Ontology – Supported Machine Learning and Decision Support in Biomedicine

Alexey Tsymbal[1], Sonja Zillner[2], and Martin Huber[1]

[1] Corporate Technology Div., Siemens AG, Erlangen, Germany
{Alexey.Tsymbal, Martin.Huber}@siemens.com
[2] Corporate Technology Div., Siemens AG, Munich, Germany
Sonja.Zillner@siemens.com

Abstract. Nowadays, ontologies and machine learning constitute two major technologies for domain-specific knowledge extraction which are actively used in knowledge-based systems of different kind including expert systems, decision support systems, knowledge discovery systems, etc. While the aim of these two technologies is the same – the extraction of useful knowledge – little is known about how the two sources of knowledge can be successfully integrated. Today the two technologies are used mainly separate; even though the knowledge extracted by the two is complementary and significant benefits can be obtained if the technologies were integrated. This problem is especially important for biomedicine where relevant data are often naturally complex having large dimensionality and including heterogeneous features, and where a large body of knowledge is available in the form of ontologies. In this paper we propose one approach for improving the performance of machine learning algorithms by integrating the knowledge provided by ontologies. The basic idea is to redefine the concept of similarity for complex heterogeneous data by incorporating available ontological knowledge, creating a bridge between the two technologies. Potential benefits and difficulties of this integration are discussed, two techniques for empirical evaluation and fine-tuning of feature ontologies are described, and an example from the field of paediatric cardiology is given

1 Introduction

Ontologies and machine learning constitute two major technologies for domain-specific knowledge extraction actively used in knowledge-based systems of different kind including expert systems, decision support systems, knowledge discovery systems, etc. By establishing an explicit formal specification of the concepts in a particular domain and relations among them, *ontologies* provide the basis for reusing and integrating valuable domain knowledge within applications [13]. *Machine learning* algorithms are also actively applied in order to extract useful knowledge in different problem domains by searching for interesting patterns (dependencies) in large volumes of data [21].

The principal difference between the two technologies is that the first is usually expert-driven (ontologies are a result of the knowledge elicitation process from a

S. Cohen-Boulakia and V. Tannen (Eds.): DILS 2007, LNBI 4544, pp. 156–171, 2007.

domain expert by knowledge engineers, and data is not necessarily involved in this process); while the latter is data-driven (the search for patterns is usually automatic and does not involve substantial interaction with the expert). While the aim of these two technologies is the same – the extraction of useful knowledge – little is known about how the two sources of knowledge can be successfully integrated.

Traditional machine learning algorithms are not able to incorporate background domain knowledge, but instead work with a sequence of instances, where each instance is represented by a single feature (attribute) vector describing the instance [21]. This limitation of traditional machine learning techniques is widely acknowledged today. The issue of learning from more complex data, and in particular similarity for complex heterogeneous data with rich background knowledge was, for example, in focus at the recent International Workshop on Mining Complex Data, MCD 2006 [31].

The principle of instance similarity is fundamental to the vast majority of machine learning algorithms. The main assumption in supervised, unsupervised and semi-supervised machine learning algorithms is that the instances of the same class (cluster) are more similar to each other than the instances of different classes (clusters). In this paper, we propose an approach to improving the performance of machine learning by redefining the concept of similarity with incorporating constraints provided by ontological domain knowledge, that is instead of simply providing the machine learning algorithm with unrelated features in the form of a single vector or a vector set, we will semantically enhance them by integrating the graph structures of relevant domain ontologies.

We see two main benefits that can be obtained from this procedure: first and the most important, is that the performance of machine learning algorithms will be improved by incorporating knowledge provided by domain ontologies. For example, the predictive accuracy of *k*-nearest neighbour classification can be improved. Second, a more practical and application-oriented advantage, is that an ontology, describing the interrelations between the features in a machine learning problem, can be presented to the user of a knowledge-based system via a Graphical User Interface, and provide an effective means of feature control and manipulation for decision support. Thus, the ontology will not be fixed, but will rather be integrated as a flexible wrapper for more efficient machine learning and knowledge discovery. Changes in the feature ontology, initiated by the user and leading to an increase in machine learning performance, may serve as an important source of novel knowledge in the domain.

This paper is organised as follows. In Section 2 we briefly analyse major existing medical ontologies. In Section 3 we give an overview of related work in mining complex data with taking into account feature semantics. In Section 4 we introduce the concept of feature ontology, consider how instance similarity can be redefined with it and discuss potential benefits of its use, and in Section 5 we consider one example application – the problem of predicting Atrial Septal Defect development. In Section 6 we present two techniques for the empirical evaluation of distance functions that can be used for the validation and fine-tuning of feature ontologies, and in Section 7 we conclude with a summary and directions for future research.

2 Ontologies in the Biomedical Domain

Clinical and biomedical applications often have to deal with large volumes of complex information originating from different sources, with different structures and different semantics. There is a long tradition of structuring clinical and biomedical information producing a vast number of standards and conceptual vocabularies that are reused in various medical applications. The efficient reuse of medical information requires the automatic processing, semantic integration, and semantic enhancement of medical knowledge resources enabled by an efficient and adequate knowledge organisation mechanism. There exists a variety of knowledge organisation systems that can be used for capturing semantic knowledge, including taxonomies, thesauri, and ontologies. All of these knowledge organisation systems express, either implicitly or explicitly, a more or less detailed semantic model of the world [14].

