Enriched Property Ontology for Knowledge Systems

A thesis presented in partial fulfilment of the requirements for the

Degree

of

Master of Information Systems

In

Information Systems

Massey University, Palmerston North, New Zealand

Conducted at LBD EPFL, Switzerland

Robert Minchin

October 2006

Contents

1 Introduction

"It is obvious that every individual thing or event has an indefinite number of properties or attributes observable in it and might therefore be considered as belonging to an indefinite number of different classes of things" [Venn 1876].

The world in which we try to mimic in Knowledge Based (KB) Systems is essentially extremely complex especially when we attempt to develop systems that cover a domain of discourse with an almost infinite number of possible properties. Thus if we are to develop such systems how do we know what properties we wish to extract to make a decision and how do we ensure the value of our findings are the most relevant in our decision making. Equally how do we have tractable computations, considering the potential computation complexity of systems required for decision making within a very large domain. In this thesis we consider this problem in terms of medical decision making.

Medical KB systems have the potential to be very useful aids for diagnosis, medical guidance and patient data monitoring. For example in a diagnostic process in certain scenarios patients may provide various potential symptoms of a disease and have defining characteristics. Although considerable information could be obtained, there may be difficulty in correlating a patient's data to known diseases in an economic and efficient manner. This would occur where a practitioner lacks a specific specialised knowledge. Considering the vastness of knowledge in the domain of medicine this could occur frequently. For example a Physician with considerable experience in a specialised domain such as breast cancer may easily be able to diagnose patients and decide on the value of appropriate symptoms given an abstraction process however an inexperienced Physician or Generalist may not have this facility.

Accordingly Physicians may be precluded from providing a correct or rapid diagnostic that ultimately has adverse affects on the patient or leads to the requirement of possibly unnecessary medical tests. Historically diagnostic KB Systems have not been tremendously successful within the medical practice, other than as simple support tools. This is thought to be caused by:

a) the limited scope (useful for a small specific domain only) of such tools

b) the inability to handle conflicting symptoms

c) the lack of consideration of the diagnostic process used by doctors.

In order to overcome these barriers, we propose the use of an extensible property rich ontology for mapping domains of decisions, with each sub-class/domain associated with a reference class and a set of KB systems. This approach guides the system user to the core set of properties that should be targeted in decision making or querying and should increase the relevance of each property used in decision making. This approach uses the existing knowledge of ontologies, decision systems such as Bayesian networks and statistics. It combines these fields so that we are able to:

a) Query an ontology that maps the domain of decision via properties, not only by sub domains. Enabling the potential of scope to map very large domains.

b) Increase the power of our decision systems within the applicable sub domain. Allowing inference of diagnosis when conflicting symptoms exist.

c) Build domain knowledge with domain experts (Physicians) thus the specific abstraction or decision making process may be mapped.

We proposed that this methodology presented in detail in section 4 could equally be used in conjunction with existing KB Systems to increase their scope and precision, for example; integrating a specialised diagnosis system Athena (for hypertension) with a general diagnostic system DXplain.

1.1 Motivation

In this thesis we propose that a property rich ontology may represent a domain and map expert abstraction to sub domains of decision. This enables the possibility to define key property variables for classifying an unknown thing in a large domain. This would make each finding obtained potentially more valuable for decision making. Equally we propose that an effective method of defining what a thing is in machine computational terms and in a large domain would enable a considerable advancement in knowledge based systems i.e. navigation systems, aids for the disabled, security systems, Medical KB and other KB systems.

Vast/complex classification systems cannot be effective or efficient if the system does not have a method of targeting what properties it wishes to consider. Current systems are too limited in scope, do not offer solutions of objectively defining specific property extraction and are essentially non extensible. Difficulties in overcoming problems of classification in a wide domain are illustrated by the limitations of use of belief revision methods and pattern recognition as shown in section 2.

As a case study we are considering medical expert systems. There has been considerable development of medical expert, decision support systems or KB systems since the 1970s. However, KB systems have still only had a very limited effect on the medical practice largely because these systems are either very specialised, are only accurate in specific domains and unreliable in others or the systems are just too simplified [9], (in terms of scope). In addition these systems may not have the possibility of deduction of error when a suspect incorrect classification has been made.

Episodic skeletal plan refinement (ESPR) – A problem solving method that classifies and provides output on a defined protocol logic (skeletal plan) that is hierarchical and often time based to match medical treatment protocols. The skeleton is refined to the appropriate level of abstraction on an episodic basis (E.g. each patient visit).

Computational Complexity – Evaluation of the required resources used during computation to solve a given problem, considering how many steps it takes to solve a problem (time) and how much memory is required. Time or space required to solve the problem is considered as a function of the size of the input problem; for example the difficulty of finding a particular disease will become harder as we have a greater number of possible diseases and symptoms.

Current medical KB systems that use a system of episodic skeletal plan refinement (ESPR) [33] may well represent the temporal nature of patient treatment and medical guidance. However, these systems are only used for specific diagnosis/guidance process e.g. AIDS. In addition these systems may not consider reference class problems. A reference class is a like group or class having similar referenced attributes. The problem is that probability/inference is specific to a group or to a referenced class and should be interpreted according to the appropriate group. For instance, if we consider two different references, e.g. European Middle Aged Female vs. Polynesian Adolescent Male, the symptom inferences could be quite dissimilar. By using the reference class information we are potentially using known verifiable statistics to have more powerful variables in our classification and decision making systems for each reference group.

There is an obvious cause of limitations that affects all decision based systems, that is the complexity of making decisions in a large domain. Medical diagnoses processes are likely to have an extremely large set of properties, and in order to cope with this complexity experienced Physicians may work in different levels of abstraction by refining target symptoms. Thus if we are to improve the scope and accuracy of diagnostic KB systems we firstly need to look for a method of defining properties (symptoms and characteristics) that are relevant for given patient scenarios. Secondly we need to ensure that the relevant KB systems maximise the significance of variables in accordance with available findings, i.e. we need to consider the inference of reference class information.

In order to tackle these two issues we propose the use of a related four staged-approach:

a) The 1st stage being the design/formation of an extensible ontology considering the natural domains of decisions associated with a reference class or expert defined abstraction trees.

b) The 2nd stage being the collection of the information or statistics applicable to the class references.

c) The 3rd stage being the development of a set of KB systems associated with each domain from the reference class information.

d) The 4th stage being the enrichment of the ontology classes with attributes defined in the associated KB systems to create a property enriched ontology.

The three binary relationships between the ontology, Reference Classes and KB systems are illustrated in Figure 1-1. The ontology classifies the sub-domains of the universe of discourse and contains the properties that can be applied to each class or sub-domain (enriched). The reference classes are the statistics or information extracted considering the conditional implications of the considered sub-domain. The KB system(s) contains the decision formula(s) or model(s) used to define the next level of abstraction, constructed from the reference class information. We refer to KB systems in a plural sense with each domain as we put no restriction on the type of decision systems used in the methodology.

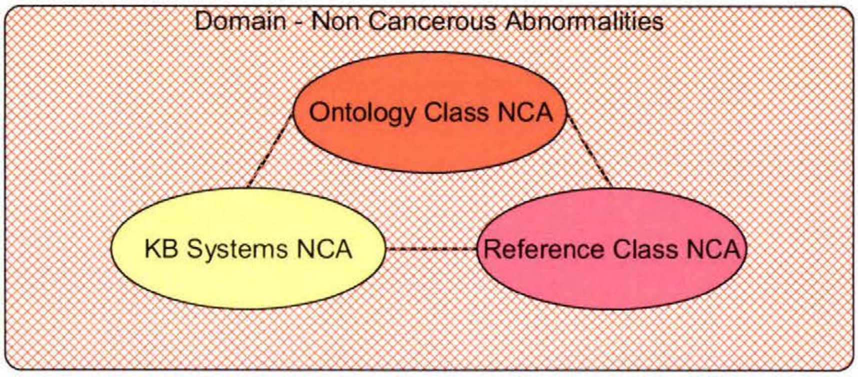


Figure 1-1 : Domain Triangle

The structure of querying implied is that we first verify whether a sufficient property exists to define a disease or a disease group (domain). Alternatively the disease domain can be established by the non sufficient properties or characteristics. From a specific domain or class reference the associated KB systems are used to target key properties for the decision. These target properties and findings then define the next level of decision or abstraction.

The symptoms ontology is static while being extensible i.e. a classification of disease symptoms is essentially fixed and will extend as new knowledge is learnt. The KB system or sets of worlds are likely to be reactionary to belief change and to the probabilities defined from reference class statistics, i.e. as we learn new information about a patient's characteristics or symptoms the implications of decision or diagnosis change. Thus an ontological mapping structure could enable the development of a vast database of diagnostic properties that could be queried effectively by symptoms/characteristics to classify patient disorders. This is possible because the data is stored in domain granules linked via the ontology and defined by class reference information. Such a structure manages the complexity of diagnostic methods. The structure increases the value of findings because unimportant symptoms for the domain are not requested and decisions or beliefs can be based on the associated reference class. In addition, such a system could know inherently when it has made an inappropriate domain allocation decision as new findings are added and could dynamically adjust i.e. when a conclusive decision cannot be defined.

In terms of medical diagnosis this could mean faster, more accurate diagnosis and reduced cost by reducing the number of medical tests to form an acceptable certainty in diagnosis. To demonstrate this methodology we have developed a prototype SOMKS (Symptoms Ontology for Mapping Knowledge Systems) that maps Knowledge domains of Breast Health using Stanford's Protégé tool and Netica developed Bayesian Networks to represent domain specific Knowledge bases.

1.2 Related Work

Classification systems generally avoid global decision domains in order that assumptions and prior knowledge about the domain can be applied; i.e. the problem is overcome by avoidance in creating systems that have very defined and limited domain of operation such as a sensory based quality control in chip manufacturing. Methods of belief revision. are introduced in later sections, have not had a large impact because they are restricted in terms of computational complexity or they are unable to provide rational revisions, as developed in [28]. Concepts of granular computing introduce the human decision making process of hierarchy and abstraction that we attempt to better map in our system.

For our case study we consider specifically Medical KB systems. The medical informatics community has built a considerable number of KB systems to aid medical practitioners in many ways. Recent systems address extensively temporal nature of medicine and use ontologies in defining medical protocol through processes including episodic skeletal plan refinement (ESPR) see [33]. The common complaints about these systems are that they are highly domain specific or are excessively general.

Rational Revision – Revision of a belief that meets the basic AGM postulates (section 2), for example when a revision is added to a belief formula and then removed the original belief formula should be obtained.

1.3 Our Contribution

The methodology that we have developed uses the pillars of existing knowledge concerning ontologies, KB systems and statistics. We combined these approaches to develop a manageable method on increasing the scope of classification or decision systems.

We recognise that there are many systems and proposals for managing classification operations. However it has been generally concluded that these systems, either, do not manage classifications/decisions well in a large domain, are excessively complex to be practically used or are just too simplified. In order to overcome these limitations in KB Systems, we propose the use of an extensible property rich ontology that maps target reference classes that have associated KB systems. The KB systems defined from reference class information directs the system to a set of target properties that can lead the system to a more precise abstraction or lower level reference class.

We are applying a granular approach to specify target properties, to increase the value in decision making of each property defined (finding) and to allow a system to potentially know when an inappropriate reference class has been defined and dynamically correct this. In addition the core of the ontology is essentially static and extensible. For example, LCIS is likely to continue to be defined as a type of non-invasive cancer with specific symptoms and if a new type of invasive cancer is defined it can be added to the super-class of Invasive cancer without affecting LCIS (see the complete ontology in appendix 1). The KB systems in turn could be dynamically adjusted via traditional methods while limiting impact of computation complexity.

1.4 Thesis Outline

In the preliminaries we introduce issues in pattern recognition, class reference and belief change affected by the limitations of scope in classification, computational complexity and KB systems. We then introduce Bayesian networks and OWL, that are used in our prototype SOMKS. We further discuss Medical Expert systems and potentially why they have had a less than expected impact on medical practice.

In the related work section we review the historic developments in belief change and discuss their limited use due to either not being rational or having complexity limitations. We introduce some of the principle medical KB systems. We also introduce the granular computing whose objectives relate strongly to our methodology.

In the theory section we review the advantages that our design and querying approach of property enriched ontology for mapping domains, would bring to KB systems. We then consider such an approach in conjunction with medical KB systems.

In section 5 we review the prototype SOMKS functionalities, and the tools/software used for its construction. In section 6 we introduce our case study using 'SOMKS' for the diagnoses of breast abnormalities. SOMKS uses a breast health ontology containing classifying properties of symptoms and patient characteristics that are then used to lead the system to a class reference domain. The user is then focused on a finite set of features that can be defined to diagnose patient abnormalities.

The prototype is outlined in figure 1-2. SOMKS ontology reasoner finds the most appropriate sub-class or domain granule from initial specified symptoms/characteristics. If SOMKS can not distinguish between a possible disorder and a healthy patient from initial information, it then requests additional information based on the defined key variables for the domain of decision using the knowledge reasoner. The knowledge reasoner then defines the next level of abstraction. Theoretically SOMKS should also enable the re-querying of the ontology with the new findings.

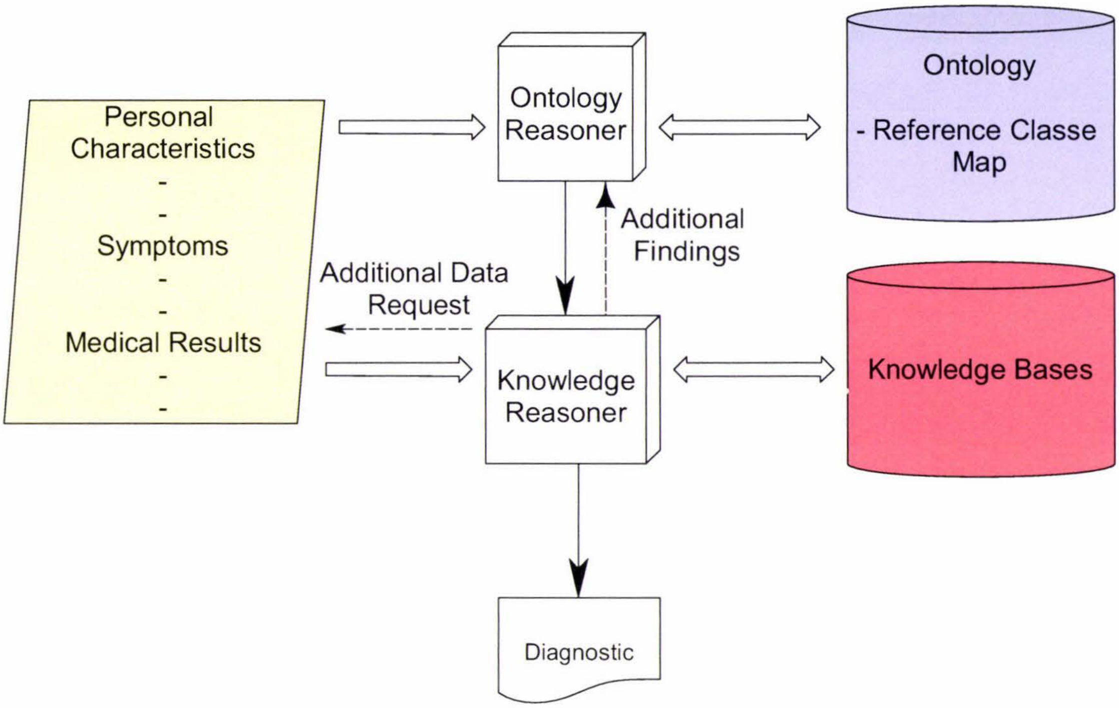


Figure 1-2: SOMKS Outline

2 Preliminaries

In this section we will review traditional KB systems and their limitations with reference to pattern recognition techniques, in order to demonstrate challenges in querying/decision making for unstructured and undefined data. Next we provide further explanation on Bayesian networks and RDF/OWL as they are the principle methods used in our prototype SOMKS.

