π from the vertices set of G to the vertices set of H, which maps concept vertices of G into concept vertices of H, relation vertices of G into relation vertices of H, preserves adjacency (if the concept vertex v in V_C^G is the ith neighbor of relation vertex $r \in V_R^G$ then $\pi(v)$ is the ith neighbor of $\pi(r)$) and furthermore $\lambda_G(x) \geq \lambda_H(\pi(x))$ for each vertex x of G.

4.1 Statistical Conceptual Graph Rules

This section describes how to exploit the statistical correlations contained in textbooks to select appropriate classifiers for HealthAgents. Statements such as "Medulloblastoma account for 20% of all pediatric tumors" or "85% of medulloblastoma occur by the age of 15" are translated into Conceptual Graph (CG) based rules (as described in [2]) with the corresponding associated degree of belief. We provide the definition for such rules below.

If S is a fixed support, then a *rule* defined on S (see [2]) is any CG H, over the support S, having specified a bipartition $(Hyp, Conc)$ of its set of relation nodes V_R^H. The subgraph of H spanned by Hyp, $[Hyp]_H$ is called the *hypothesis of the rule* H, and the subgraph spanned by $Conc$, $[Conc]_H$, is the *conclusion of the rule* H.
Applying a rule H *to a CG* G means to find a projection π from $[Hyp]_H$ to G, to add a disjoint copy of $[Conc]_H$ to G, and finally to identify in this graph each concept node $v \in V_C^{[Conc]_H} \cap V_C^{[Hyp]_H}$ to $\pi(v)$, its image by π. The new CG

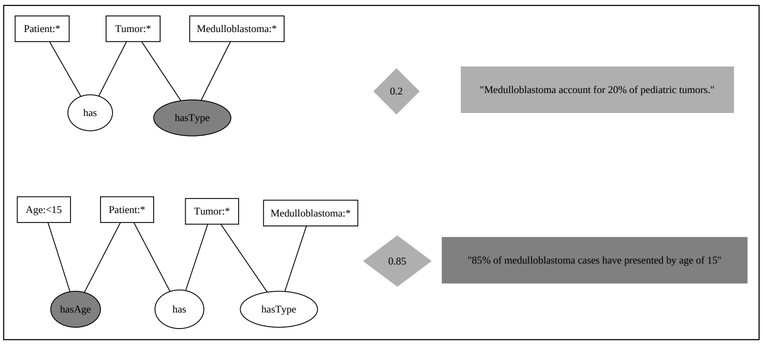
Patient:*
Tumor:*
Medulloblastoma:*
has
hasType
0.2
"Medulloblastoma account for 20% of pediatric tumors."
Age:<15
Patient:*
Tumor:*
Medulloblastoma:*
hasAge
has
hasType
0.85
"85% of medulloblastoma cases have presented by age of 15"

Figure 2. Conceptual Graph Probabilistic Rules

obtained, G', is called an *immediate derivation* of G, by the application of rule H, and following π. A *probabilistic rule* is pair $(R, p(R)$, where R is a rule and $p(R)$ is its probability.

In Figure 2 two such probabilistic rules for the tumor type medulloblastoma are presented. The first rule states that if the patient has a tumor (as encoded by the white labelled relation "has") then the tumor type is medulloblastoma (as encoded by the grey labelled relation "hasType") with a probability degree of 0.2. Similar, the second rule states that is a patient has a tumor and that tumor is of the type medulloblastoma then the patient is under 15 with a probability degree of 0.85. The support for these rules has been omitted for simplicity reasons. These two rules have been extracted from a pediatric study on tumor types and are the only two available rules for the tumor type medulloblastoma. This is an important fact, as it shows that the number of such correlation rules is not large, thus not affecting the computational effectiveness of our approach. We will show how these rules are applied for HealthAgents in the next section.

4.2 Conceptual Graph Derivation Tree

This section will detail how the rules introduced in the previous section can be used on a specific instance of a patient case. All of the relevant rules for the patient instance will be applied and a derivation tree built. The derivation tree will be used for the clinician to have an overview on potentially useful information prior to classifier selection. The weights on the tree edges will only be used as an indication of correlations in the field. Please note that due to the way we defined the derivation tree the same rule can be applied twice, therefore not ensuring independency. This is the main reason why we do not use the derivation tree for probabilistic inference, but rather for an organized exploration of the available information relevant to a particular case. It is also important to mention that the derivation tree cannot get potentially very large due to the number of available rules for each of the tumor types.