A taxonomy establishes a classification hierarchy of terms [25] by subsuming similar objects under distinct classes and subclasses. In contrast to taxonomies, thesauri provide additional means for refining the established classification hierarchies by constituting a fixed set of predefined relations between the concepts, enabling, for instance, the specification of similar or synonymic concepts [22]. Thus, by specifying a terminology of a particular domain, thesauri allow for the sophisticated and detailed annotation of objects of interest.

In computer science, an ontology is defined as "an explicit, formal specification of a shared conceptualisation" [13]. Through the specification of rules, ontologies enable the formulation of constraints, negations, logical functions, and mathematical operations. As taxonomies and thesauri are less expressive than ontologies, their captured content can easily be represented with ontological structures.

As already mentioned, in the domain of healthcare and biomedical informatics, a number of different knowledge repositories have been developed. Figure 1 provides an overview of relevant medical knowledge bases ordered by their size, i.e. the number of concepts. As one can see, the knowledge bases vary in the size (from 900,000 in UMLS to 40 in BioPax), in the way of knowledge organisation (ontology, meta-thesauri, thesauri, and taxonomy), in the covered subject domain, and in the format.

The Unified Medical Language System (UMLS) [8] originated in 1986 at the US National Library of Medicine (NLM) as a terminology integration project. It is a controlled compendium of medical vocabularies enhanced by mappings between them, with over 900 thousand concepts and 12 million relations between them. UMLS has three major components:

- the UMLS Meta-thesaurus being a repository of interrelated biomedical concepts integrating more than 60 families of biomedical vocabularies;
- the UMLS Semantic Network providing high-level categories for classifying every concept from Meta-thesaurus;
- the SPECIALIST lexicon yielding lexical resources and programmes for generating lexical variants of biomedical terms that enable the identification of lexically similar concepts.

Name	Domain	Size	Format	Type	Licensing	Institution
UMLS	Biomedical and Health domain	900 000 concepts	Relational files, OO-model, web access	Meta-Thesarus	UMLS Licensing	NLM/NIH
SNOMED CT	Healthcare Terms	400 000 concepts	proprietary format, web access	Thesaurus	Commerical & in UMLS	SNOMED International
ICD	Diseases and Injuries	60 000 concepts	book format, proprietary format, web access	Taxonomy	Commerical & in UMLS	WHO
FMA	Human Anatomy	70 000 concepts	Protege-Frames, web access	Ontology	Free	Univ. of Washington
MeSH	Medical Terms	22500 concepts	XML, ASCII, Tree, MARC, RDF/OWL, web access	Thesaurus	Free	NLM/NIH
Gene Ontology	Genetics	22 000 concepts	OWL, XML, OBO, Text, MySQL, web access	Thesaurus	Free	Collaborative
MGED	Microarray Experiments	230 concepts	OWL	Ontology	Free	MGED Society
BioPax	Biological Pathway Data	40 concepts	OWL	Ontology	Free	Collaborative

Fig. 1. Overview of major biomedical knowledge bases

UMLS concepts and relations are captured in a proprietary relational format and can either be accessed online via a web browser or are distributed on a CD-ROM or via FTP for offline usage. Although the access to the UMLS knowledge resources is free of charge, UMLS users have to sign a license agreement authorising them to use the UMLS content for research purposes.

The International Classification of Diseases (ICD) [16] is published by the World Health Organization (WHO). By providing means for the classification of known diseases and other health-related problems, the ICD enables the storage, retrieval and statistical analysis of diagnostic information. It is a taxonomy covering approximately 60 thousand concepts organised in 22 chapters of different classes of diseases. Its focus is to subsume similar diseases under classes, and infrequent diseases are sometimes combined without indicating profound similarity. ICD is commercially available on a CD or as a book. It can also be accessed free of charge with a web browser[1] and is a part of the UMLS knowledge repository.

Medical Subject Headings (MeSH) [23] is a thesaurus used for indexing and annotating journal articles and books in the PubMed database of biomedical literature. It establishes a set of poly-hierarchically structured concepts providing the basis for searching annotated medical literature at various levels of specificity. MeSH is created and maintained by the US National Library of Medicine (NLM). The MeSH Thesaurus establishes approximately 22,500 concepts (e.g., Disease, Cardiovascular Disease, Congenital Heart Defect, Atrial Septal Defect) and 83 qualifiers (e.g., Diagnose or Ultrasonography). Both concepts and qualifiers are hierarchically structured ranging from the most general to the most specific ones. The qualifiers provide means for addressing a particular view of a concept, e.g. by attaching the qualifier Ultrasonography to the concept Atrial Septal Defect (ASD) one can emphasise the ultrasonography-related diagnostic aspects of ASD. The MeSH thesaurus can be downloaded from the US National Library of Medicine[2] in the XML, ASCII, MeSH

[1] See www.who.int/classifications/apps/icd/icd10online/
[2] See www.nlm.nih.gov/mesh/MBrowser.html

Tree[3], or MeSH MARC[4] formats; it has also been converted to the RDF/OWL format [1][27]. It is also freely accessible through the UMLS knowledge repository.