We will be demonstrating how ontologies associated with reference classes can reduce the complexity associated with KB systems and belief change. Accordingly we examine the historic problem of class reference, computational complexity and methods of belief change. Finally in this section we review medical experts systems and the barriers they have had in their development.

2.1 Pattern Recognition

Pattern recognition is the act of sensing and gathering raw data, recognising patterns and taking an action based on the category of the data, e.g. classifying visual and sound patterns. Pattern recognition uses the techniques of KB systems to build knowledge bases and make decisions. The limitations of pattern recognition systems are illustrated in figure 2-1 below, which outlines pattern recognition process, demonstrating the requirement for prior knowledge of the domain we are considering.

Objects are sensed and are broken down into processable segments, which could be extremely difficult if the system has not first defined what it is looking for i.e. what shape or form. Next features are extracted by a sensory system, however the system must again know what it is looking for to define what type of sensory system e.g. shapes, colours, sounds, odours, etc. Then based on this information the system must decide on the importance of extracted features for its classification, but how does it do this in a human meaningful manner if it does not have prior specialised knowledge of the domain? Finally the system makes adjustment for cost and missing features and context, which would not be possible unless the system is operating in a specialised and known domain. Therefore we propose that useful pattern recognition systems are limited to being highly specific and applying to a known domain. For example; when considering facial pattern recognition, these systems are generally designed for a frontal sensed extraction of the facial features but challenges of accuracy may occur for a side or angular views.

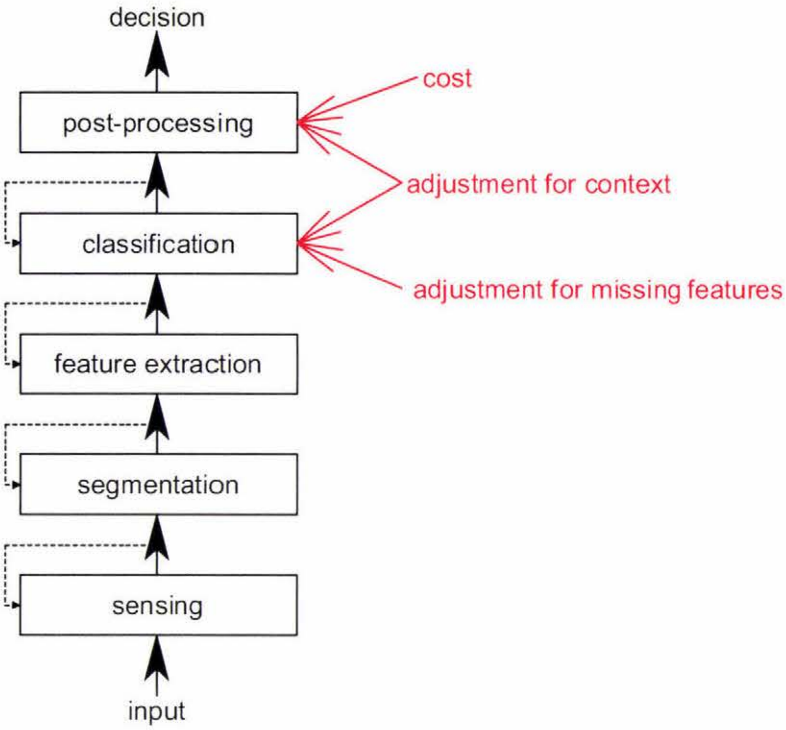


Figure 2-1 : Pattern Recognition [14 Duda]

We are proposing in our methodology to address these limitations by making use of a property enriched ontology to:

◊ Make decision making transparent.

◊ Portable between domains.

◊ More accurate in the domains in which we make categorisation.

◊ More easily able to adjust for context because context is meaningfully defined.

Pattern recognition techniques use a number of methods in which knowledge is stored and recognised but these methods are largely local or specialised for a particular domain or operation and are normally not necessarily based on classification as found in reality [12]. Fuzzy systems approach pattern recognition in a similar method to human expression i.e. allowing for generalisations, but systems depend often on correlation functions to form logic that may not necessarily relate to reality.

When a human looks at a road and sees a moving object of the appropriate size he/she is able through a process of abstraction to determine that the object is an automobile. When he/she sees a moving object at a horse race he/she would assume that it is a horse whether or not he/she has been able to adequately extract key features. These assumptions are based on a level of prior knowledge and knowing something and associating objects with certain boundaries and context through a process of abstraction. Thus if we want KB systems to mimic humans we need to have a method of defining our sub-context or sub-domain and relate this to possible objects to expected features through abstraction.

In pattern recognition the effectiveness and efficiency of extraction of features depends on the level of prior knowledge. Improved prior information improves opportunities for feature extraction of what a certain object is likely to be. Better Knowledge about object and sub-object features is likely to lead either reduced post processing or reduced probability of error in the classification of an extracted object. The more that is known about the situation of all the attributes of a particular pattern the easier it is to define the appropriate and most efficient analysis method for extracting patterns. For this reason pattern recognition is associated with very specialised functions.

Figure 2-2 below illustrates defining a sub-domain dynamically and applying a domain specific diagnosis system that follows the process of pattern recognition.

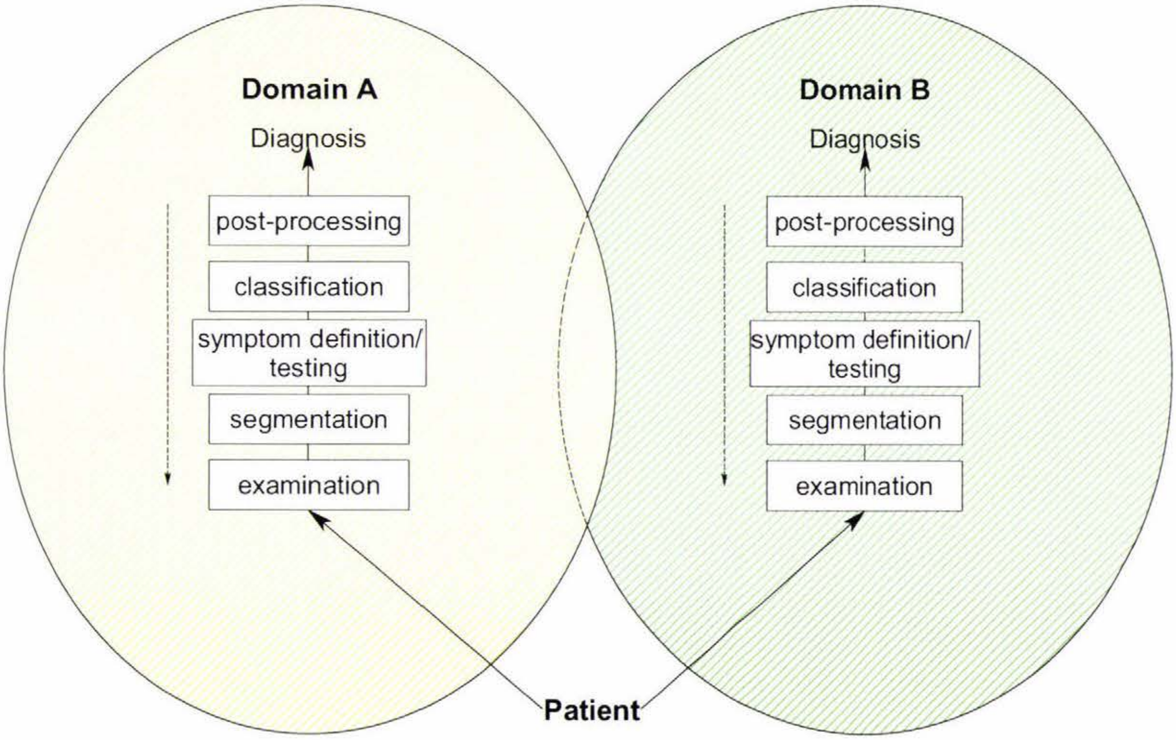


Figure 2-2: Dynamic Pattern Recognition

Later we demonstrate these proposals by building a medical diagnostic system for breast cancer. However first we will briefly review some of the methods of object classification used in pattern recognition and further introduce Bayesian Networks considering they are used in our Prototype.

2.1.1 Object Classification & Knowledge Representation

Pattern recognition is used for regional and object classification. Almost always when information about an object or region class is available some pattern recognition information is used. Indeed, no recognition is possible without knowledge. E.g. in [42] a Doorman is directing baseball players and jockeys to their appropriate meeting point. The Doorman would first start by asking the guests which meetings they are attending, from the guest's answers and their physical characteristics the doorman could soon form conclusions as to which meeting each guest is attending. This paper suggests that ontology's about objects will improve this knowledge base and therefore allow utilisation of sensed information much more effectively.

The main knowledge representation methods are formal grammars and languages describing features used to classify a pattern. Grammars, languages are made up of a structural description formed from existing primitives and the relations between them. Alternatives include:

Predicate Logic

Predicate logic is a type of formal logic that evaluates the validity of propositions in Boolean operations. The requirement of pure truth represents the main weakness as it does not allow for uncertainty or incomplete information that commonly occurs in sensing.

Statistical Pattern Recognition

Classifiers decide on classes from sensed object properties. Object descriptions (properties) are defined statistically based on defining features. With appropriate property selection, similarity of objects results in proximity of their patterns in a pattern space and thus makes possible assignment of class. However, the majority of pattern recognition problems do not have separable classes and in such cases the classes can not be separated correctly. Classes can often be better indicated by a property that is not statistically represented e.g. a knee being part of a leg.

Classification is normally based on a process of learning where the optimum classifiers settings are extracted from examples. The problem is "how to choose the best features from a set of available features and how to detect the features with the highest probability of recognition success" [42].

Bayesian Networks

Bayesian networks or belief networks use the theory of Reverend Thomas Bayes who defined the relationship between posterior probability and conditional probability, which has formed Bayes Theorem.

"The basic concept in the Bayesian treatment of certainties in causal networks is conditional probability. Whenever a statement of the probability, P(A), of an event A is given , then it is conditioned by other know factors." [25] Thus a Bayesian network is a model that reflects the states of some part of a world that is being modelled and it describes how those states are related by probabilities to define possible outcomes in terms of probabilities.

Neural Net

Neural Net is based on combinations of interconnected elementary processors (neurons) which take a number of inputs and provide output to other neurons until a final output is reached. Input is normally associated with a weighting and the output is a function of weighted sum. The concept mimics the high level of interconnection of elementary neurons found in the brain that are thought to map human decision making and explain the damage resistance and recall capabilities of humans.

Syntactic Methods

Syntactic methods use numeric quantitative descriptions of objects. The interrelationships or interconnections of features yield import structural information, for example, music. Elementary of syntactically described objects are called primitives, for example musical notes. "Where primitives have been successfully extracted all inter-primitive relations can be described syntactically as n-ary relations" [42], (tree grammars), for example tunes. Grammars can be used for representation of all patterns in their classes; a syntactic classifier can then be designed to assigns patterns to classes (classical, rock, etc).

Patterns are processed by either by Top down /Bottom up approaches. "The pure top-down approach is not efficient because too many incorrect paths are generated. Consistent approaches (tree pruning) can be designed that improve efficiency however this depends on the level of prior knowledge, "Many more consistency tests can be designed that take advantage of prior knowledge" [42].

Fuzzy System

"Fuzzy systems represent diverse non-exact, uncertain and inaccurate knowledge or information. They are qualifiers that are very close to the human way of expressing knowledge, e.g. bright, dark." [42]. Thus the system of logic is considered similar to human systems, i.e. formed where x and A represent properties and y and B are linguistic variables, if x is A then y is B. Rarely can a recognition problem be solved using a single fuzzy set (concept of partial membership of a set or assertion of possible membership) and associated single membership relation (indicating the degree of truth). Therefore tools are made that combine various fuzzy sets and allow one to determine membership functions. Fuzzy non-conditional and conditional (if-then) rules represent how fuzzy associated rules (knowledge) are stored.

Ontologies

"Ontology is a term borrowed from philosophy that refers to the science of describing the kinds of entities in the world and how they are related." [41]. An ontology is a systematic arrangement of all of the important categories of concepts which exist in a domain, showing the relations between them e.g. is a Sub-class, is a Super-class. It is a categorization of all of the concepts in some field of knowledge, including the classes and properties, relations, and functions needed to define the objects and specify their actions. Ontologies enable queries on non numerical basis i.e. based on categorisation, having a certain property or a sufficiency condition.

Other Systems

A considerable number of other alternative systems are detailed in [42] these include:

- ◊ Graph matching
- ◊ Optimisation techniques
- ◊ Production rules
- ◊ Semantic nets.

2.2 *Bayesian Networks*

In our Prototype SOMKS we use Bayesian networks as the Knowledge Bases for each domain or class reference and for that reason we have included this section to further explain their use. A Bayesian network is a directed graph of nodes representing variables and arcs representing dependence relations among the variables [25]. A node can represent any kind of variable, an observed measurement, parameter, latent variable, or hypothesis. Nodes are not restricted to representing random variables. Bayesian Networks utilise the law of condition probability from Bayes Theorem.

Bayes Theorem was named after Reverend Thomas Bayes, 1702-1761, a British theologian and mathematician who wrote down a basic law of conditional probability:

> For any two events, A and B,
>
> p(B|A) = p(A|B) x p(B) / p(A)
>
> Where p(A) is the probability of A, and p(A|B) is the probability of A given that B has occurred.

Bayes theory implies a degree of belief interpretation of probability, as opposed to frequency or proportion or propensity interpretations. This is particularly relevant for our ontology map because our class references do not necessarily define completely the possible outcomes but there is an inherent overlap or disjunction between our reference classes and accordingly we do not need to consider ever possible outcome.

Bayesianism is the philosophical tenet that the mathematical theory of probability applies to the degree of plausibility of a statement [20]. This also applies to the degree of believability contained within the rational agents of a truth statement. This is in contrast to frequentism, which rejects degree of belief interpretations of probability; frequentism considers probabilities only to random events according to their relative frequencies of occurrence. An example of a simple Bayesian Network shown in Figure 2-3 is taken form our SOMKS development

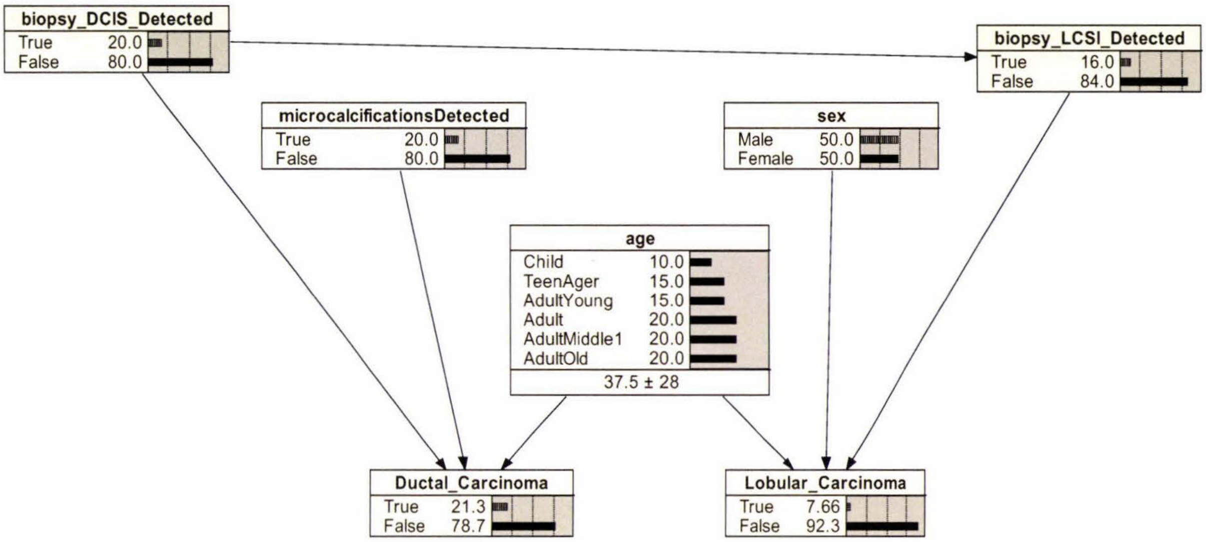


Figure 2-3: Bayesian Network for Non-Invasive Cancer

Each node in the network corresponds to some condition or characteristic of the patient, for example, "biopsy_DCIS_Detected" indicates whether a biopsy has taken place that resulted in DCIS being detected. The links between any two nodes indicate that there are probability relationships that are know to exist between the states of those two nodes.

