Ontology-Based Tools to Expedite Predictive Model Construction

Peter Haug, MD\textsuperscript{1,2}; John Holmen, PhD\textsuperscript{1}; Xinzi Wu, PhD\textsuperscript{1}; Kumar Mynam\textsuperscript{1}; Matthew Ebert, MS\textsuperscript{1}; Jeffrey Ferraro, PhD\textsuperscript{1}

\textsuperscript{1}Intermountain Healthcare, Salt Lake City, Utah; \textsuperscript{2}University of Utah

Abstract

Large amounts of medical data are collected electronically during the course of caring for patients using modern medical information systems. This data presents an opportunity to develop clinically useful tools through data mining and observational research studies. However, the work necessary to make sense of this data and to integrate it into a research initiative can require substantial effort from medical experts as well as from experts in medical terminology, data extraction, and data analysis. This slows the process of medical research. To reduce the effort required for the construction of computable, diagnostic predictive models, we have developed a system that hybridizes a medical ontology with a large clinical data warehouse. Here we describe components of this system designed to automate the development of preliminary diagnostic models and to provide visual clues that can assist the researcher in planning for further analysis of the data behind these models.

Introduction and Background

Electronic medical record (EMR) systems have become a standard contributor to modern healthcare. In many clinical settings, they capture all of the information recorded as a part of the documentation of care. In the process, they provide tools to streamline medical processes as well as opportunities to enhance the medical decision making process both indirectly, through effective organization and presentation of patient data, and directly, through clinical decision support (CDS) technologies.

The clinical data that accumulates as a result of computer-assisted healthcare has an additional role to play. As it is collected and curated over time, it can be analyzed to yield insights into the diagnosis and treatment of disease. The process of mining this data can result in new medical knowledge that can lead to changes in care. It can also support new CDS-based interventions built upon models derived from a combination of clinical data and medical knowledge.

We believe that medical data mining focused on the development of clinically useful models has large potential value as a way to increase the ability of EMR systems to standardize and expedite care. To support research in this area, we have constructed a system whose focus is the use of data from a large enterprise data warehouse (EDW) combined with medical knowledge stored in a disease-oriented ontology. This combination is used to automate the construction of computable diagnostic models. We call this system the Ontology-driven Diagnostic Modeling System (ODMS)\textsuperscript{1}. Here we will describe features of this system designed specifically to support a researcher as she/he generates and evaluates tools for real-time clinical diagnosis.

The system described below takes advantage of an extensive EDW created by Intermountain Healthcare, a large healthcare delivery system in Utah. This data warehouse captures data representing ~3 million visits each year collected while serving ~1 million patients. The data represents outpatient, inpatient, and emergency services and are a common focus for a variety of clinical research activities.

In recent years, we have brought added attention to the development of decision aids in the Emergency Department, notably a diagnostic and therapeutic CDS system for community acquired pneumonia\textsuperscript{2}. The resulting system currently provides daily care in a group of 4 Utah hospitals. It is a prototype for the systems the ODMS is designed to produce.

The ODMS provides a group of embedded tools whose goal is to reduce the effort necessary to build diagnostic models. Key components of this process are addressed by the system including:

- Cohort development: the identification of patient subgroups with and without the target condition.
- Feature selection/extraction: the selection and extraction of those data elements relevant to the diagnostic problem.
- Model specification: the choice of a predictive modeling approach for the required diagnostic model.
- Model evaluation: tools that provide an initial exploration of the quality of the diagnostic models developed.
The goal of this system is to provide a semi-automated tool to generate and return to the user a collection of outputs designed to provide a good starting point from which to continue the construction of a functioning diagnostic model. The results of this initial evaluation assist the user in determining whether the available data in the EDW is indeed adequate to produce a usable clinical tool. Below we describe the components mentioned above and illustrated some of the features of the ODMS.

Methods
The ODMS uses as input the EDW and a diagnostic ontology. It outputs four key products. These outputs include:

- A diagnostic predictive model produced by the system.
- A list of the features found to be relevant to the diagnostic problem addressed.
- An analysis of the quality of the predictive model.
- The raw data generated by the system and used to develop the diagnostic model.

These products are described and illustrated below. We will use a project where we have begun building a model for the diagnosis of pulmonary embolism (PE) to illustrate some of the features available in the ODMS.

A disease ontology is central to the operation of the ODMS and is used in the creation of the data sets necessary to the modeling effort. This ontology is designed to capture relationships that support identification of patient subgroups with and without the target disease and to identify clinical features important in the diagnosis of this condition. It has been developed largely through a combination of various existing sources followed by the manual introduction of relevant properties connecting classes with the needed taxonomic components.

Figure 1 contains a fragment of the ontology defining a group of patients with PE. This fragment is displayed in the ontology management tool, Protégé. It indicates the categories of patients for which to search. In this case, the ODMS uses taxonomic explosion to expand the “All_Pulmonary_Embolism” concept to include the 11 ICD-9 codes necessary to completely represent this concept. These concepts are harvested from the taxonomic trees embedded in the ontology (figure 2).