The Systematised Nomenclature of Medicine Clinical Terms (SNOMED CT) [29] is a thesaurus of healthcare terms, covering clinical data for various diseases, clinical findings, and procedures. SNOMED CT is supported and maintained by SNOMED International, a division of the College of American Pathologists (CAP). It covers approximately 400 thousand concepts with formal logic-based definitions organised in 18 top-level hierarchies. Besides the classical "is-a" relations, it specifies more than 50 other relation types and encompasses more than 900 thousand instantiated relations. Being a very comprehensive standard, SNOMED CT cannot be provided and used in a classical book format, but has to be integrated into some access software. The SNOMED CT content is commercially distributed on CDs with or without additional access software and can be accessed free of charge via the SMOMED CT Browser[5] or through the UMLS knowledge repository.

The Gene-Ontology (GO) [3] project is a collaborative effort to provide a set of structured vocabularies for labelling gene products in different databases. Aiming to establish a controlled vocabulary for describing the functions of genes in a species-independent manner, the GO comprises of three independent vocabularies establishing terms for annotating molecular functions, cellular components and biological processes in gene products. In short, molecular functions detail what a gene product does at the biochemical level, biological processes capture broad biological objectives and cellular components specify the location of a gene product within cellular structures and within macromolecular complexes. Its approximately 22 thousand concepts are organised as a directed acyclic graph, i.e. a hierarchical structure with concepts having one or more parents, and with two relations, "is-a" and "part-of", linking the concepts. However, the GO specifies no associative relations across its three hierarchies. Being free of charge, the GO can be downloaded[6] in many different formats, such as OWL, XML, OBO, free text, and MySQL, as well as can be accessed online via the GO browser AmiGO[7].

The Microarray Gene Expression Data (MGED) ontology [30] provides standard terms for the annotation of microarray experiments. The ontology was created and is maintained by the MGED Society, an international organisation of biologists, computer scientists, and data analysts whose goal is to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. It encompasses 229 concepts and 110 properties. The concepts are defined and structured by formal-logic-based constraints, such as existential restrictions (specifying the existence of at least one relation of a given property to an individual being a member of a specific concept). Besides, MGED contains 658 instantiated concepts (instances) covering terms that are common to many microarrray experiments. MGED ontology is available for free in the OWL format.

The Biological Pathway Exchange (BioPAX) project [4] is a collaborative community effort aiming at the developing of a common exchange format for biological

[3] See www.nlm.nih.gov/mesh/mtr2007abt.htm
[4] See www.loc.gov/marc/specifications/speccharmarc8.html
[5] See snomed.vetmed.vt.edu/ sct/menu.cfm
[6] See www.geneontology.org
[7] See www.godatabase.org/cgi-bin/amigo/go.cgi

pathway data, capturing the key elements of data models from a wide range of popular pathway databases. The established BioPax ontology covers metabolic pathway information, molecular interactions, protein post-translational modifications, and supports the Proteomics Standards Initiative (PSI). To cope with the complexity of pathway data, the BioPAX working group has decided to use a multi-level development approach, i.e. BioPAX Level 1 is focused on the representation of metabolic pathway data, Level 2 expands the scope of Level 1 by including the representation of molecular binding interaction and hierarchical pathways, and further levels are also planned. The BioPAX Level 2 establishes 40 concepts and 33 properties. BioPAX is freely available and is currently implemented in the OWL format, but other implementations, such as XML Schema may be developed in the future.

The Foundational Model of Anatomy (FMA) is the most comprehensive ontology of human "canonical" anatomy [26]. It is developed and maintained by the School of Medicine of the University of Washington and the US National Library of Medicine (NLM). Beside the specification of anatomy taxonomy, i.e. an inheritance hierarchy of anatomical entities, the FMA provides definitions for conceptual attributes, part-whole, location, and other spatial associations of anatomical entities. By additionally allowing for attributing relations (i.e. relations can be described in more detail by attaching additional attributes) FMA is particularly rich with respect to the specification of relations and, thus, can cope with the requirements for the precise and comprehensive capturing of the structure of the body. FMA covers approximately 70 thousand distinct anatomical concepts and more than 1.5 million relations instances from 170 relations types. The FMA is freely available as a Protégé 3.0 project or can be accessed via the web browser Foundational Model Explorer (FME)[8]. Moreover, there exist research approaches focusing on the conversion of the frame-based Protégé version of FMA to the OWL DL format [12].

3 Related Work: Mining Complex Data and Data Mining with Ontologies

Medicine is a domain where large complex heterogeneous data sets are commonplace. Today, a single patient record may include, for example, demographic data, familiar history, laboratory test results, images (including echocardiograms, MRI, CT, angiogram etc), signals (e.g. EKG), genomic and proteomic samples, and history of appointments, prescriptions and interventions. And much if not all of this data may be relevant and may contain important information for decision support [19]. A successful integration of heterogeneous data within a patient record thus becomes of paramount importance. Various techniques for mining complex data that try to take into account feature heterogeneity and inter-feature relations were recently suggested.

Perhaps, the most straightforward way to construct a predictive model from heterogeneous data is simply to merge the heterogeneous features into a single heterogeneous feature-vector, neglect possible inter-relation among the features, and to employ some conventional inductive learning technique that is able to work with features of different types. For example, Berrar *et al.* [7] integrate clinical and

[8] See fme.biostr.washington.edu:8089/FME/index.html

transcriptional data in order to get improved classification performance for lung cancer survival prediction. Different learning algorithms are compared; boosted C5.0 decision trees, SVMs, probabilistic neural networks, *k*-nearest neighbour (*k*-NN), and MLP. MLP proved to be the most sensitive and less efficient with large heterogeneous feature vectors, while *k*-NN (somewhat surprisingly) and SVMs were the most robust classifiers resulting in the best predictive performance. Drawbacks of this "naïve" approach include a high risk of overfitting, the need in relatively low dimensionality ("the curse of dimensionality"), and the fact that not every technique supports feature heterogeneity.