To diagnose a patient, findings can be added to the nodes as they become known; all findings are by default unknown and have a base probability of occurring. From this information the Bayesian network will calculate possible probability results. For example, a new patient sex could be set to Female- 'True', Biopsy DCIS Detected – 'True'; this would result in an inference that the patient has a high probability that she suffers from ductal carcinoma and a low probability that the patient suffers from lobular carcinoma. This method is different to joint distribution that requires that we have a table of all the probabilities of all the possible combinations of states in a model, mean that the data could be very large and obtaining such data in accurate form would be very difficult. Accordingly Bayesian networks can adjust easily to belief change information without having a complete list of probability inference.

Given many of the advantages of Bayesian networks it is unfortunately inference of the most probable explanation are associated with computational complexity of NP- Hard [11] (see section 2.6).

2.3 Ontologies OWL/RDF

Ontologies are used to capture knowledge about a domain; they describe the concepts in the domain and also the relationships that hold between those concepts. They are very useful for non numerical operations for example if objects were defined in a form with related parts and environmental setting (general knowledge), systems would much better be able to form hypothesises of object identities with limited or related property information. For example a human knows that a hand is normally part of a human body thus assumes that the global object is a human which has physical features that would be proportion to the hand size.

The development SOMKS uses OWL descriptive logic; firstly because on the consideration of a possible web based development of SOMKS and secondly because of the wish to use Protégé OWL Java API. However Protégé or an alternative Ontological language could have been used.

The World Wide Web Consortium (W3C) recently contributed the web ontology language OWL standard that extends Recourse Description Framework Scheme (RDFS). OWL/RDF allows for complex concepts to be built up out of simpler concepts. OWL/RDF is machine understandable and the descriptive logic and lite versions allow the use of reasoners which can check the consistency of statements and definitions in the ontology (e.g. RacerPro). This paper proposes the use of ontologies in W3C form for pattern recognition this section reviews the basic concepts of OWL/RDF.

2.3.1 RDF

RDF is structured as one or more Triples: [42] the subject (what the data is about), [4a] the property (an attribute of the subject) and [7] the actual value. RDF is commonly expressed in XML (RDF/XML).

What RDF attempts to do is provide a resource description or a record of statements about a URI that a machine can understand and therefore classify. RDF allows one set of statements to be merged with another set of statements; even though the information contained in each set of statements may differ dramatically i.e. RDF statements can be stored in one file and accessed for different sections of information.

2.3.2 RDF Schema or Vocabulary

RDF properties may be thought of as attributes of resources, but RDF provides no means of describing the relationships between these properties and other resources. RDF Schema describes related resources and the relationships between these resources.

The RDF vocabulary description language class and property system is similar to the type systems of object oriented languages. However RDFS differs from OO systems; indeed, instead of defining a class in terms of the properties its instances have, RDFS describes properties in terms of the classes of resource to which they apply.

RDFS Classes

Resources are divided into groups called classes. The members of a class are known as instances of the class. Classes are themselves resources and may be described using RDF properties. Associated with each class is a set, called the class extension of the class, which is the set of the instances of the class. Two classes may have the same set of instances but be different classes (e.g. Canton Vaud and the Post). A class may be a member of its own class extension and may be an instance of itself.

The group of resources that are RDF Schema classes is itself a class called rdfs:Class. If a class C is a subclass of a class C', then all instances of C will also be instances of C'. The rdfs:subClassOf property may be used to state that one class is a subclass of another. The term super-class is used as the inverse of subclass. If a class C' is a super-class of a class C, then all instances of C are also instances of C'.

RDFS Properties

In addition to classes of things RDFS is able to describe specific properties that characterize those classes of things (such as four wheels to describe a car). RDFS uses the RDF class rdf:Property , and the RDF Schema properties rdfs:domain, rdfs:range, and rdfs:subPropertyOf to describe properties.

All properties in RDF are described as instances of class rdf:Property. So a new property is described by assigning the property a URIref, and describing that resource with an rdf:type property whose value is the resource rdf:Property, e.g.

```
ex:Person rdf:type rdfs:Class .

ex:author rdf:type rdf:Property .

ex:author rdfs:range ex:Person .
```

[24]

An rdfs:range - is an instance of rdf:Property that is used to state that the values of a property are instances of one or more classes. rdfs:domain is an instance of rdf:Property that is used to state that any resource that has a given property is an instance of one or more classes. The rdfs:subPropertyOf property may be used to state that one property is a sub property of another.

2.3.3 Ontologies / OWL

An ontology is a definition of a classification of something and associating properties with this definition. Ontologies are used to capture knowledge about a particular domain of discourse, describing the concepts within the domain and also the relationships that hold between those concepts.

OWL (Web Ontology Language) is a mark-up language for publishing and sharing data using ontologies on the Internet. OWL is a vocabulary extension of RDF (the Resource Description Framework). OWL has three variants of the language OWL Lite, OWL DL, and OWL Full [www.w3c.org]. As SOMKS uses OWL DL we do not consider OWL Lite and OWL Full further. OWL DL (Descriptive Logic) limits the expressiveness of OWL Full to remove the possibility of endless loops. This allows the use of reasoners that can check whether statements and definitions in the ontology are mutually consistent and can also recognise which concepts fit under which definitions. Thus a reasoner can control the hierarchy of the ontology and enable tractable querying that might not be impossible with OWL full.

The OWL extension of RDF addresses the following shortcomings to RDF schema [7]:

- ◊ Cardinality constraints on properties, e.g., that a Person has exactly one biological father.
- ◊ Specifying that a given property (such as ex:hasAncestor) is transitive, e.g., that if A ex:hasAncestor B, and B ex:hasAncestor C, then A ex:hasAncestor C.
- ◊ Specifying that a given property is a unique identifier (or key) for instances of a particular class.
- ◊ Specifying that two different classes (having different URIrefs) actually represent the same class.
- ◊ Specifying that two different instances (having different URIrefs) actually represent the same individual.
- ◊ Specifying constraints on the range or cardinality of a property that depend on the class of resource to which a property is applied.
- ◊ The ability to describe new classes in terms of combinations (e.g., unions and intersections) of other classes, or to say that two classes are disjoint.

2.4 Class Reference

Reference classes first identified by Venn [20] where Venn referred to the difficult is assigning probabilities and concluded:

> "This variety of classes to which the individual may be referred owing to his possession of a multiplicity of attributes, has an important bearing on the process of inference".

An individual or thing may be associated with many reference classes from which different probabilities will result. The reference class problem arises when we want assign a probability to a single event E, without considering unconditional probability. We often like to apply unconditional probability of E but there are in fact many conditional probabilities of the form P(E, given A) P(E, given B), etc

> "We cannot recover P(E) from these conditional probabilities by law of total probability since we lack the unconditional probabilities for A, B, etc" [20].

For the relevance of our consideration of Bayesian inference;

> "The Reference class approach is to equate the degree of belief in propositions about an individual with the statistics from a suitable chosen representation class i.e. a set of domain individuals with the statistics from a suitable class reference class" [27].

For example if a Physician wishes to describe a degree of belief of invasive cancer he would first try to obtain the most suitable reference class that he has statistics. Intuitively the reference class is a set of individuals of which the patient is a typical member [27]. E.g. the patient could be a male or a child, in which the probability would be sufficiently affected.

However there are explicit problems that the Physician will have in defining his reference classes:

- ◊ The patient will belong to several reference classes and the approach of finding the smallest possible set is not always appropriate i.e. there may be no record of male children having invasive breast cancer living in the Commune of Vevey, implying a degree of belief of zero.
- ◊ How does a Physician decide what information is relevant or irrelevant? Living in Vevey may not be important but the fact that the child is of European decent and parents have a cancer history could well be.
- ◊ How does a Physician decide what class he should use if he has 2 competing reference classes that have similar appropriateness?

The class reference problem is considered further in methods of belief change or models of adapting our beliefs to new information or events.

2.5 *Belief Change*

Belief revision an important topic for our consideration of the adaptation of KB systems given certain prior beliefs. Belief revision represents a rational agent's process of changing beliefs (old knowledge) to take into account new information (new knowledge) without generating an inconsistency. If we have new knowledge we must integrate this knowledge with old knowledge by considering minimisation of lose of old knowledge and managing confliction between the old and the new beliefs. What makes belief revision non-trivial is that several different ways for performing this operation may be possible, "Unfortunately there is no definite right way of relating statistical information to the degree of belief" [27].

The main assumption of belief revision, minimal change (the knowledge before and after the change should be as similar as possible) is affected by the relationships of the beliefs, for example:

> 'Top Gun' is to be played in one of two cinemas; where a and b indicate that the 'Top Gun' will be performed at the Odeon or at the Embassy, respectively ($a \vee b$). If we are advised 'Vanilla Sky' will be played at the Odeon, $\neg a$ holds; if we assume that we are referring to the same date and time, we could conclude that "Top Gun" will be played at the Embassy and not at the Odeon, which could be represented by ($\neg a \wedge b$).

Thus revising the belief with the new information can produce two different results $\neg a$ or ($\neg a \wedge b$) depending on how we relate the events a and b.

2.5.1 The AGM postulates

The AGM postulates (Alchourron, Gärdenfors, and Makinson) are properties that normative theories of belief change of what an operator that performs revision should satisfy to be rational [1]. In the AGM framework the main goal of an agent is to maintain the coherency of a belief when new information is inserted, and for this reason they are considered to be an example of the coherency approach to belief change. AGM considered 3 operations are considered given a theory K:

> Expansion - addition of a belief 'P' to the theory 'K' without a consistency check that results in an expanded set which is closed under logical consequences.
>
> Revision - proposition P inconsistent with a given theory K, is added to K under the requirement that the revision theory be consistent and closed under logical consequence (amendment).
>
> Contraction - removal of a belief where a proposition which was earlier in a theory K is rejected. The problem being to define what propositions should be rejected along with 'P' so that the contracted theory will be closed under logical consequence.

The principle formal problem of belief revision and contraction is to define ideal forms of change. For AGM, the current set of beliefs is represented by a deductively closed set of logical formulae K called belief base, the new piece of information is a logical formula P, and revision is performed by a binary operator * that takes its operators as the current beliefs and the new information and produces as a result a belief base representing the result of the revision. The + operator denoted expansion: K + P is the deductive closure of K $\cup \{P\}$. The AGM postulates for revision are:

I. $K * P$ is a belief base (i.e., a deductively closed set of formulae);

II. P is in $K * P$

III. $K * P \subseteq K + P$

IV. if $-P \notin K$, then $K * P = K + P$

V. $K * P$ is inconsistent only if P is inconsistent

VI. if P and Q are two logically equivalent formulae, then $K * P = K * Q$

VII. $K * (P \wedge Q) \subseteq (K * P) + Q$

VIII. if $-Q \notin K * P$ then $(K * P) + Q \subseteq K * (P \wedge Q)$

2.5.2 Contraction

The operators of revision and contractions are related by the Levi and Harper identities:

$$K * P = (K - \neg P) + P$$

$$K - P = K \cap (K * \neg P)$$

Whenever a revision operator satisfies the eight postulates for revision, its corresponding contraction operator satisfies the eight postulates for contraction, and vice versa [18].

The recovery postulate has been greatly discussed: $K = (K - P) + P$. According to this postulate, the removal of a belief P followed by the reintroduction of the same belief in the belief base should lead to the original belief base. However this is not always reasonable: in particular, the contraction by a general condition leads to the removal of more specific conditions from the belief base; it is then unclear why the reintroduction should also lead to the reintroduction of the more specific condition. For example, if Jo is believed to come from Lausanne, it would be believed that she is Swiss. Contracting the belief that Jo comes from Switzerland implies the retraction of the belief that Jo is Lausannoise. If it is then concluded that Jo comes from Zurich, the fact that she is Swiss is also reintroduced. According to the recovery postulate, however, the belief that she also comes from Lausanne should also be reintroduced. For Swiss Romand readers, the famous 'Roestigraben' has no relevance to this example.

2.5.3 Foundational revision

Working with deductively closed sets of formulae of the AGM postulates leads to the implicit assumption that equivalent belief bases should be considered equal when revising. According to foundational approach, retracting a non-derived piece of knowledge should lead to retracting all its consequences that are not otherwise supported (by other non-derived pieces of knowledge). This approach can be realized by using knowledge bases that are not deductively closed and assuming that all formulae in the knowledge base represent self standing beliefs.

For example the two equivalent sets {a, b} and {a $\wedge$ b}, revised by $\neg$a should produce different results. Case 1, a and b are two separate beliefs; therefore, revising by $\neg$a would have no effect on b, and the result of revision is {$\neg$a, b}. Case 2, is a single belief, 'a' being false contradicts this belief, which should therefore be removed from the belief base, thus giving the result {$\neg$a}.

2.5.4 Non-Monotonic Reasoning

Monotonicity indicates that learning a new piece of knowledge cannot reduce the set of what is known. Thus a monotonic logic cannot support reasoning tasks such as reasoning by default (facts may be known only because of lack of evidence of the contrary), abductive reasoning (facts are only deduced as most likely explanations). In belief revision new knowledge may contradict old beliefs. There are frame, ramification and qualification problems in this approach however non monotonic logic [5] is required to make temporary predictions explicit i.e. we want to maintain default conclusions e.g. when I stand up, my shoes are still on my feet.

Default reasoning

An example of a default assumption is that the typical bird flies. As a result, if a given animal is known to be a bird, and nothing else is known, it can be assumed to be able to fly. This fact must however be retracted if it is later learned that the considered animal is a penguin.

Abductive Reasoning

Abductive reasoning is the process of deriving the most likely explanations of the known facts. An abductive logic should not be monotonic because the most likely explanations are not necessarily correct. For example, the most likely explanation for seeing wet grass is that it rained; however, this explanation has to be retracted when learning that the real cause of the grass being wet was a sprinkler. Since the old explanation (it rained) is retracted because of the addition of a piece of knowledge (a sprinkler was active), any logic that models explanations is non-monotonic.

2.5.5 Merging

The assumption implicit in the revision operator is that the new piece of information P is always to be considered more reliable than the old knowledge base K. This is formalized by the second of the AGM postulates: P is always believed after revising K with P. More generally, one can consider the process of merging several pieces of information (rather than just two) that might or might not have the same reliability. Revision becomes the particular instance of this process when a less reliable piece of information K is merged with a more reliable P.

When merging a number of knowledge bases with the same degree of plausibility, a distinction is made between arbitration and majority. Arbitration involves maintaining as much of the original beliefs as possible and can be considered to be a union between the old and new belief. Majority implies that as we add new beliefs the old beliefs can be replaced when a majority of opinion is obtained. If the opinions of the new beliefs are contradicting then the old belief may remain in place if a majority of opinion supporting the original belief is maintained.

2.5.6 Complexity

One of the principle problems of belief revision in terms of computational complexity is confirming that Q is a logical consequence of K*P. That is if P is true, can Q be derived from K*P. To answer this problem we need to consider the whether 'K' is entailed by the result of a belief revision, which could be the result of an update, merge, revision, iterated revision, etc. The computational complexity of entailment for mainstream belief revision methods is considered in detail in "On the Complexity of Propositional knowledge base revision, updates and counterfactuals" [15]. It is shown that Polynomial time only occurs for Model based approaches where 'P' is bound and K, P & Q are taken to be conjunctions of Horn clauses and not at all for formula based approaches.

Another area of considerable research is model checking, which relates to model based revision methods that are described in section 3, where a knowledge base 'K' can be represented by a set of models M(K). Here complexity of model checking is evaluating whether a model M' $\in$ M (K * P). Liberatore [28] stated that "model checking for almost all operations is in the second level of the polynomial hierarchy" and concluded that "model checking for belief revision and update is far more complex than for classical proposition logic".