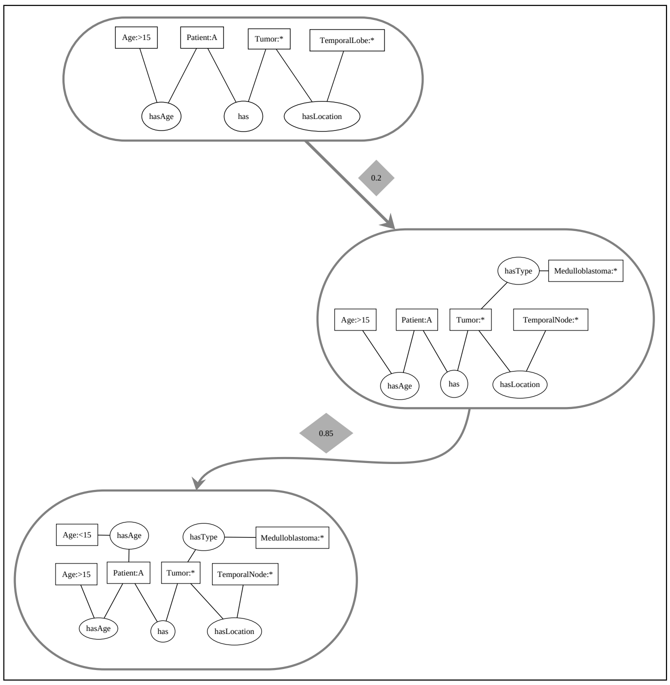

Figure 3. Patient Case Example

Let $\mathcal{R}$ a set of rules defined on S and G a CG over S. Then G, $\mathcal{R}$ derives a CG G' if there exists a sequence of immediate derivations leading to G' by applications of rules in $\mathcal{R}$. The set of all CGs G' which can be derived from a CG G using $\mathcal{R}$ by means of sequences of immediate derivations of length at most k is denoted by $\mathcal{R}^k(G)$ and can be described as a derivation tree having as nodes CGs, rooted in G and having as directed edges pairs of CGs representing immediate derivations. If the rules in $\mathcal{R}$ are probabilistic, then each such directed edge has assigned as weight the probability of the rule used.

Figure 3 presents such derivation tree obtained from a patient case of over 15, with a tumor in the temporal lobe. The clinician intuition (based on the MRS scan) is that medulloblastoma is an potential diagnosis and the two rules previously shown for medulloblastoma have been applied. As a consequence a contradiction was obtained: given the fact that medulloblastomas account for 20% of cases, 85% of those will be on patients under 15, and the patient was over 15.

Please note that if the clinician would not have any intuition on the tumor type, then all the rules relevant to tumor types and further consequences would have been applied. Even if the rule will state that for the particular instance tumor location a tumor type is not possible, the outcome will be presented to the clinician. The motivation is that a potentially new type of tumor could be under discursion and by performing "reasoning" this aspect would be ignored. It is therefore very important, in the context of this domain, to present the clinician with all possible information related to the patient case.

5 Conclusion and future work

In this paper we provided a methodology for integrating probabilistic information to enhance the HealthAgents decision support system. We have shown how the probabilistic rules retrieved from textbooks can be translated into a Conceptual Graph formalism and then how they can be applied for building a derivation tree.

In advancing out work we have to keep the knowledge representation and reasoning research tightly coupled with the clinician feedback in the domain. So far, the clinician have proved reluctant to discarding information as part of the reasoning process. However, future work will look at pruning the derivation tree based on contradiction and reorganizing information based on such pruning. We would also like to facilitate intuitive navigation of such tree and current work is looking at addressing such design problems.

References

1. C. Arús, B. Celda, S. Dasmahapatra, D. Dupplaw, H. González-Vélez, S. van Huffel, P. Lewis, M. Lluch i Ariet, M. Mier, A. Peet, and M. Robles. On the design of a web-based decision support system for brain tumour diagnosis using distributed agents. In *WI-IATW'06: 2006 IEEE/WIC/ACM Int Conf on Web Intelligence & Intelligent Agent Technology*, pages 208–211, Hong Kong, December 2006. IEEE.
2. J.-F. Baget and M.-L. Mugnier. Extensions of Simple Conceptual Graphs: the Complexity of Rules and Constraints. *Jour. of Artif. Intell. Res.*, 16:425–465, 2002.
3. M. Croitoru and E. Compatangelo. Conceptual graph projection: a tree decomposition-based approach. In P. Doherty, Mylopuolos, and C. Welty, editors, *Proc. of the 10th Int'l Conf. on the Principles of Knowledge Representation and Reasoning (KR'2006)*, pages 271–276. AAAI, 2006.
4. M. Croitoru, B. Hu, S. Dashmapatra, P. Lewis, D. Dupplaw, and L. Xiao. A conceptual graph description of medical data for brain tumour classification. In *Conceptual Structures: Knowledge Architectures for Smart Applications, 15th International Conference on Conceptual Structures, ICCS 2007*, 2007.
5. J. Halpern and D. Koller. Representation dependence in probabilistic inference. *JAIR*, 21:319–356, 2005.
6. T. Lukasiewicz. Expressive probabilistic description logics. *Artif. Intell.*, 176:852–883, 2008.