A more sophisticated though not always applicable approach is to build an ensemble of models, one for each type of data. Futschik *et al.* [11] claim to be the first to focus on the combination of clinical and microarray-based classifiers. The hypothesis is that clinical information could be enriched with microarray data such that a combined ensemble predictor would perform better that a classifier based on either microarray data alone or clinical data alone. A Bayesian network was built on clinical data and an Evolving Fuzzy Neural Network (EFuNN) on microarray expression data in order to get an improved prediction accuracy for risk group prognosis in patients with lymphoma cancers. This approach has a number of advantages; 1) the heterogeneous data may be physically located at different sites and the computation can be parallelised; 2) there is relatively less risk of overfitting; 3) there is a possibility to apply more suitable techniques to a particular type of data (e.g., gene expression data), with a larger variety of available techniques. The main drawback of this approach is that it is usually applicable only when the different sources of data are representative enough of the problem, so that two or more relatively strong (better than a random guess) models can be constructed for the problem at hand.

Another common approach to take account of feature semantics for complex data consists in aggregating partial (dis)similarities computed on features of the same type possessing certain conceptual commonality. For example, Camps-Valls *et al.* [10] consider the use of composite kernels in order to combine spatial and spectral information for the enhanced classification of hyperspectral images. The main assumption is that the composite representation will allow modelling the dependencies between the extracted features to some extent and this will lead to a more intuitive definition of similarity between instances. It was demonstrated that the use of such composite kernels leads to a significant increase in predictive performance. However, the representation of feature interrelation is limited here to one-level grouping only (a grouping into non-overlapping feature subsets).

Another important related branch of research is focused on the use of taxonomies and ontologies in order to improve data mining results (normally ontologies are used in order to redefine similarity in data mining). Usually, such studies are based on taxonomies which help to structure the instance space in homogeneously represented classification problems (such as texts, annotated images and genomic data). Normally, in the core of such studies there is a concept of taxonomic or semantic distance which depends on the location of two concepts/instances in the taxonomy (ontology). Perhaps, the most well-studied area in this context is text mining where each document is often represented as a set of concepts (so called "bag-of-words" approach). The ontology used in this case can be a predefined graph-based model that reflects semantic relationship between concepts [24] or it can be derived from the texts themselves

using some unsupervised learning (one-level or hierarchical clustering) techniques (perhaps Baker and McCallum [5] were the first to apply this to text classification using so-called distributional clustering). Similar studies are done in order to find semantic similarity between annotated images for improved image retrieval based on the ontological representation of relations between the labels (see e.g. [17]).

With the appearance of the extensive Gene Ontology (GO) and the more and more acknowledged role of personalised genomic medicine, there emerged studies that tried to use the GO in order to define semantic similarity between genes in a similar way as it was done before for texts and images. Thus, Azuaje and Bodenreider [2] demonstrate that there is a significant correlation between the semantic similarity between a pair of genes and the probability of finding them in the same complex (cluster) in the analysis of gene expression data. This is claimed to be an assessment confirming to some extent the quality and consistency of the knowledge represented in the GO. In a related study, Bolshakova *et al.* [9] suggest to use the GO as the domain knowledge in order to validate clustering results and to determine the number of clusters in gene expression analysis.

A similar attempt to enhance inter-case similarity with the domain semantics, for the field of medicine, was performed by Melton *et al.* [20]. The ontology used to define semantic similarity was SNOMED CT, and each patient was represented by a "bag of findings" (compared to the "bag-of-words" representation of texts), where findings included SNOMED CT concepts extracted from free texts (clinical notes, discharge summaries, etc) and coded procedure and diagnosis data (ICD9-CM codes), from the Columbia University Medical Center (CUMC) data repository in 1989 – 2003. Patient cases included various disorders treated in the Medical Center. The use of taxonomic distance defined in SNOMED CT helped to improve the similarity assessment a little in comparison with the simple "bag of findings" similarity, checked by the correlation with the expert-perceived similarity. Although being an interesting research about inter-patient similarity, this study is still quite far from its practical application, as long as this similarity assessment is quite noisy and still poorly correlated with the expert-perceived similarity, and, on the other hand, most interesting for data mining medical data sets are rather disease-focused, where the "bag of findings" representation would not be suitable.

In summary, most related studies on the use of ontologies in data mining are focused on homogeneously represented cases and concentrate mainly on taxonomic distance and ontologies with "is_a" relations. These techniques are not particularly suitable for mining complex medical data, as long as medical data are usually heterogeneous and disease-focused, where it does not often make much sense to split the instance space into hierarchical concepts. On the other hand, to the best of our knowledge, no study focuses on mining disease-focused medical records with complex inter-feature relations.

The use of domain semantics in order to improve similarity search and decision support is also under active study in the Case-Based Reasoning community [28]. Although, the focus in the so-called knowledge-intensive similarity measures is on creating a customised distance function for each particular feature, instead of the conventional Euclidean and Manhattan (city-block) metrics, and not on the total aggregated distance (similarity). This research in CBR is rather complementary with

regards to the study presented in this paper in that the customised feature distances may be used as components in combination with a feature ontology which structures the feature space.