Complexity is equally an issue when considering whether the belief revision should include a preference relation. A preference relation can be represented by a sequence of formulae whose models are increasingly preferred. Storing the relation as a set of pairs of models is often not a compact representation of preference because the space required is exponential in the number of propositional letters.

2.6 Computational Complexity

Computational complexity evaluates the required resources used during computation to solve a given problem, considering how many steps it takes to solve a problem (time) and how much memory is required. Time or space required to solve the problem is considered as a function of the size of the input problem.

Computational problems are evaluated in terms of complexity classes. The class P denotes the set of problems whose solution can be found in polynomial time, while NP denotes the class of problems that can be resolved in polynomial time by a non-deterministic Turing. Polynomial time refers to the computational time is no greater than the polynomial function $O(n^2)$ by a deterministic Turing machine. 'n' being the number of nodes being or bits in the problem. Problems in P are considered to be tractable whereas problems in NP are considered to be intractable (problems solvable in theory but not in practice).

Although complexity analysis is important for our consideration, it should also be noted that the problems of open domains considered globally are likely to have a infinite or very large set of 'n' implying that even if the problems are tractable they may not be realistic to compute without granulation. Computational complexity is covered in detail in "Computational Complexity" [34].

2.7 Decision Support Systems & Expert Systems

Decision Support System (DSS) is information and planning system that provides the ability to interrogate data on an ad hoc basis. Expert Systems simulate expert problem solving; these systems need to process data and numeric relationships and by reasoning transform this data into opinions, judgement, evaluations and advice. The functionality of these systems is combined and called Knowledge-based Decision Support Systems (KB-DSS) [27]. For the sack of simplicity we will refer to systems as KB Systems.

2.7.1 Medical Expert Systems Problems

Computerised clinical decision support systems CDSSs or expert systems or Knowledge Based (KB) systems attempt to mimic the decision making capabilities of human experts.

KB systems contain clinical knowledge, usually about a very specifically defined task, and are able to reason with data from individual patients to provide conclusions. KB systems have been developed to provide diagnostic assistance, treatment guidance, systems of alarms warnings/reminders and the often interact with medical reference data. Although there are many barriers to the success of KB systems there is a recognition of their potential usefulness. For example when a patient's case is complex, rare or the person making the diagnosis is simply inexperienced or a non specialist in that domain of medicine, a KB system can help in the formulation of possible diagnoses based on patient symptoms.

> "Diagnostic support is often needed with complex data, such as the ECG, where most clinicians can make straightforward diagnoses, but may miss rare presentations of common illnesses like myocardial infarction, or may struggle with formulating diagnoses, which typically require specialised expertise" [9].

There are numerous reasons why more CDSS are not in routine use [9]. Some require the existence of an electronic patient record system to supply their data, and most institutions and practices do not yet have all their working data available electronically. Others suffer from poor human interface design. Much of the initial reluctance to use CDSS is considered to occur because KB systems did not fit naturally into the process of care, "and reluctance or computer illiteracy of some healthcare workers" [9].

The potential benefits of using electronic decision support systems in clinical practice fall into three broad categories [9].

- ◊ Improved patient safety e.g. through reduced medication errors and adverse events and improved medication and test ordering;
- ◊ Improved quality of care e.g. by increasing clinicians' available time for direct patient care, increased application of clinical pathways and guidelines, facilitating the use of up to date clinical evidence, improved clinical documentation and patient satisfaction;
- ◊ Improved efficiency in health care delivery e.g. by reducing costs through faster order processing, reductions in test duplication, decreased adverse events, and changed patterns of drug prescribing favouring cheaper but equally effective generic brands.

KB Systems have had problems in their take up and have had a less than expected impact on medical practice because medical expert systems were found only to be accurate within narrow ranges of medical topics, and they sometimes fail badly ("limited in range") . Medical KB Systems or knowledge bases are considered not to contain "world knowledge" and do not necessarily have in depth knowledge that a doctor may have. The potential drawbacks of Medical KB Systems, described by the medical practice include (from /www.openclinical.org):

◊ Potential 'deskilling' effect

◊ Can be perceived as a threat to clinical judgment

◊ Can be considered too inflexible (can appear prescriptive, can appear to direct proceedings; can be difficult to depart from ordered, pre-prepared paths)

◊ Promote over-reliance on software; limit clinicians' freedom to think?

◊ Difficult to evaluate - lack of accepted evaluation standards

◊ Can be time-consuming to use, possibly lead to longer clinical encounters and create extra work

◊ Uncertain and untested ethical and legal status

◊ Costs: maintenance, support and training required after initial outlay

◊ A clinician's experience and imagination cannot be duplicated in a computer application.

These factors and legal restrictions must be considered by a medical organisation before such systems can be implemented in practice. This thesis addresses methodologies for improving decision systems and does not consider the individual ramifications of their use. We will review some of the historic Medical KB systems in the next section.

3 Related Work

Our methodology is aimed at management of complexity given initially obtained attributes or evident sample information, for example the patient is a man. Methods of classification or decision given a base set of information are general considered in terms of belief change and granular computing and accordingly we review work in these areas in this section. In addition to have further background on our case study we also review milestone medical KB systems.

3.1 Belief Change

Referring to Winsletts [45] framework for comparison of semantic update methods we review principle update methods Model based versus Formula based approaches.

3.1.1 Model Based Approach

Model based approaches operate by selecting the models of P on the basis of some notion of proximity to the models of K. In model based approaches K is a single formula, thus if K is a set of formula it is implicitly interpreted as the conjunction of the elements. Model based approaches are based on the individual models of K rather than formulas within K. P is applied to each model individually, when P is applied to a model 'M' a set P(M) of models is produced. Winslett [45] classifies model based approach as satisfying the below rules, where 'x' is the update formula and 'P' is the update:

- ◊ MB1: the formula 'x' must be true in every model P(M).
- ◊ MB2: the result of applying P to K is a theory whose set of models is that set resulting from applying P to each member of Models (K).
- ◊ MB3: the effect of P on M is independent of the other Models i.e. no other information other than M and P need be considered.
- ◊ MB4: assuming that there are no formula rules in K asserting truth of a given variable. Then for any model M' in P(M), M and M' agree on the truth valuations of all propositions not appearing in 'x'.

According to these rules, the result of revising/updating a formula K by another formula *P* is expressed by the set of models of *M* that are the closest to the models of K. Different notions of closeness can be defined, leading to the difference among proposals, for example:

- ◊ Dalal 1988 [13]
- ◊ Satoh 1988 [38]
- ◊ Winslett 1990 [46]

3.1.2 Formula Based Approach

The Formula Based approach does not examine the models of K but rather the formulas in K, the formulas are methods of change and not propositional truth variations as in the model based approach above. Informally P is obtained by adding 'x' to K, unless the result is inconsistent. If {K, x} are inconsistent the inconsistent formulas (as few as possible) are removed from K. before adding 'x' so that {K, x} is consistent.

I.e. Given a theory K and an update "insert" 'x', a subset S of formulas of K is a minimal set if

{K-S, x} is consistent and,

{K-S', x} is inconsistent, for S' any proper subset of S.

However the problem being that the principle of removing only a minimal set, conflicts with a number of proposals where multiple minimum sets of formula for removal, for example:

- ◊ The set Theories Approach [16]
- ◊ The Cross Product Approach [17]
- ◊ WIDTIO (When In Doubt Throw it Out) [45]

3.1.3 Criticisms Belief Change

Belief change methods are mainly theoretical concepts versus systems that can be effectively used [8, 1], relating to limitations including the following:

- ◊ Belief change essentially deals with the case where the world has not changed, only the agent learnt more about the world. But an agent may also receive new information concerning changes that have occurred in the world [21].
- ◊ Belief change normally assumes that last thing learnt is the most important; however with such a methodology we cannot consider the strength of a belief, i.e. certain beliefs may be retained despite new conflicting information [21]. In addition where we do represent strength of beliefs, deciding when an update is accepted is non trivial.
- ◊ Finally and most importantly belief revision approaches do not overcome the computation limitations in decision making systems and generally imply exponential increases in complexity [28]. Most revision systems suffer from the problem of representational blow up: i.e., revision may exponentially increase the size of the knowledge base [45].

3.2 Granular Computing

Granular computing is a category of multi-disciplinary methodologies and techniques that make use of information granules in the process of decision making. It concerns processing of complex information granules, which arise in the process of abstraction of data and derivation of knowledge from information (information granulation). Information granules are a collection of entities (e.g. female patients aged between 45 and 55), that are arranged together due to their similarity, functional adjacency, coherency, etc. The notions of granular computing may be interpreted in terms of abstraction, generalization, clustering, levels of abstraction, levels of detail, and so on in various domains.

A hierarchy represents different levels of granularity in granular computing. The basic ingredients of a hierarchy are levels; the levels are linked together by a partial order. A level is populated by, granules whose properties characterize the level. Levels may be considered as parts, and the partial order describes the relations between, or dependency of, parts. Under the partial order, parts are arranged inside a whole described by a hierarchy. The general principles of hierarchical analysis and granular computing, understanding of the whole in terms of its parts and understanding of the system based on its inherent internal structures, are almost universally applicable. In practice, when describing a specific system, one may impose additional system dependent constraints and interpretations.

Granular computing is considered advantageous for the following reasons [23,48,35]:

- ◊ Good representation of the real world. Many natural and artificial systems are naturally organized into hierarchical systems.
- ◊ Consistent with human problem solving. Human problem solving is based crucially on levels of granularity and change between granularities.
- ◊ Simplification of problems. By omitting unnecessary, irrelevant details and focusing on the right level of abstraction, we are able to simplify a complex system, or a complex problem.
- ◊ Economic solution. By considering a problem at different levels of granularity, we ignore some details, and potential save processing time and/or cost of unnecessarily information extraction.

In 1985 Hobbs proposed a theory of granularity [23], which proposes that we as humans perceive and represent the world under various grain sizes, and abstract only those things that serve our present interests. The ability to conceptualize the world at different granularities and to switch among these granularities is fundamental to our intelligence and flexibility. This enables us to map the complexities of real world into computationally tractable simpler theories and apply to consider what is relevant and to ignore irrelevant details (abstraction).

Zhang (1997 - 2004) developed the quotient space theory of problem solving based on hierarchical description. The quotient space theory proposes to conceptualize computing at different granularities and consider a problem space from a hierarchy and abstraction levels.

3.3 *Medical KB Systems*

In this section we review briefly some of the milestone Medical KB systems used for diagnosis and treatment guidance.

3.3.1 Leeds Abdominal System

LAS was developed in the 1960s using disease incidence data for defining signs and symptoms, that were related in a bayesian network. Obtained patient data was then applied to the bayesian network to calculate the probability of possible explanations for acute abdominal pain (Appendicitis, Diverticulitis , Perforated ulcer, Cholecystitis, Small bowel obstruction, Pancreatitis). [www.openclinical.org].

304 patients were analysed and the diagnosis from the system was compared the results provided by normal physicians. LAS was 91.8% accurate compared with physicians who were 80% accurate. However when the LAS was tested outside of Leeds the results were never as accurate.

LAS unlike the SOMKS addresses a very controlled problem set only and would likely be impractical in a larger domain.

3.3.2 MYCIN

MYCIN is a simplistic system by today's standards however it is considered a milestone in medical KB systems. MYCIN developed in the 1960s by Dr Edward Shortliffe, incorporated knowledge in packets called production rules i.e. it contained 500 rules (If-Then) [32]. MYCIN was designed to diagnose infectious blood diseases and recommend antibiotics, with the dosage adjusted for a patient's body weight.

Although research found MYCIN to have a accurate diagnosis rate of 65%, which at the time was considered superior to non specialist, MYCIN was never used in practice because of ethical and legal issues related to the use of computers in medicine i.e. who would be accountable for a faulty diagnosis.

MYCIN compared to the SOMKS prototype is a simplistic system; it does not have any method of abstraction. This type of system would be limited to very specific diagnostic domains because of the computation issues of combining production rules.

3.3.3 DXplain

DXplain is a KB system developed at the Laboratory of Computer Science at the Massachusetts General Hospital, based on a modified form of Bayesian logic [3]. DXplain was first developed in 1984, containing information on 500 diseases. DXplain currently supports more than 2200 different diseases; with an average disease description of 53 findings. Each disease/finding pair has two numbers describing the relationship, one representing the frequency with which the finding occurs in the disease and the other the degree to which the presence of the finding suggests likelihood of the disease. There are over 230,000 individual data points in the KB representing disease/finding relationships.

DXplain uses inputted sets of clinical findings (signs, symptoms, and laboratory data) to produce a ranked list of diagnoses which might explain the clinical manifestations. Then, DXplain is able to provide details of possible diseases, and target additional clinical information collection that could be used for verifying each suspect disease.

DXplain unlike the SOMKS prototype is an altered Bayesian network only. Bayesian networks have a comparatively high build & maintenance cost and are likely to be limited in scope without introducing a method of abstraction.

3.3.4 Onconcin

Onconcin was developed at Stanford University as a decision support system for cancer chemotherapy management. Onconcin was first employed at the Stanford Oncology Clinic in 1981 where it provided treatment advice to Physicians prescribing experiment chemotherapy to cancer patients [26].

Onconcin involved the use of Knowledge representation or structured representation of clinical medicine in particular the representation of clinical treatment protocols based on if-then rules Protocol decisions such as drug dosage current treatment plan, current patient laboratory, toxicity findings and pertinent past patient features. Onconcin was one of the first KB Systems which attempted to model decisions and sequencing actions over time. It extended the skeletal-planning technique to an application area where the history of past events and the duration of actions are important i.e. dosage decisions would be made on information relating to last visits.

Onconcin like the SOMKS includes a method of abstraction. However, only on a pure temporal bases for very a specific treatment path.

3.3.5 Athena DSS & EON

ATHENA (Assessment and Treatment of Hypertension), provides guidelines for hypertension using EON architecture considering blood pressure control. ATHENA consists of two principle components, a KB (a protégé produced Ontology) that models Hypertension knowledge independently of its use and a guideline interpreter that creates patient specific use [19].

The EON project a NLM funded project from Stanford Medical Informatics, is a guideline modeller that uses the Protégé environment to create and maintain domain concept ontologies of concepts and relations in patient information model guideline KBs. It uses a meta-class facility to define the top-level categories (e.g., a laboratory-test result) and their attributes. It uses the class/subclass relationship to define hierarchies. From the ontology of guidelines and protocols, it generates forms for the acquisition of individual guidelines.

EON applies episodic skeletal plan refinement (ESPR) in the a EON based theory planning system for aids patients example from [33] for

- ◊ Determining whether a patient is eligibility for various protocols and guidelines
- ◊ Determining temporal based therapy recommendation for each clinical visit of a given patient.

The facility that the developers have in this system is that they have first defined a general diagnosis and they can then apply specific context rules with in that domain. However how do we first get a patient into a domain and adjust knowledge or probabilities based on a specific decision domain?

EON systems offer standardized ontologies for temporal based diagnostic and treatment. Unlike the SOMKS prototype EON probably does not offer an abstraction method that could be used in a complex diagnostic situation (large domain).

3.3.6 Others

There have been a development of a large number of diagnostic and guideline tools, all using a range of methods to store knowledge and provide decision support to Physicians. As the subject is vast, we introduce only a few of the principle models. Similar to the Onconcin and EON these models do not propose a method of abstraction in a diagnostic decision system where the patient has an unknown disorder or disease:

Asbru – collaboratively developed by Vienna University of Technology and Stanford Medical Informatics, released in 1998. Asbru is time oriented, intention based, skeletal-plan specification language that is used to represent clinical protocols. Skeletal plans capture the core procedure and leave room for execution time flexibility in the achievement of particular intentions.

GUIDE – developed at the University of Pavia, Italy, as illustrated in Figure 3-1 supports integrating modelled guidelines into organisational workflows and using decision analytical models such as decision trees and influence diagrams, and stimulating guideline implementation.

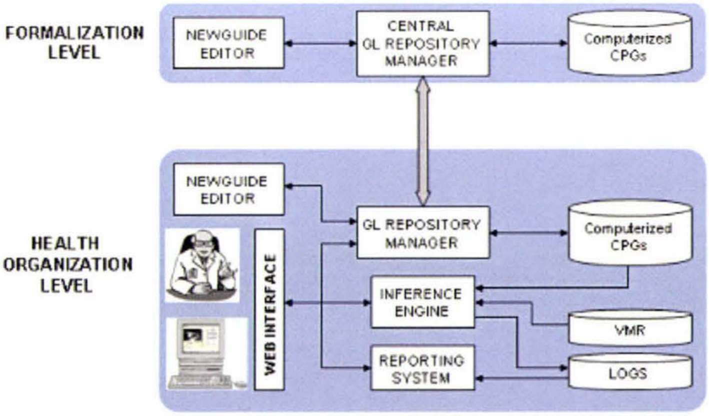


Figure 3-1: Guide [http://www.openclinical.org]

PRODIGY – developed at the University of Newcastle upon Tyne for the support of chronic disease management in primary care. It aimed at supporting simplest most readily comprehensive model necessary to represent class of guidelines. Essentially, it supports guideline modelling, a series of decisions that a Physician may have to make in different patient encounters. The model enables a guideline to be organised as a network of patient scenarios, management decisions and action step which produce further scenarios. Scenarios are patient states defined by the patient's condition and current treatment. Scenarios are associated with, firstly; a consultation template that describes the best-practice workup for a patient in that scenario, secondly; a choice between alternative courses of action. The management over time of a patient according to a guideline specification can be viewed as the traversal of a number of selected scenarios and associated actions and further decision points along a single path. Sequencing of actions is achieved by defined followed by relations.

PROforma – was developed at the Advance Computational Laboratory of Cancer Research, UK. Applications built using PROforma software are designed to support the management of medical procedures and clinical decision making at the point of care. The notion of a task is central to the model, the PROforma task model divides from the keystone into four types: plans, decisions, actions and enquiries:

◊ Plans are the basic building blocks of a guideline and may contain any number of tasks of any type.

◊ Decisions are taken at points where options are presented, e.g. whether to treat a patient or carry out further investigations.

◊ Actions are normally clinical actions.

◊ Enquiries are typically requests for further information or data, required before the guideline can proceed.

4 Theory

We as humans are able to handle the complex issues of applying classification continuously. We use complex structures of classification that have been learnt to make decisions. These classifications are applied in levels that are adjusted as we almost instantaneously extract property information to reclassify and make decisions [23]. Unfortunately currently we don't have machine abstraction processes as efficient as human processes.

The methods of belief change or using Episodic Skeletal-Plan Refinement (EON etc) are indeed very valid but may lack levels of prior decision if we want to produce KB systems that combine accuracy, global scope and dynamic adaptability.

We propose that in order to have computation systems that can mimic and approach the efficiency of human processing in a large domain, there are three fundamental problems that must be overcome.

◊ Firstly we need to select a granule of the universe of discourse based on available findings, in order to reduce the decisional space.

◊ Secondly we have to have a method of maximising the importance of our decision variables used (increased implication of property variables to the decision), given certain circumstances. According we propose the harnessing of the power of reference classes from the reference class problem.

◊ Thirdly we need a reusable and extensible method of defining hierarchy and relationship in a non numerical sense that can guide us to the applicable reference class by querying based on properties and classification. Accordingly we propose that a property enriched ontology may be used to carry out this functionality.

In this section we review our proposed methodology, query system and consider the advantages.

4.1 Reference Class

Our approach for domain modelling includes the use of reference classes. The reference class specifies a probability of an event given conditional circumstances. For example if the patient is a man the probability that he suffers from Fibroadenomas is 0% thus a Physician need not consider symptoms that are associated with this abnormality, however a women can suffer Fibroadenomas and accordingly a Physician may want to consider related symptoms.

Now if we want to define KB systems for domains we must find the appropriate reference class for our decision making process [20, 10]. For example if a Physician wishes to describe a degree of belief of invasive cancer he would first try to obtain the most suitable reference class that he has statistics. Thus we propose to use the reference class problem to our benefit by defining ontology class domains that are associated with defined reference class e.g. "Non Cancerous Abnormality Defined". The decision rules can then be defined from obtained statistics for this domain.

4.2 Ontology Domain Mapping

If we are to have KB systems that can manage expansive decision making we also need to mimic humans in the preliminary classification process of defining the domain in which they are making the decision or the class reference. To construct a similar decision making process we propose the development of an extensible ontology that determines the classification based on human defined properties. This ontology would then be used to imply different sets of contexts that can be controlled and applied within the limits of computational complexity.

4.3 Methodology

We are trying to have KB systems that can manage expansive decision making that mimic human specialists in the classification process of defining the domain in which they are making the decision. This allows us to target specific findings and apply decision rules that are specific to a reference class. To construct a similar decision making process, we propose the development of an extensible ontology that defines the classification based on human defined properties. The class being obtained by a recursive query operation to define the lowest common super class of the input symptoms or characteristics. The methodology of design that we propose is as follows:

- ◊ Domain experts define an ontology with the ontology classes representing reference classes of diagnosis e.g. patients with breast abnormalities detected.
- ◊ The associated statistics from each reference class are used to specify KB System sets for each domain.
- ◊ Attributes of ontology classes are extended to include associated attributes from the KB Systems that can imply further classification.

By introducing domain granules we are reducing the problem space that reduces the complexity of KB system inference, the problem space being the nodes (classes), relations and attributes. As we put no limitation on the type of KB system we will refer to the algorithm of complexity of KB system inference as $O(g(n))$. O notation is from complexity theory i.e. $f(n) = O(g(n))$, implies f grows at a rate $\leq$ g, allowing the generalisation of computational system used.

If 'n' is the problem space for the global domain and 'm' is the problem space of a domain granule, then $n \geq m$, and $O(g(n)) \geq O(g(m))$. m is normally an improved version of n' (where n' is a domain granule with unaltered probability implications), given that the reference class information is used to increase the importance of each variables decision implications within the selected granule (increased precision). Now if 'p' is the problem space for the ontology mapping the properties, normally $n \neq p$ because 'p' contains classification properties & restrictions and n will contain properties that can't be used in classification inference, thus $n+p \geq n$. Thus the route problem space has potentially increased.

The ontology expressiveness 'E' together with the granulised KB system expressiveness K' is normally more expressive than K alone ($E+K' \geq K$). Expressiveness in this case being the types of properties, relations and restrictions that can be used within the domain being modelled. Similarly our approach is more powerful than a hybrid Bayesian network using a two step ontology/KB System querying process, because KB(m) is more precise than KB(n'). Considering each variable in KB(m) will have greater inference given the improved reference class information.

The methodology initially increases the problem space and the expressiveness of the universe of discourse. Therefore design restrictions are made on granules such that $O(g(m))+O(f(p)) \leq O(g(n))$, where f(p) is the computational complexity algorithm of querying the ontology.

4.4 Querying

The system of querying that we propose is as follows:

◊ Initial symptoms obtained are queried in the ontology to define the lowest level super-class that contains all classes with the specified symptoms. For example from figure 4-1, a lump symptom being identified would imply the class 'Abnormalities Detected' and not the alternative class 'No Abnormalities Detected Screening'.

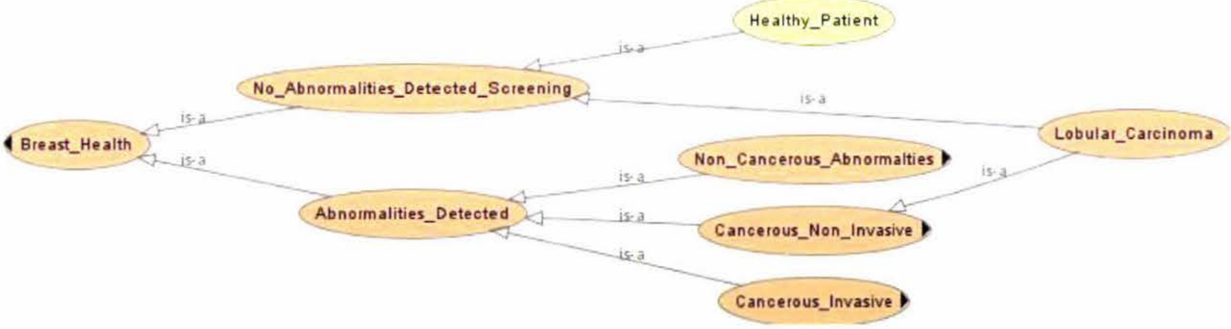


Figure 4-1: Breast Health Ontology, Upper Level Super-classes

◊ Next, the associated KB systems define targeted attributes that are significant in forming a recommendation for the choice of class for the next lower level within the super class. For example from Figure 4.2, the patient characteristics would be used below as indication of which subclass to proceed to i.e. here the indication is cancerous invasive.

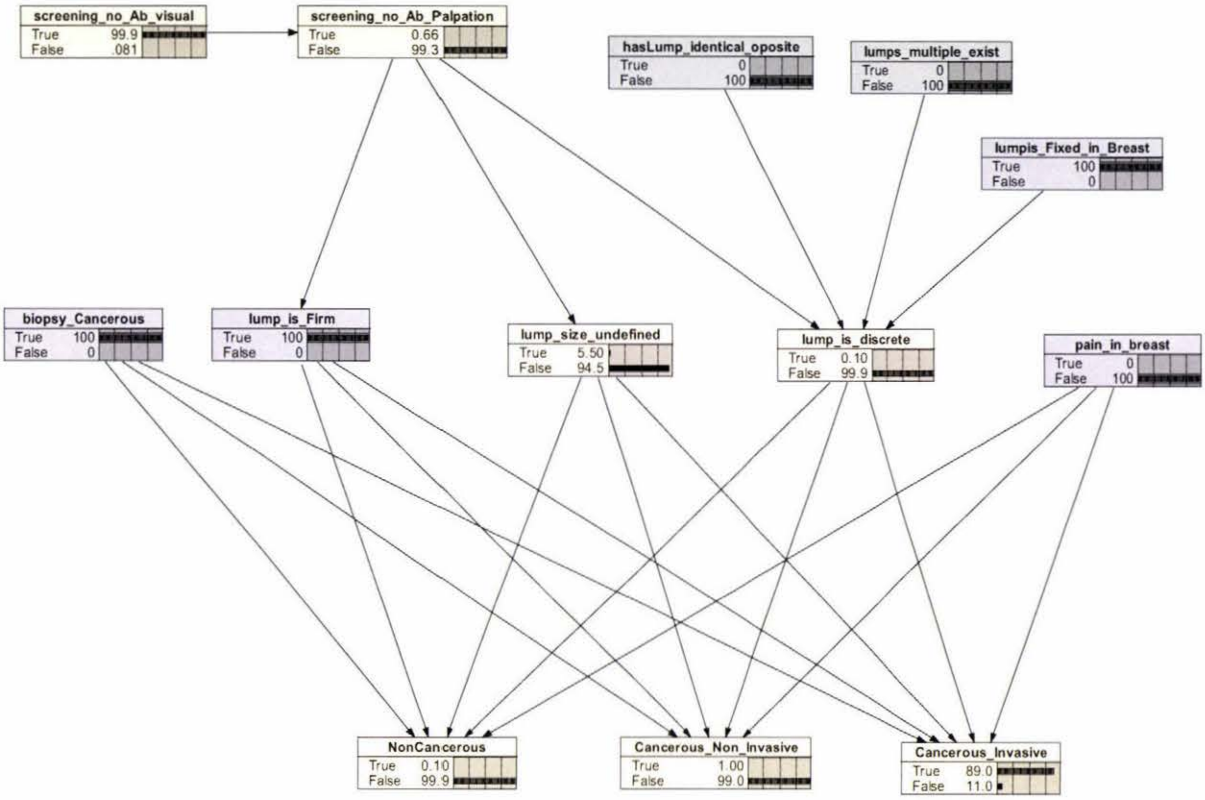


Figure 4-2: Bayesian Network, Abnormalities Detected

◊ Evaluation may then proceed to KB system sets associated with the recommended subclass. For example figure 4-3 subclass invasive cancer, the indication is a high probability of Stage 4 Invasive Cancer.

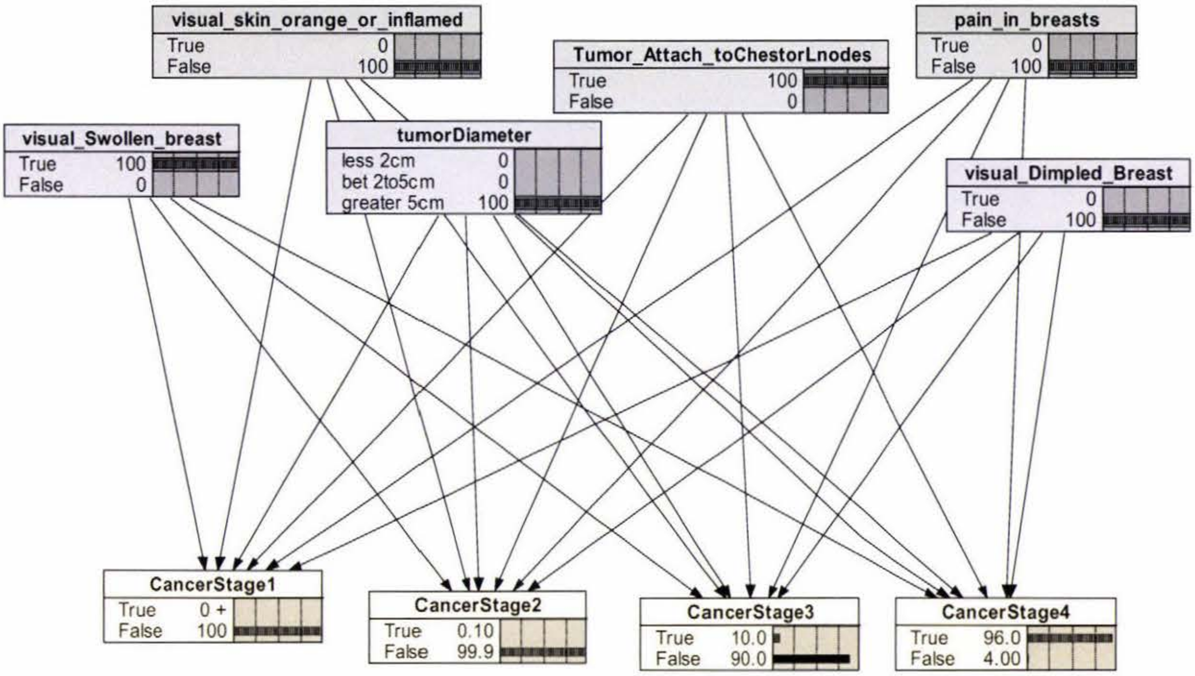


Figure 4-3: Bayesian Network, Invasive Cancer

- If certainty of classification is not acceptable the newly obtained findings may be re-queried in the ontology, with findings that are not applicable to the ontology being stored for use in the next applicable set of KB systems.

4.5 Advantages

By finding a specific domain granule we are able to limit our further decision making to a set of input variables with a decision value that are normally more meaningful than considering the whole domain. That is by processing available properties in an ontological form we can specify a reference class/KB system with a limited set of variables that have potentially enhanced values and thus each new (directed) finding will have a greater importance to the decision making process than if attribute selection was undirected.

Let's consider a simple example of a diagnostic of a patient with scaly skin located on one breast:

1) Universal Domain (no initial classification based on findings)

- Large collection of possible groups of disorder that include skin disorders and breast tumours.
- What symptoms do we select from the universe of possible symptoms?
- Key determining properties may be difficult to initially define.
- How do we build decision rules or have an idea of likelihood/probability when we have an infinite set of possible symptoms?
- Difficult for a system to know when the system conclusions are wrong or possibly not correct.