4 Feature Ontology: Redefining the Distance with Complex Data

As discussed above, today the most advanced ways of taking into account feature semantics in complex data consist in one-level feature grouping and either building a separate model for each semantic group (ensemble learning) or aggregating partial distances calculated within each group; or in the use of taxonomic distance over the hierarchical clustering of homogeneous features.

The basic idea in our suggested approach is to improve the performance of machine learning by redefining the concept of similarity with incorporating constraints provided by ontological domain knowledge, that is instead of simply providing a machine learning algorithm with features in the form of a single vector or a set of vectors, they will be semantically enhanced by integrating the graph structures of relevant domain ontologies. This can be achieved through the integration of all related ontological knowledge in a single so-called Feature Ontology, systematically structurising the feature space. The task of ontology integration is lately under active study in the area of ontology mapping [18]. Although a number of different solutions were proposed that may help in automating the integration in some cases, the process still remains routine and largely manual. The idea of structuring the feature space with a Feature Ontology is somewhat similar to object-oriented representation in CBR [6].

The main contribution of the feature ontology in terms of machine learning performance is in a more logical distribution of weights in the feature space, reflecting the semantics of the domain. To give a simple example, imaging features should not outweigh clinical features just because their number can be more than a thousand. They should be considered equally important for determining the distance if they are situated at the same level of the feature ontology (unless the expert intentionally specifies that for the current task a particular branch of features is more important).

A schematic distribution of weights in a feature ontology is shown in Figure 2. Feature ontology is a hierarchical structure in the form of a tree graph, where the nodes (N_n^l, where l is the level at which the node is situated, and n is the ordinal number of the node at level l) correspond to a group of features with common semantics, starting from the root node N^0 combining all the relevant features, and the leaves (f_n^l) include features. The tree structurises the feature space into k+1 levels. Leaves (features) can be situated at any level of the tree (although in the figure they are shown at level k+1 only for the sake of simplicity). The graph is weighted; weights are assigned to its edges (branches of the tree). Weight w_n^{lm} corresponds to the n-th child edge originating from the m-th node at level l.

The weights of all child branches of a node in such a feature ontology should sum to one: $\sum_n w_n^{lm} = 1$. By default, if no prior knowledge is available, the weights of child branches should be equal. The weight of a particular feature f_n^l is defined as

the product of the weights in the tree on the path towards this feature: $w(f_n^l) = \prod_{i=o}^{l-1} w_*^i(f_n^l)$, where $w_*^i(f_n^l)$ is the weight of an ancestor branch of level *i* for feature f_n^l. According to this definition, the deeper a node (or a feature) is in the hierarchy, the less influence it will have in the similarity assessment.

The weights in the feature ontology can be established by an expert (satisfying to the defined constraints) and/or they can be fine-tuned with some machine learning algorithm (e.g. using a form of genetic search). The resulting feature weights can be used in combination with any distance function supporting feature weighting. In the simplest case, the overall distance can be calculated as the weighted average of contributing partial distances corresponding to each relevant feature. Each partial distance may be different and may take into account the type and semantics of a particular feature but should be normalised (i.e., it should be in the range from 0 to 1).

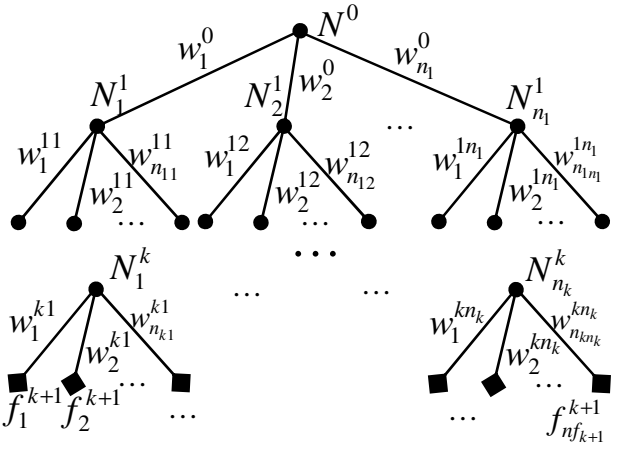

Fig. 2. Schematic distribution of weights in a feature ontology

One of the most common machine learning techniques where similarity between instances is explicitly calculated is instance-based learning (e.g., *k*-nearest neighbour classification, *k*-NN) [21]. The distance function that lies in the core of *k*-NN is normally defined for a single set of unrelated features representing the problem. By semantically enhancing the set of relevant features by integrating medical domain knowledge and redefining the distance function, the patient diagnostic (classification) accuracy can be improved. This is the most important expected benefit from the use of the feature ontology. Presumably, feature ontology will also be useful for other learning techniques, implicitly taking instance similarity into account, in order to improve their performance.

Besides the improved predictive performance, the graph-based representation of the feature ontology can be convenient for an expert in order to establish different feature weights by changing the weights of branches corresponding to a certain semantic group of features, instead of assigning importance to each particular feature.

The feature ontology can be presented to the expert as part of the system's GUI and might provide an effective way for feature control and manipulation for decision support. Thus, the ontology will not be fixed, but will rather be integrated as a flexible wrapper for more efficient machine learning and knowledge discovery. Changes in the feature ontology, initiated by the user and leading to an increase in machine learning performance, may also serve as an important source of novel knowledge in the subject domain.

5 Example: Prognosis of Atrial Septal Defect Development

The authors of this paper are participants of the EU's 6th Framework Programme's (FP6) Integrated Project "Health-e-Child" (*www.health-e-child.org*), which was started in 2006. The present study is motivated by the main objectives of the project. The focus of the project is on the vertical integration of biomedical data, information and knowledge spanning the entire spectrum from genetic to clinical to epidemiological with the aim of gaining a comprehensive view of a child's health and providing the basis for improving individual disease prevention, screening, early diagnosis, therapy and follow-up of paediatric diseases. Health-e-Child focuses on some carefully selected representative diseases in three different categories: paediatric heart diseases, inflammatory diseases, and brain tumours.

Atrial Septal Defect (ASD) which is characterised by a hole in the atrial septum is a congenital heart defect, is perhaps the most common cause of Right Ventricle Overload (RVO) and is among the most common paediatric heart diseases [15]. Usually, the intervention to treat ASD is performed at a pre-school age (4-6 years of age). However, the size of the hole is constantly changing with time and in some cases the defect may get worse, so that time can be lost to do device closure (trans-catheterisation), and only an open-heart surgery can be performed. On the other hand, in some cases the hole in the septum (even a moderate-sized one, even at the age of 4-6) may close on its own [15]. Up to know the phenomenon of ASD development is rather unclear to physicians and data-driven decision support will be of great help here. Another problem where decision support might be useful is possible complications after trans-catheterisation. E.g., there are cases where tissue erosion and rupture is reported, which might need another trans-catheterisation procedure, or even surgery. Distinguishing potentially high-risk patients in terms of possible complications after ASD treatment is another important task in this context.

Using different examinations and tests, such as echocardiogram, chest X-ray, electrocardiogram, Doppler study, MRI, and cardiac catheterisation, a physician collects all available information for determining the diagnose and the most suitable treatment. As the prognosis of ASD development depends on heterogeneous features of different kind representing clinical data, genetic data, ECG, and imaging data, the resulting feature space becomes quite complex. Therefore, we represent the features in a hierarchical semantically enhanced structure by establishing a feature ontology. By mapping and relating the concepts of the feature ontology to existing medical ontologies (see Section 2), valuable medical background knowledge, such as

relations between concepts, constraints, and axioms can be used for refining the feature ontology, providing the basis for improving the predictive performance of decision support.

Integrating machine learning algorithms with feature ontologies is especially important and beneficial in problem domains where the structure of the feature space is complex and significantly unbalanced, where the features are diverse and represent heterogeneous concepts. This is often the case with biomedical problems, and the task of prognosis of cardiological disease (ASD) development is a good representative of such a problem.

Figure 3 demonstrates an excerpt from the feature ontology for the problem of prognosis of ASD development. Some branches are marked with different weights, reflecting the relative importance of corresponding features (these weights are arbitrary and are used for the purpose of illustration only). The weights can be fine-tuned both by the expert and in an automated way using a machine learning technique (e.g., a genetic algorithm). Fine-tuning weights corresponding to different branches in the feature ontology for a particular problem may lead to the discovery of important problem-specific knowledge.

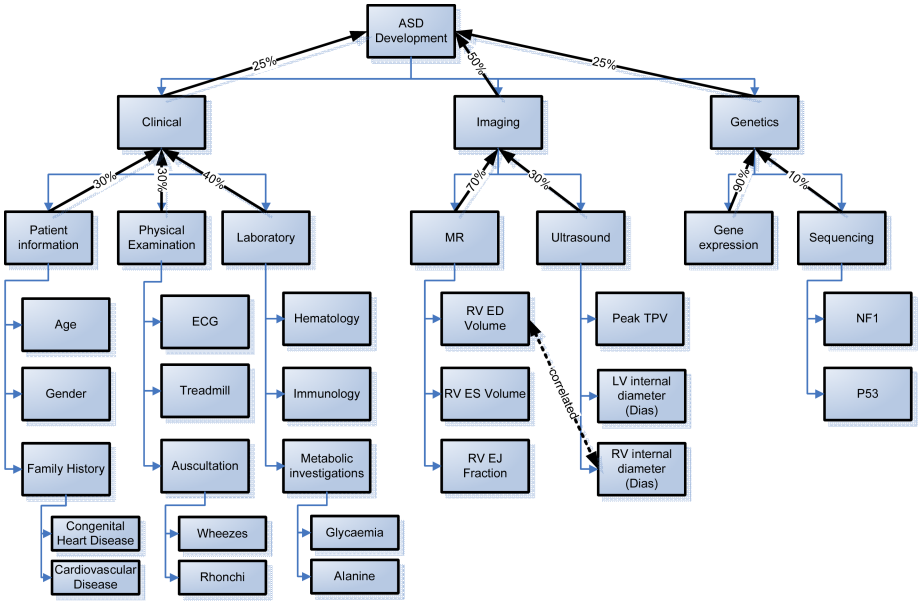

Fig. 3. Example Feature Ontology for the problem of prognosis of ASD development

Other information from the existing medical ontologies can be useful as well. For example, normal value ranges for different standard medical features can be extracted from ontologies (a light-blue box at the bottom demonstrates the normal range for Alanine in the figure). The normal value ranges are important for outlier removal and they may influence the distance metric as well. The feature ontology may also represent correlated or redundant features, which may have an influence on determining

the inter-patient distance. In the given example, the blood volume of the right ventricle cavity can be determined using both ultrasound and magnetic-resonance images (MRI). The estimate received with MRI is usually more exact, and the ultrasound estimate may be ignored when MRI information is available.