2) Class Reference = Super-Class or Union of Classes Skin Disorder – Breast Tumours (initial classification based on findings)

   ◊ Smaller finite set of possible disorders.

   ◊ Possible symptoms normally limited and manageable.

   ◊ Key symptoms could be more easily defined e.g. are there breast lumps / are there skin disorders in other areas of the body.

   ◊ Likelihood / probability / functional rules are applicable and normally significant.

   ◊ Easier to know when the class reference is inappropriate and dynamically make a new domain hypothesis.

Further it is important to note that we do not propose that the KB system associated with an ontologically defined domain or class is limited to the properties defined in the ontology. For example let us consider the characteristic patients age:

Ontology: Patients age for many disorders does not necessarily exclude them from being affected from a disorder. Thus age may not be a useful classifier.

KB System: Patients age may be very useful in considering the likelihood that the patient suffers several specific possible disorders.

Thus the characteristic age is clearly important for our KB system but in many cases inapplicable for classification in our ontology. Another good example is ethnic group where people of African origins raised in Africa may have a lower probability of suffering breast cancer than a European with a similar age.

The primary proposed use of the ontology is to define the domain of consideration or the reference class through a recursive querying process and to direct the system to the most probable features of consideration. Thus even if the ontology is re-queried based on the additional findings obtained from operation with the first KB system the findings not used in the ontology are not relevant and could be stored for reuse in the next domain's KB system.

As secondary benefits:

◊ The mapping ontology would be inherently stable while being extensible to new knowledge.

◊ A property enriched ontology could be beneficial in linking domain specific KB systems whatever the KB system functionality.

The primary proposed use of the ontology is to define the domain of consideration or the reference class and to direct the system to the most probable features of consideration. Thus even if the ontology is re-queried based on the additional findings obtained from operation with the first KB system the findings not used in the ontology are not relevant and could be stored for reuse in the next domain's KB system.

Considering our ontologic Map once we have reached a degree of extraction that the user or system defines as the lowest level that can be obtained we then have a domain in which medical KB systems such as Eon, Belief Revision, and pattern recognition could be potentially applied.

4.6 Ontological Domain Mapping With Belief Change Methods

Belief change systems are addressing change in belief with new information coming to light. Thus there is clearly a relationship between the propositions of ontology based domain determination. We see the domain determination as a more global change of the class reference base where an event relates to a known or defined classification. Belief change methodologies are hindered by complexity and computation challenges however they could be useful in reference class/domain specific belief change and thus altering the rules of the applicable KB system set. A global ontology of defining structure of domain connections is potentially challenging to create but once created updating is likely to be less complicated as classification is relatively static where as reaction to new information is dynamic. Thus we could consider further the use of belief change theory in our methodology as a means to improve the system.

5 Prototype System

5.1 *System Overview*

For the sack of this case study we will refer to the development as a Symptoms Ontology for Mapping Knowledge Systems (SOMKS).

5.1.1 Functionality

SOMKS is conceived to carry out the following tasks:

- Query the ontology to define a domain that is defined by a necessary and sufficient condition.
- Query the domain to define the lowest level supper class that can be extracted by a recursive query based on the initially specified properties (symptoms).
- Define domain specific knowledge bases that define and use the symptoms that need to be obtained / extracted to form a recommendation within the system.
- Forming a recommendation for the next class reference level based on the findings defined.

SOMKS this version does not carry out the following functions:

- Dynamically reapplying applicable obtained findings to the ontology and querying to obtain a new class reference domain.
- Definition of associated properties when a Query of the ontology results with a root class.

SOMKS conducts queries using Protégé OWL API and Norsys JNetica308 API

5.1.2 System Architecture

The SOMKS process consists of the following steps:

- Loading the defined domain ontology.
- Querying whether a necessary or sufficient condition exists.
- Querying classes defined by the properties specified and removing subclasses.
- Where more than one class is obtained in the initial property query, querying recursively to obtain the lowest super class level where only one class is defined.
- Obtaining an associated domain model mapped as a Bayesian network and request input information important for decision making within the defined reference class.

◊ Compiling the network with inputted findings and receiving probability outputs for the associated class reference.

◊ The user is then directed to the Bayesian model associate with the most probable class reference level.

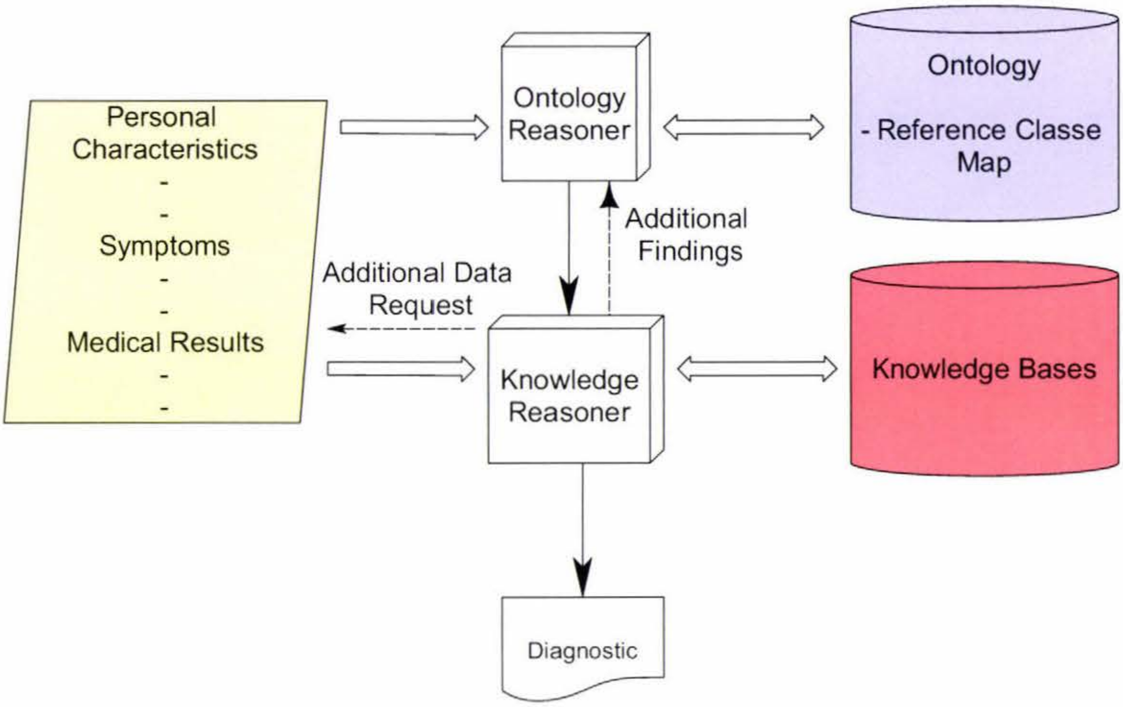


Figure 5-1: SOMKS Architecture

Considering that there are numerous technologies and systems for extracting patterns we considered that we are able to extract relevant features from the sources that are correct and produce data in a semi structured format that is processable. We do not address the requirement of storage/access of patient data because there are many commercial available system solutions that are likely to be location / institution / utilisation specific.

5.2 Ontology

The most appropriate example would have been the use of an existing ontology that has distinguishing properties specified that can be used to define classes. The targeted ontology needed to cover a number of potential decision domains and not be limited in scope, specify a number of detectable characteristics that can be used for classification, and contain conflicting classifications for individual attributes e.g. in Fat necrosis is normally soft but may sometimes be hard.

Unfortunately there are currently few existing ontologies that have already included physical properties particularly in the area of bioinformatics some are very vast but stick to categorisation i.e. SuperClassOf, SubClassOf e.g. http://obo.sourceforge.net/browse.html. Accordingly an example ontology was developed for diagnosis of breast cancer referencing the Unified Medical Language System http://www.nlm.nih.gov/research/umls/. The ontology is not proposed to be an expert representative of the real world domain that breast cancer diagnosis takes place, nor does it consider temporal decision process often used in the medical practice expert systems i.e. Eon, Athena or Oncocin. However the ontology does provide a legitimate example of how a patient might be classified and considers that there are certain stages in a diagnosis process. The ontology super-class map can be referenced in Appendix 1.

Our symptoms ontology for Breath Health takes the applicable attributes (attributes that can include or exclude classes) defined in our KB systems. As we move down the ontology the specification of the symptoms become more specific and can define a specific abnormality.

5.3 Implementation Considerations

5.3.1 Programming language Java

Java was considered the appropriate language for this development because it is an object oriented programming language that is platform independent, reliable and robust. Java Object Oriented framework means that SOMKS can easily be extended with new classes to manipulated extended knowledge bases (new domains. Other considerations were that the Java SDK 1.5 could be used in conjunction with APIs from both Netica and Protégé packages that made this prototype development feasible.

5.3.2 Ontology Construction Protégé

Protégé is an open-source platform software for building domain models and knowledge based applications with ontologies. Protégé was developed by the Stanford Medical Informatics department at the Stanford University School of Medicine. Protégé is implemented in Java, and runs on a broad range of hardware platforms, including Windows, Mac OS, Linux, and Unix.

Protégé has been a primary development environment for a number of ontologies in the life sciences. These projects include the Foundational Model of Anatomy, Cerner's Clinical Bioinformatics ontology, the DICE TS, and MGED ontology.

The basic Protégé infrastructure provides the following features

- ◊ An extensible knowledge model
- ◊ Protégé's representational primitives provide classes, instances of these classes, properties representing attributes of classes and instances.

- Ability to import ontologies in different formats. Including OBO, DAG-EDIT, XML, RDF, and OWL.
- Ontology authoring and management tools.
- An extensible architecture that enables integration with other applications.
- A Java Application Programming Interface (API). We can use the Protégé API to access and programmatically manipulate Protégé ontologies.
- Protégé can be extended by way of a plug-in architecture and a Java-based Application Programming Interface (API) for building tools and applications.

The Protégé platform supports two means of modelling ontologies:

The Protégé-Frames editor enables users to build and populate ontologies that are frame-based, in accordance with the Open Knowledge Base Connectivity protocol (OKBC).

The Protégé-OWL editor enables users to build ontologies for the semantic web, using W3C's Web Ontology Language (OWL).

It was decided to develop this ontology in OWL in order to consider the potential use of the development over the web however Protégé frame could have equally been used.

5.3.3 Netica Belief Networks & Netica J API

For belief network design it was decided to use Netica software from Norsys Software Corporation considering that Netica is considered to be the world's most widely used Bayesian network development software and has been extensively used in medicine and biology. In addition the Netica Java API for the development of the belief network interface is well developed and documented.

SOMKS uses the latest pre-release versions of Netica, Netica 308 for designing and building belief networks and Netica J306 API. Netica is a program for building belief networks and influence diagrams. Netica can use the networks to perform various kinds of inference. Given a new case of which we have limited knowledge, Netica will find the appropriate values or probabilities for all the unknown variables. Netica can use influence diagrams to find optimal decisions which maximize the expected values of specified variables. Netica can construct conditional plans, since decisions in the future can depend on observations yet to be made, and the timings and inter-relationships between decisions are considered, however in SOMKS we have not yet consider the actual decision process of conditions.

6 Case Project Breast Cancer Diagnosis

In this section we review breast health and breast cancer diagnosis process to develop a tangible example of the use of SOMKS as an extensible model of Breast Health analysis. This case study illustrates how SOMKS works to define a reference class of decision and then applies an associated Knowledge base to the defined class reference.

Cancer is a group of diseases that occur when cells become abnormal and divide without control or order. Benign tumours are not cancerous. The cells in benign tumours do not invade other tissues and do not spread to other parts of the body. Malignant tumours are cancer. The cancer cells grow and divide out of control. They can invade and damage nearby tissues and organs. Also, cancer cells can break away from a malignant tumour and enter the bloodstream or lymphatic system. Breast cancer is the most common type of cancer among women, breast cancer also affects men.

The most common type of breast cancer begins in the lining of the ducts and is called *ductal carcinoma.* Another type, called *lobular carcinoma*, arises in the lobules. When breast cancer spreads outside the breast, cancer cells are often found in the lymph nodes under the arm (auxiliary lymph nodes). If the cancer has reached these nodes, it may mean that cancer cells have spread to other parts of the body, other lymph nodes and other organs, such as the bones, liver, or lungs.

6.1 Overview Diagnostic Process

The examination for breast cancer normally includes inspection (looking) and palpation (feeling) of the entire breast/chest area including the lymph node areas above and below the collarbone and under each arm. Women 40 and older are encouraged having breast examination annually and after the age of 50 the examination should include a mammogram (a mammogram is a kind of x-ray, using very low levels of radiation).

A small percentage of cancers will not be detected by a mammogram some of these cancers can be detected by palpation.

Warning signs of breast cancer are as follows [http://imaginis.com]:

◊ Any new lump found in the breast or armpit

◊ Any lump or thickening that does not shrink or lessen.

◊ Any change in the size, shape or symmetry of your breast

◊ A thickening or swelling of the breast

◊ Any dimpling, puckering or indention in the breast

◊ Dimpling, skin irritation or other change in the breast skin or nipple

- ◊ Redness or scales of the nipple or breast skin
- ◊ Nipple discharge (fluid coming from nipples other than breast milk), particularly if the discharge is bloody, clear and sticky, dark or occurs without squeezing the nipple
- ◊ Nipple tenderness or pain

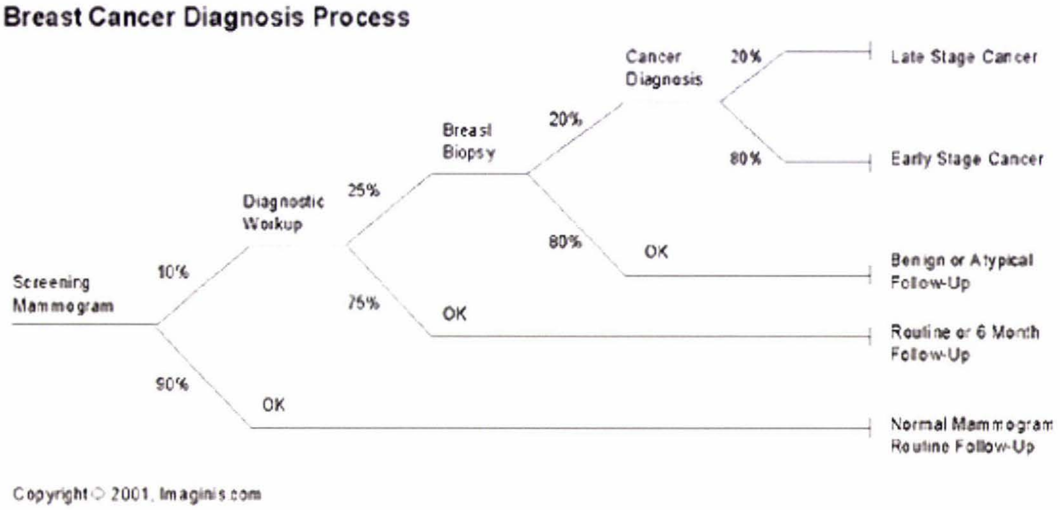

Breast Cancer Diagnosis Process
Screening Mammogram
10%
90%
OK
Diagnostic Workup
25%
75%
OK
Breast Biopsy
20%
80%
OK
Cancer Diagnosis
20%
80%
Late Stage Cancer
Early Stage Cancer
Benign or Atypical Follow-Up
Routine or 6 Month Follow-Up
Normal Mammogram Routine Follow-Up
Copyright © 2001, Imaginis.com

Figure 6-1: Breast Cancer Diagnosis Process, Percentage of Patients [www.imaginis.com]

A biopsy is the process of removing a sample of breast tissue to define whether it is cancerous or non-cancerous, the type of tumour, and if cancerous the degree of cancer. The types of Biopsy are, Fine Needle Aspiration (FNA), Core Needle Biopsy, Vacuum-Assisted Biopsy, Large Core Biopsy and Open Surgical Biopsy.

6.2 *Walking Through SOMKS*

Step 1: Run SOMKS

The user selects the properties (Symptoms) and the Instance (Likelihood) where applicable that relate to the patient. Here we have only used two alternative properties because of the complexity in programming with Protégé OWL API and it was considered that more properties does not necessary improve the demonstration of this example prototype.