6 Two Approaches to Distance Evaluation

There are two basic approaches to distance evaluation that can be applied to the validation and fine-tuning of any distance function in general, and a feature ontology-based distance in particular; evaluation based on expert-perceived similarity, and automatic data-driven wrapper-like evaluation.

The first approach was used in [20]. Its main idea consists in ranking a set of instances by a group of experts in a subject domain, according to the perceived similarity to another control instance (this process can be repeated for a number of control instances). Then the resulting rankings can be compared with the one produced by the distance function under study. For example, Spearman's rank correlation coefficient can be used for the comparison. The quality of the distance function is assumed here to be proportional to the average expert-function rank correlation (the bigger the average correlation between the expert- and distance function-produced ranks the better).

A serious drawback of this approach is the fact that the expert-perceived similarity may be rather subjective and context-dependent. However, experiments show that inter-expert rank correlation is usually significant enough even for very heterogeneous complex domains as in [20], so that such a comparison is appropriate. Inter-expert rank correlation may serve as a measure of expert agreement and partly validity of such an approach. Due to the experts' involvement, this approach may be applied to relatively small data sets only, which raises a question about the generality of findings.

Another approach is to use the distance function under study as an element in a learning algorithm that is used as a wrapper. The assumption is that the quality of the distance function will then be reflected by the performance of the learning algorithm on validation data. This approach is often used in machine learning research; it is applied for parameter selection and tuning in machine learning algorithms [21]. Its advantage is that the distance function is evaluated (or updated) in the context of the task being solved. Thus, for our example from the previous section, a good distance function should result in a better predictive performance (classification accuracy) of ASD classification. Any appropriate data-driven validation technique can be used in combination with an appropriate learning algorithm for wrapper-based distance evaluation. For example, cross validation together with *k*-NN classification can be used in our example.

One drawback of this approach is that enough data is needed in order to avoid potential overfitting and to provide valid evaluation. Thus, in some domains there might be simply not enough data for a separate validation set, and sometimes even for cross validation (for ASD behaviour prognosis, the number of instances is normally of the order of 10, which is significantly exceeded by the number of features, which are of the order of 10^4 or even 10^5). When enough data is available, this approach can be applied iteratively, to search for a better distance function in the space of valid distance functions.

7 Conclusions

In this paper we identify a problem, give a review of related work and propose one solution for the task of integration of machine learning techniques with existing ontological knowledge, that is especially important for biomedical domains where data is often naturally complex and is represented by a large heterogeneous feature-vector. Our main assumption is that structurising the feature space into a so-called *Feature Ontology* will reflect semantics of the domain and thus may help in improving the performance of machine learning techniques, through the redefined distance (similarity) function. We give an example for the task of prognosis of ASD development in children, and analyse two techniques that can be used for the evaluation and refining of a feature ontology. Beside the benefit in terms of improved predictive performance, the feature ontology may also become an important element of the Graphical User Interface, providing a means to data access and manipulation, in the context of the classification problem under consideration.

We would like to emphasise here that, in the context of this integration, the task of the creation of feature ontology becomes *central*, and this task is, unfortunately, not trivial at all as it may seem (especially taking into account the common complexity of biomedical problem domains). Some techniques were developed, in the area of ontology mapping, that may help to partially automate this process, though this process still remains largely routine and manual, needs a skilful expert and is based on the expert's knowledge and intuition. The feature ontology needs to be carefully developed, and it needs to focus on the classification task under study (the feature space should be structured with the classification task in mind), otherwise it is difficult to expect an improved similarity measure. The usual computer science principle "GIGO" (Garbage In, Garbage Out) works here as well. If enough data is available, data-driven feature ontology refinement may be applied, taking the expert ontology as a starting point in the search.

Our future work includes the creation and evaluation of feature ontologies for the medical problems within the Health-e-Child project. Another interesting direction for further research is the incorporation of various relations available in the existing ontologies in the distance calculation. For example, many ontologies include information about correlation between relevant features.

Acknowledgements. This work has been partially funded by the EU project Health-e-Child (IST 2004-027749). The authors wish to acknowledge support provided by all the members of the Health-e-Child consortium in the preparation of this paper.

References

1. Assem, M., Menken, M., Schreiber, G., Wielemaker, J., Wielinga, B.: A method for converting thesauri to RDF/OWL. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 17–34. Springer, Heidelberg (2004)
2. Azuaje, F., Bodenreider, O.: Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study. In: Proc. IEEE Symposium on Bioinformatics and Bioengineering, BIBE 2004, pp. 317–324. IEEE Press, Los Alamitos (2004)