Step 2: Go

Once the user has selected the properties and instances, he should then click 'Go'. After a short moments delay (due to ontology extraction) for the first query, SOMKS will return possible abnormalities if an abnormality can be assumed from the symptoms and a proposed reference class. In the event the user is not content with the domain specified the user may input alternative symptoms and rerun a query.

Step 3: Run Class

In the Menu Bar select the menu item 'RUN CLASS' this will take the user to the GUI interface associated with the domain obtained in the ontology query. The user can then specify the findings required for decision making in this domain, i.e. for choosing the possible direct subclasses of the obtained ontology class level.

Step 4: Go Knowledge Base

Once the user has selected either Unknown (default), True, or False for each finding the user should then select the button 'Go'.

The GUI will provide the following output

- ◊ The base probability in the domain of each outcome.
- ◊ The probability of each outcome given the inputted findings.

The user can rerun the query based on new findings; however there is a bug in the Netica API that leads to system exit after 3 runs or 6 network compiles.

Step 5: Go to Next Layer

Once the user has obtained a level of probability that is considered appropriate the user may chose to go to the next decision level or class reference layer (where a further decision level exits).

In the Menu Bar select the menu item 'Go to Next Layer' this will take the user to the GUI interface associate with the domain having the highest probability of occurring based on the findings specified in the first domain.

The user may repeat this operation multiple times; however there is a bug in the Netica API that leads to system exit after 3 runs or 6 network compiles. In the event this occurs the user will need to re-run SOMKS.

7 Conclusion and Future Work

7.1 Conclusion

The methodologies of KB systems and ontology based decision systems have a well developed and large portfolio of means for managing domain specific problems. Unfortunately most of these systems seem to either imply excess computation complexity or do not address complexity challenges of application in an open domain.

In this thesis report we have presented a methodology that could be applied to address complexity of open domain decision making situations by increasing the significance of attributes and directing finding extraction. In addition our system could potentially manage implications of general knowledge by second guessing classification as new findings are obtained. This approach would be used to build open KB systems and specifically we have demonstrated that this system could be used for diagnosis processes to better enable Physicians to conduct rapid diagnosis and limit unnecessary patient testing.

The thesis work does not fully address the proof of these concepts considering that the system proposed is design dependant. In order to fully justify these proposals the decisional structure must be designed by medical experts and must be fully tested with real medical data/environment.

7.2 Future Work

In this work we have proposed the enhancement of decision and knowledge based systems through controlling complexity and by providing a proof of concept development. This solution could lead to some exiting new applications in navigation, medial/bio knowledge Systems, human sensory aids, etc.

In order to further justify our proposals a case study diagnostic system should be designed with medical experts, using structured medical data and tested in field situations. In addition further research is required in the areas including the following:

◊ Property Inference From Findings

A method for targeting properties to be evaluated where the initial findings do not allow the system to classify below the route class. This research considers finding a lowest level of Class Reference in a human defined extensible ontological hierarchy. However the system does not yet consider a calculation or method of property inference that is required to define most likely properties when a route class or inappropriately high reference Class is defined and a new query would most logically be applied on the ontology i.e. minimising complexity in obtaining unknown properties in a global domain.

◊ Domain experts and learning systems

The case model SOMK is based on defined structure and plug values from medical references. The structure of the ontology and class reference Knowledge Bases need to be developed in conjunction with domain experts in this case Pathologists and Oncologists. The statistic needed for defining probabilities or decision rules used in the class reference Knowledge Bases should be learnt from recorded medical data.

◊ Managing Class Disjunction

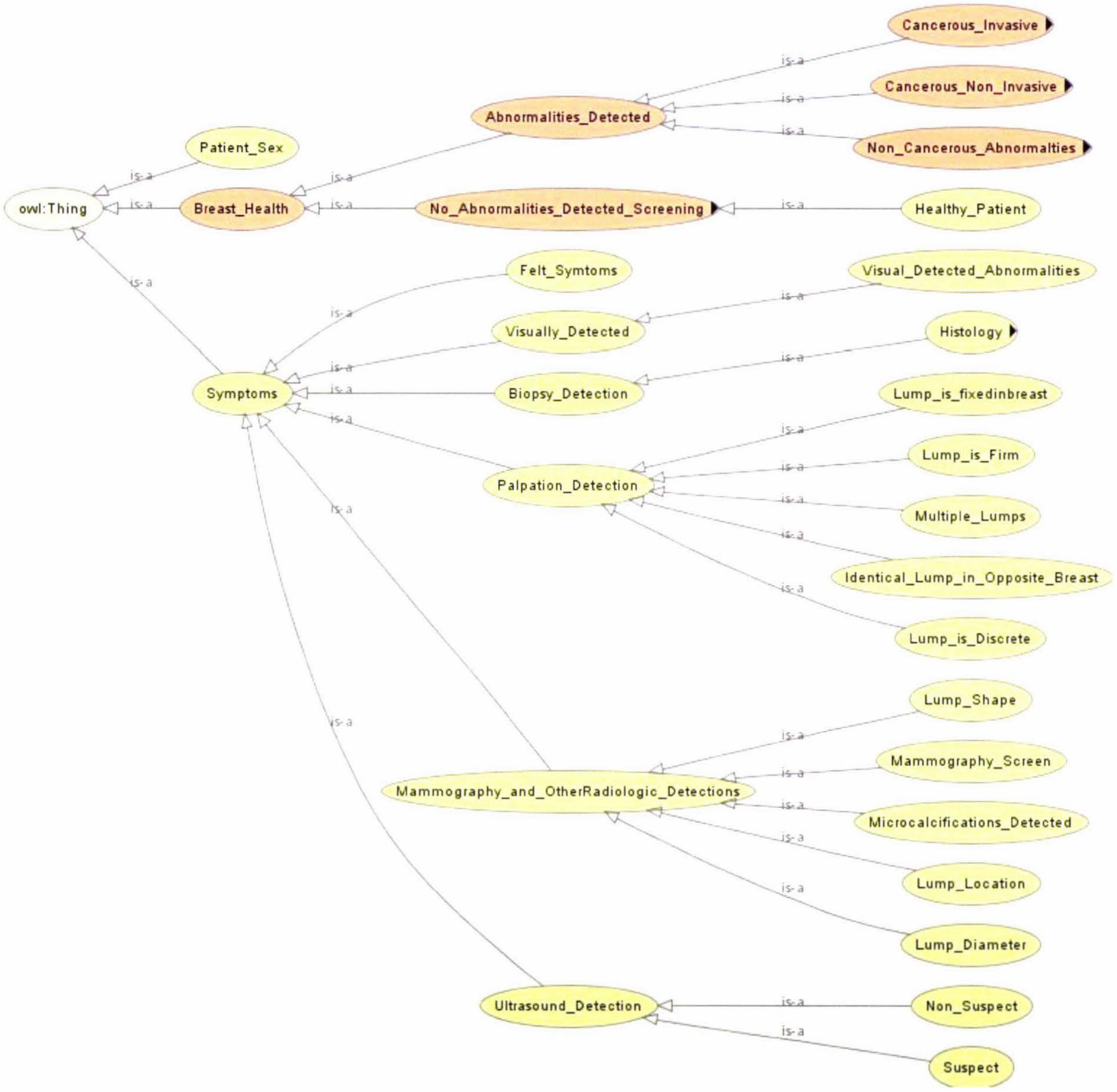
This model assumes that the outcome will be in the domain defined. However that is not necessarily the case as there is potentially of an overlap between different reference classes. In our example we might define that a patient has no abnormalities but in reality he could have a non-cancerous tumour that was simply not detected. This could probably be implied by key findings; however, we need to have a method of error control and of jumping from one class reference to another in the most logical sense.

◊ Interoperation between ontology and expert system

This system uses Bayesian networks that appear to offer many advantages. Alternative knowledge base and decision systems should be evaluated with this type of approach.

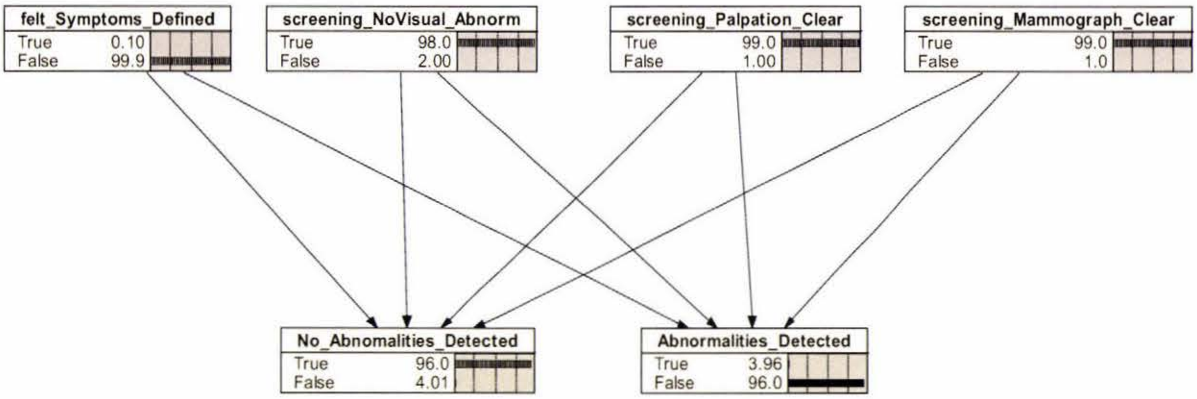
Appendix 1: SOMKS Base Ontology

The diagram of super-classes is presented below, the complete OWL file is provided with this document in electronic form.

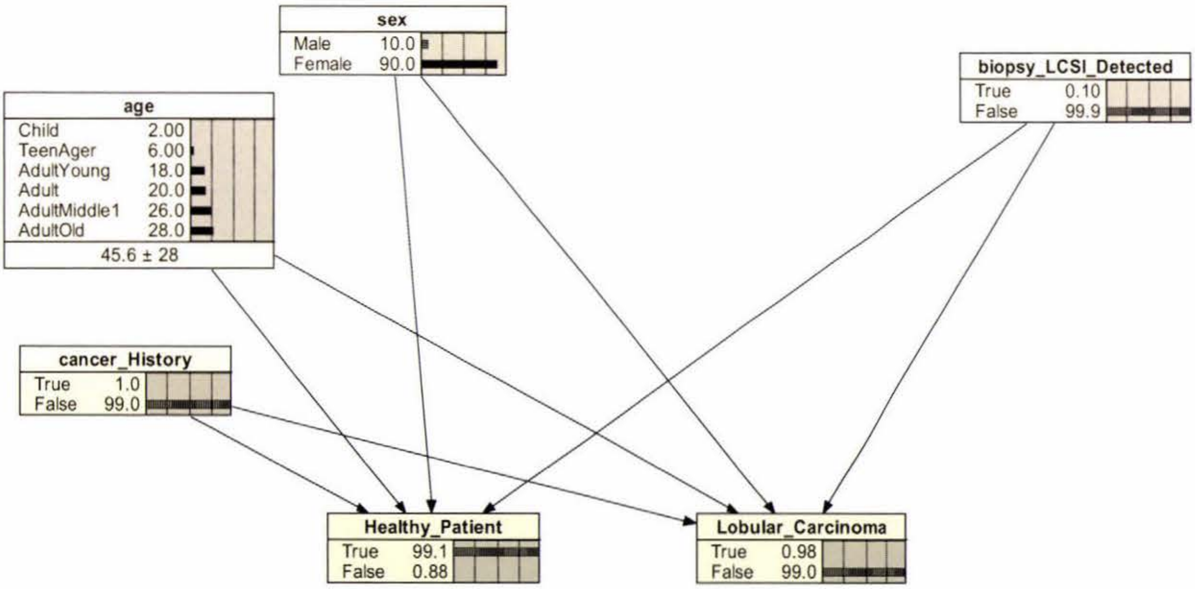


Appendix 2: SOMKS Bayesian Networks

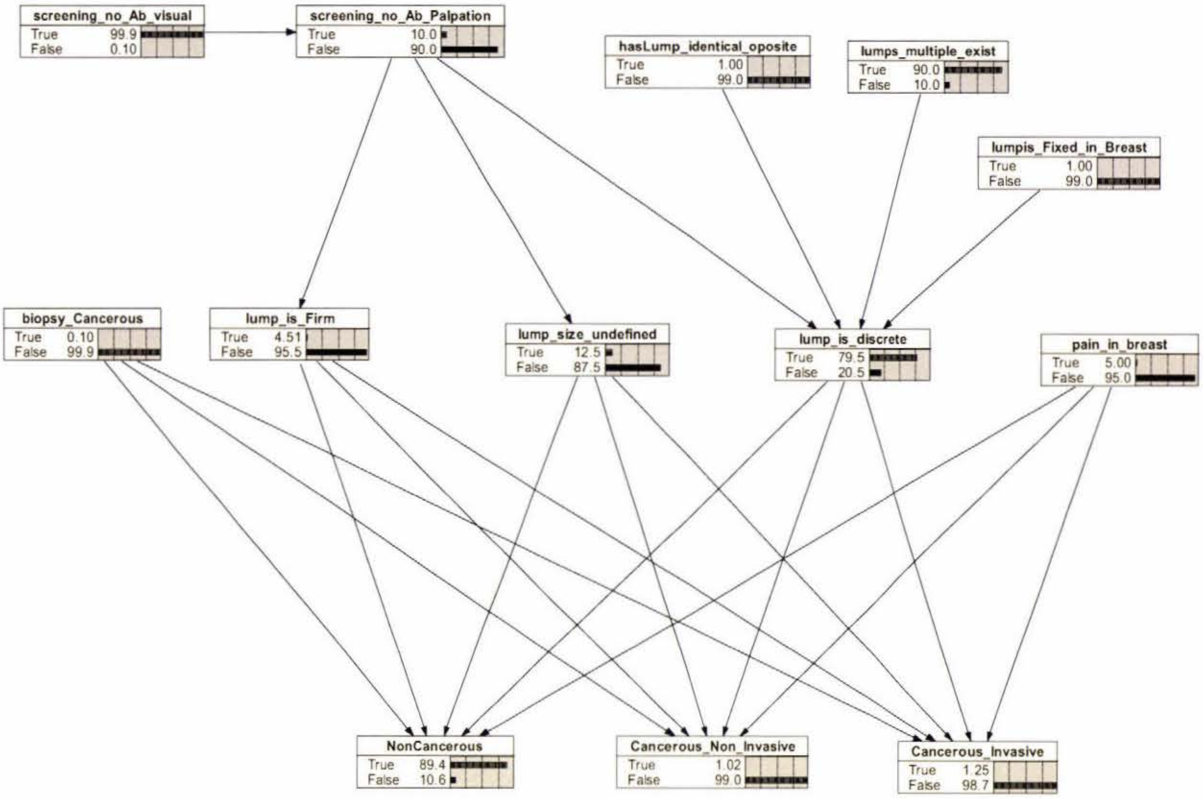
Breast Health



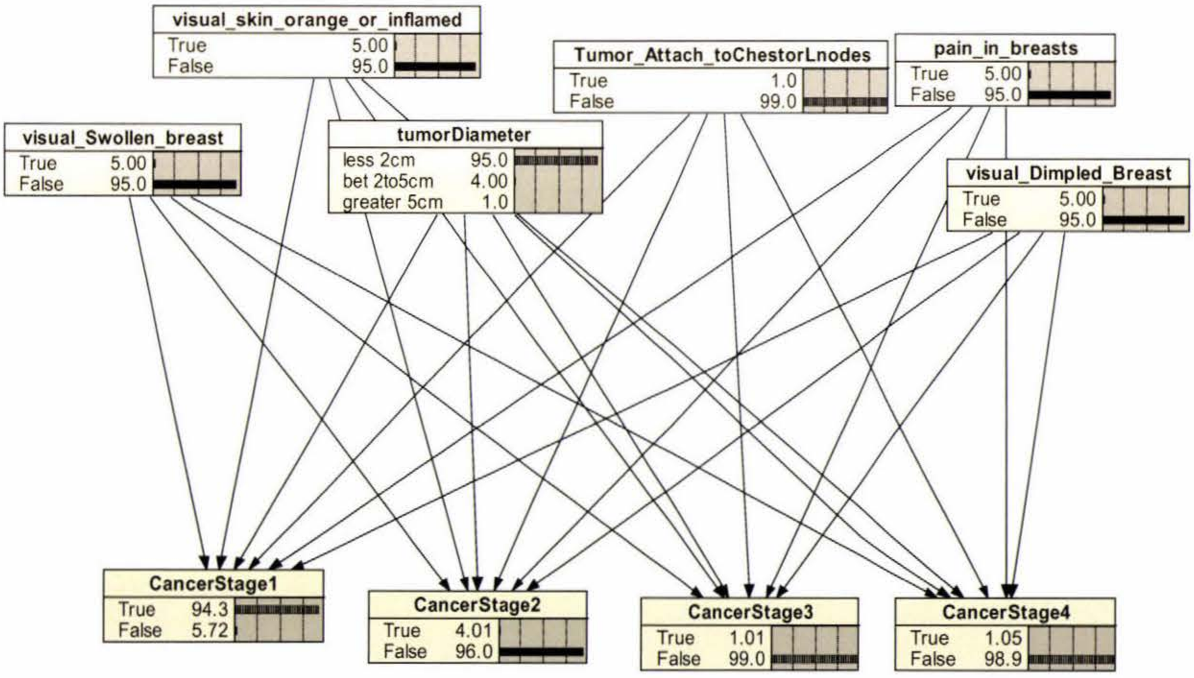
No Abnormalities Detected



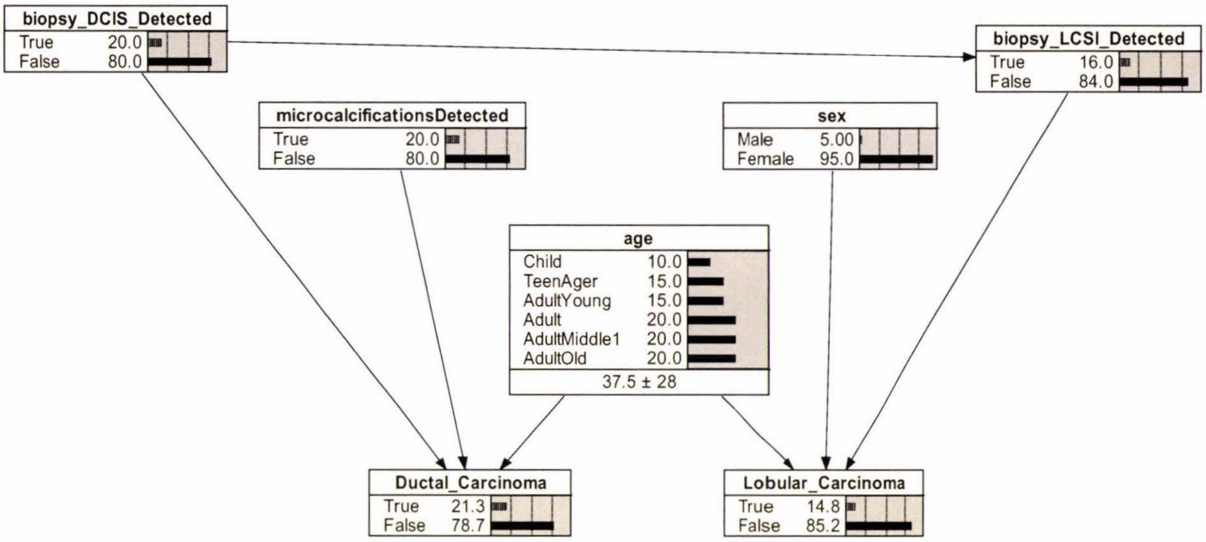
Abnormalities Detected



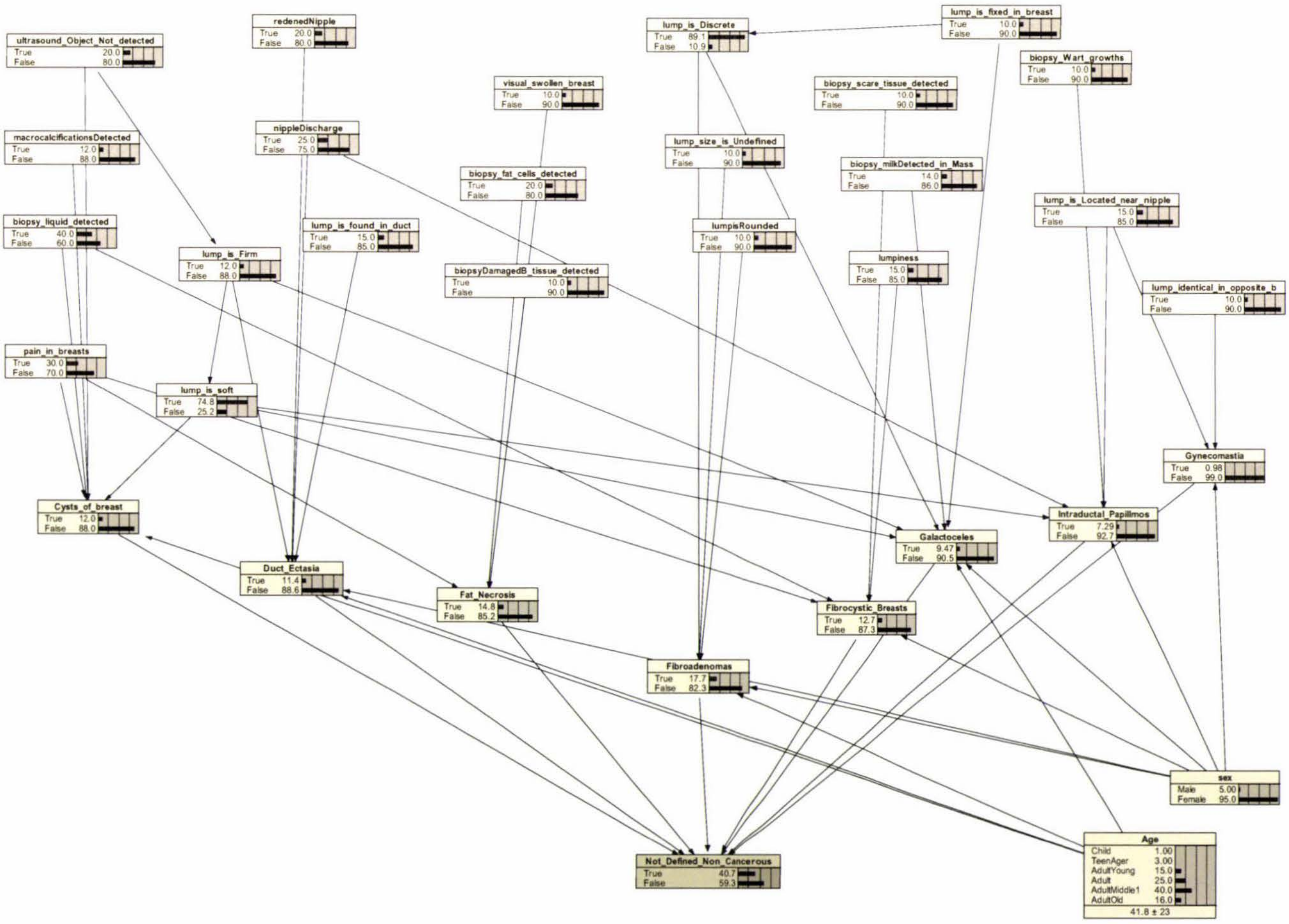
Cancerous Invasive



Cancerous Non-Invasive



Non-Cancerous Abnormalities



Bibliography

1. Alchourron, C. E., Gardenfors, P. & Makinson, D. (06.1985). “On The Logic of Theory Change: Partial Meet Contraction and Revision Functions”, The Journal of Symbolic Logic, Vol 50, No. 2, Pages 510 - 530.

2. Altman, Russ D. (09/10.2000). “The Interactions Between Clinical Informatics and Bioinformatics”, Journal of the American Medical Informatics Association, Volume 7, Number 5.

3. Barnett G. O., Hoffer E.P., Packer M. S. (1992) “DXplain-demonstration and discussion of a diagnostic decision support system”,. Proceedings Annual Symposium on Computer Applications in Medical Care, pp 822.

4. Beckett, D. (10.02.2004). “RDF/XML Syntax Specification”, W3C Recommendation.

5. Bidoit, N., & Hull, R. (1989). “Minimalism, justification and non-monotonicity in deductive databases”, Journal of Computer and System Sciences, vol 38, pp. 290-325.

6. Boudreau, T., Glick, J., Greene, S., Spurlin, V., & Woehr, J. (2003) “NetBeans The Definitive Guide”, O'Reilly & Associates, USA.

7. Brickley, D. & Guha, R. (10.02.2004). “RDF Vocabulary Description Language 1.0: RDF Schema”, W3C Recommendation. (3)

8. Chou, T. S-C. & Winslett, M. (06.1991). “The Implementation of a Model-based Belief Revision System”, ACM SIGART Bulletin (0163-5719), Volume 2, Issue 3, Pages 28 -34.

9. Coiera, E. (2003). “The Guide to Health Informatics (2nd Edition)”. Arnold, London, October 2003.

10. Colyvan, M., Regan, H. M. & Ferson, S. (2001). “Is it a Crime to Belong to a Reference Class?”. The Journal of Political Philosophy; Volume 9, Number 2, pages 168-181.

11. Cooper, G, F., (1990), “The computational complexity of probabilistic inference using Bayesian belief networks”, Artificial Intelligence, vol. 42, pp. 393-405.

12. Cox, E. (1999) “The Fuzzy Systems Handbook, Second Edition”, Academic Press, USA.

13. Dalal, M. (1988), “Investigations into a theory of knowledge base revision”, Preliminary report. In Proceedings of the Seventh National Conference on Artificial Intelligence, pps 475-479.

14. Duda , R. O., Hart, P. E., Stork, D. G. (2000). “Pattern Classification (2nd Edition)”, Wiley, USA.

15. Eiter, T., & Gottlob, G. (1992), “On the complexity of propositional knowledge base revision, updates and counterfactuals”. AIJ, 57, pages 227-270.

16. Fagin, R., Kuper, G. M., Ulman, J. D., & Vardl, M. Y. (1986) ""Updating Logical Databases," Advances in Computing Research,

17. Fagin, R., Ulman, J. D., & Vardl, M. Y. (1983) "On the Semantics of Updates m Databases", Proceedings of the 2nd ACM Symposium on Principals of Database Systems, pps 352-365.

18. Fuhrmann, A. (1991), "Theory Contraction Through Base Contraction", Journal of Philosophical Logic, vol 20, pp.175-203.

19. Goldstein, M. K., Hoffman, B. B., Coleman, R. W., Musen, M, A., Tu, S. W., Advani, A., Shankar, R., O'Connor, M. (2000). "Implementing Clinical Practice Guidelines While Taking Account of Change Evidence: ATHENA DSS, an Easily Modifiable Decision Support System for Managing Hypertension in Primary Care", Proc AMIA Symp: pages 300-304.

20. Hajek, Alan. (2005) "The Reference Class Problem is Your Problem Too", Synthese.

21. Halpern, J. Y. (10.1999). "Belief Revision: A Critique". Journal of Logic, Language and Information, Volume 8, Number 4, pages: 401 – 420.

22. Halpern, J. Y. (2003). "Reasoning About Uncertainty". MIT Press, USA.

23. Hobbs, J.R.(1985) "Granularity", Proceedings of the 9th International ,Joint Conference on Artificial Intelligence, pages 432-435.

24. Horridge, M., Knublauch, H. & Stevens, R. (27.08.2004) "A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools", The University Of Manchester.

25. Jensen, F, V. (1996), "An introduction to Bayesian Networks", UCL Press, London.

26. Kahn, M. G., Lawrence, F. M. & Shortliffe, E. H. (1986). "Time in Clinical Decision Support Systems: Temporal Reasoning in ONCOCIN and ONYX", ACM SIGBIO News Letter, Volume 8, Issue 1, Pages 13 - 16.

27. Klein, M. R. & Methlie, L. B. (1995). "Knowledge-based Decision Support Systems 2nd Edition", Wiley.

28. Liberatore, P. & Schaerf, M. (1996). "The Complexity of Model Checking for Belief Revision and Update", Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'96), Pages 556 – 561, AAAI Press/The MIT Press.

29. Loy, M., Eckstein, R., Wood, D., Elliot, J., & Cole, B. (2003) "Java Swing, Second edition", O'Reilly & Associates, USA.

30. Manola, F. & Miller, E. (10.02.2004). "RDF Primer", W3C Recommandation.

31. Mork, P., Brinkley, J. F., & Rosse, C. (2003). "Querying Agent for the Foundational Model of prototype for providing flexible and efficient access to large semantic networks". Journal of Biomedical Informatics, 36, pages 501–517.

32. Musen, M. A. (1999), "Stanford Medical Informatics: uncommon research, common goals". Medical Computing. Jan-Feb;16(1), vol 50, pp.47-8.

33. Musen, M. A., Tu, S. W., Das, A. K., and Shahar, Y. (1996). "EON: A Component-Based Architecture for Automation of Protocol-Directed Therapy". Report KSL-96-06, October, Journal of American Medical Informatics Association.

34. Papadimitrou, C, H. (1994), "Computational Complexity", Addison-Wesley, USA

35. Pawlak, Z. (1998) "Granularity of knowledge, indiscernibility and rough sets", Proceedings of 1998 IEEE International Conference on Fuzzy Systems, pages 106-110.

36. Peleg, M., Tu, S., Bury, J., Ciccarese, P., Fox, J., Greenes, R. A., Hall, R., Johnson P. D., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe E. H., Stefanelli, M. (01/02.2003). "Comparing computer-interpretable guideline models: a case-study approach". Journal of American Medical Informatics Association, pages 52-68.

37. Powers, S. (2003). "Practical RDF". O'Reilly Media.

38. Satoh, K. (1988), "Nonmonotonic reasoning by minimal belief revision.", In Proceedings of the International Conference on Fifth Generation Computer Systems, pps 455-462.

39. Schalkoff, R. (1992). "Pattern Recognition, Statistical, Structural and Neural Approaches", Wiley, USA.

40. Schildt, H., (2001) "Java2: The Complete Reference, Fourth Edition", McGraw-Hill, USA.

41. Smith, M. K., Welty, C. & McGuinness, D. L. (10.02.2004) "OWL Ontology Language Guide", W3C, 10.02.2004.

42. Sonka, M., Hlavac, V. & Boyle, R. (1999). "Image Processing, Analysis, and Machine Vision". Second Edition, PWS Publishing.

43. Tu, S. W., Kemper, C. A., Lane, N. W., Carson, R. W. & Musen, M. A. (1993) "A methodology for determining patients' eligibility for clinical trial". Methods of Information in Medicine, 32, pages 317-325.

44. Weaver, J., Mukhar, K., & Crume, J. (2004) "J2EE 1.4", Apress L.P., Berkeley, USA

45. Winslett, M. (1988) "A Framework for Comparison of Update Semantics (Extended Abstract)", proceedings of the seventh ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, Pages: 315 – 324, ACM.

46. Winslett, M. (1990), "Updating Logical Databases", Cambridge University Press, USA.

47. Yao, Y.Y. (2004) "A Partition Model of Granular Computing", Lecture Notes in Computer Science. 1, pages 232-253.

48. Zhang, L. and Zhang, B. (2004) "The quotient space theory of problem solving", Fundamenta Informatcae, 59, pages 287-298.

Internet References

1. http://imaginis.com/
2. http://java.sun.com/
3. http://www.nlm.nih.gov/research/umls/
4. http://www.openclinical.org
5. http://protege.stanford.edu/
6. http://smi-web.stanford.edu/
7. http://www.wikipedia.org/