3. Ashburner, M., et al.: Creating the gene ontology resource: design and implementation. Genome Research 11(8), 1425–1433 (2001)
4. Bader, G., Cary, M. (eds.): BioPAX – Biological Pathways Exchange Language, Level 2, Version 1.0 Documentation, BioPAX Working Group (2006) available at http://www.biopax.org
5. Baker, L.D., McCallum, A.K.: Distributional clustering of words for text classification. In: Proc. 21st ACM Int. Conf. on Research and Development in Information Retrieval SIGIR'98, pp. 96–103. ACM Press, New York (1998)
6. Bergmann, R., Kolodner, J., Plaza, E.: Representation in case-based reasoning. In: Knowledge Engineering Review, vol. 20, pp. 209–213. Cambridge University Press, Cambridge (2005)
7. Berrar, D., Sturgeon, B., Bradbury, I., Downes, C.S., Dubitzky, W.: Microarray data integration and machine learning methods for lung cancer survival prediction. In: 4th Int. Conf. Critical Assessment of Microarray Data Analysis, CAMDA, pp. 43–54 (2003)
8. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. In: Nucleid Acids Research, vol. 31, pp. 267–270. Oxford University Press, Oxford, UK (2004)
9. Bolshakova, N., Azuaje, F., Cunningham, P.: Incorporating biological domain knowledge into cluster validity assessment. In: Rothlauf, F., Branke, J., Cagnoni, S., Costa, E., Cotta, C., Drechsler, R., Lutton, E., Machado, P., Moore, J.H., Romero, J., Smith, G.D., Squillero, G., Takagi, H. (eds.) EvoWorkshops 2006. LNCS, vol. 3907, pp. 13–22. Springer, Heidelberg (2006)
10. Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters 3(1), 93–97 (2006)
11. Futschik, M.E., Sullivan, M., Reeve, A., Kasabov, N.: Prediction of clinical behaviour and treatment for cancers. Applied Bioinformatics 2(3), 53–58 (2003)
12. Goldbreich, C., Zhang, S., Bodenreider, O.: The foundational model of anatomy in OWL: experiences and perspectives. In: J. of Web Semantics: Science, Services, and Agents on the World Wide Web, vol. 4, pp. 181–195. Elsevier, North-Holland, Amsterdam (2006)
13. Gruber, T.: Towards principles for the design of ontologies used for knowledge sharing, Human and Computer Studies, vol. 43, pp. 907–928. Academic Press, San Diego (1995)
14. Hodge, G.: Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files, The Digital Library Federation (2000)
15. Hanslik, A., Pospisil, U., Salzer-Muhar, U., Greber-Platzer, S., Male, C.: Predictors of spontaneous closure of isolated secundum atrial septal defect in children: a longitudinal study. Pediatrics 118(4), 1560–1565 (2006)
16. International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10), World Health Organization [classifications/apps/icd/ icd10online/], available at http://www.who.int/
17. Janecek, P., Pu, P.: Searching with semantics: an interactive visualization technique for exploring an annotated image collection. In: Proc. On The Move to Meaningful Internet Systems 2003: OTM 2003 Workshops. LNCS, vol. 2889, pp. 185–196. Springer, Heidelberg (2003)
18. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. In: Kalfoglou, Y., Schorlemmer, M., Sheth, A., Staab, S., Uschold, M. (eds.) Semantic Interoperability and Integration, Dagstuhl Seminar Proceedings 4391, IBFI (2005) [available at drops.dagstuhl.de/opus/volltexte/2005/40]

19. Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., Tarczy-Hornoch, P.: Data integration and genomic medicine. Methodological review, Biomedical Informatics 40, 5–16 (2007)
20. Melton, G., Parsons, S., Morrison, F., Rothschild, A., Markatou, M., Hripcsak, G.: Inter-patient distance metrics using SNOMED CT defining relationships. Biomedical Informatics 39, 697–705 (2006)
21. Mitchell, T.M.: Machine Learning. McGraw Hill, New York (1997)
22. Moench, E., Ullrich, M., Schnurr, H., Angele, J.: SemanticMiner – ontology-based knowledge retrieval. Universal Computer Science 9(7), 682–696 (2003)
23. Nelson, S., Johnston, D., Humphreys, B.: Relationships in medical subject headings. In: Bean, C., Green, R. (eds.) Relationships in the Organization of Knowledge, pp. 171–184. Kluwer Academic, Boston, MA (2001)
24. Oleshchuk, V., Pedersen, A.: Ontology-based semantic similarity comparison of documents. In: DEXA Workshops 2003, pp. 735–738. IEEE CS Press, Los Alamitos, CA, USA (2003)
25. Panyr, J.: Thesauri, semantic nets, frames, taxonomies, ontologies – conceptual confusion or conceptional diversity? In: Harms, I., Luckhardt, D., Giessen, H. (eds.)Information and Language – Contributions from Computer Science, Computer Linguistics, Librarianship, and Related Disciplines, Saur-Verlag, pp. 139–152 (In German) (2006)
26. Rosse, C., Mejino, J.: A reference ontology for biomedical informatics: the foundational model of anatomy. Biomedical Informatics 36, 478–500 (2003)
27. Soualmia, L.F., Golbreich, C., Darmoni, S.J.: Representing the MeSH in OWL: towards a semi-automatic migration. In: Proc. 1st Int. Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), Whistler, Canada, pp. 81–87 (2004)
28. Stahl, A.: Learning of Knowledge-Intensive Similarity Measures in Case-Based Reasoning, Ph. D. Thesis, University of Kaiserslautern, Germany (2004)
29. Stearns, M., Price, C., Spackman, K., Wang, A.: SNOMED: clinical terms: overview of the development process and project status. In: Proc. Annual Symposium of American Medical Informatics Association, AMIA 2001, Hanley & Belfus, pp. 662–666 (2001)
30. Whetzel, P., Parkinson, H., Causton, H., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., Sansone, S., Taylor, S., White, J., Stoeckert, C.: The MGED ontology; a resource for semantics-based description of microarray experiments. In: Bioinformatics, vol. 22, pp. 866–873. Oxford University Press, Oxford, UK (2006)
31. Zighed, D.A., Ras, Z.W. (eds.): In: Proc. 2nd IASC Workshop on Mining Complex Data, in conjunction with IEEE Int. Conf. on Data Mining ICDM 2006, Hong Kong (December 2006)