The ODMS also links to information that can tell the system which clinical environments to use when extracting a study cohort (in this case, ED or inpatient locations within a regional collection of care environments called the “Urban Central Region” (UCR)). In this case, it specifies the generic “Person” as the subject type, although further restrictions using age and sex are available.

The concepts in the ontology are linked, wherever possible, to standard national or international coding systems. For instance, diagnoses are linked to ICD-9 codes, laboratory results are linked to LOINC, and medications are links to RXNORM. Thus, when the ODMS reads through the ontology to construct a query against the EDW, the resulting queries are couched in terminologies that are largely standards-based. We anticipate that this will help us generalize this system to function in other settings.

Diagnostic Model
The ODMS is constructed to support the incorporation of a variety of predictive modeling tools within its framework. At present, we have incorporated four such tools. They are Naive Bayesian Models, Tree-Augmented
Bayesian Networks\textsuperscript{4}, K2-structured Bayesian Networks\textsuperscript{5}, and Random Forest Classifiers\textsuperscript{6}. When appropriate, the system automatically discretizes continuous data using a Minimum Description Length algorithm (MDL)\textsuperscript{7}.

The modeling framework within the ODMS is architected to accommodate models from a variety of sources. Those listed above are based on components from Weka\textsuperscript{8}, a general purpose, data-mining environment; Netica\textsuperscript{9}, a Bayesian network tool; and custom predictive modeling tools that have resulted from local development efforts. When a new analysis is begun, the system provides a default protocol to guide the process. This can be reviewed and modified at system initiation. For users who do not wish to accept system defaults, alternate sets of components and procedures can be configured from a setup page.

**Relevant Features**

As indicated above, the ODMS uses ICD-9 codes extracted from an ontology-based, disease modeling system to identify groups of patients with and without the target diagnosis. The ontology also contains links from diseases to their diagnostic features. These include laboratory results, vital signs, x-ray results, nurse charting information, and chief complaints. The system interrogates the ontology for these relationships. They are used to build queries against data in the EDW and to retrieve this information for patients with and without the target diagnosis. As a part of this process, it assembles an exhaustive list of features, which typically includes dozens of variables that may contribute to the diagnostic model. In case of pulmonary embolism, this process generated seventy-four proposed variables to be used in the initial diagnostic model. Once the proposed data set is assembled, the ODMS activates a feature selection process designed to identify a subset of the features that will be the focus of further modeling efforts.

The ODMS uses an initial filtering step to reduce the number of features prior to model construction. The default number of variables to include is 15, but this parameter can be changed on the system configuration page. A simple testing procedure is used to rank these features by discrimination power. The system employs a Chi-squared algorithm to give an initial assessment of the degree of association between each feature and the disease. Figure 3 displays the ranked list used by the system to display these associations while figure 4 depicts the graphical output of this ranking procedure.

![Figure 3: A ranked list of the diagnostic features to be used in development of the diagnostic model. The features in the list are those that will be used to suggest the diagnosis of pulmonary embolism prior to definitive testing.](image)

![Figure 4: A graphical representation of the ranking for features based on their discriminating power (Chi-square score) for diagnosing pulmonary embolism. Within the 74 extracted features the $\text{Chi}^2$ approaches 0 after the 40th variable.](image)

The Chi-squared test statistic is used as an initial filter for the proposed features. The user determines the number of variables that should be included in the analysis from this feature selection process. Additional feature selection algorithms may further reduce the number of variables as a part of the model generation process.

**Model Evaluation**

As a part of its initial analysis, the ODMS produces the predictive model specified at the start of the procedure. It also does an initial evaluation of this model. This is useful to reassure the user of the validity of the results. Issues
such as over-fitting are typically avoided in this way. Two algorithms are offered as a part of this initial evaluation step, N-fold cross validation and bootstrapping. In each case the system defaults to a 10-fold approach, but the user can modify this to include any number of steps deemed appropriate.

Visualization is a vital part of model evaluation. For modeling tools that return a numeric value representing the likelihood that a case will fit a specific category (diagnosis), the system provides a number of graphical outputs including ROC curves (figure 5), recall vs. precision graphs (figure 6), and others. The goal is to both give an overall sense of the quality of the predictive model and to allow exploration of the characteristics of this model. Tools of this sort allow the user to estimate the operating characteristics that a diagnostic system will have when used in a real-world clinical setting.

The ODMS not only produces an initial predictive model with a minimum of user effort, it also supports refinement of this model under control of the user. To support this process of refinement, the ODMS includes graphical tools for visually comparing different models. Figure 7 compares two ROC curves from an initial and a refined diagnostic model developed for pulmonary embolism.

**Raw Modeling Data**

As mentioned above, the ODMS returns the raw data used in the initial modeling effort along with the model and evaluation. This allows the system’s user to inspect this data, to modify it when appropriate (e.g. create derived variables, reduce redundancy, etc.) and to resubmit it to the modeling system for re-evaluation. The user may also process this data using other data mining systems that provide access to predictive algorithms not yet available in the ODMS.

The data extracted from the enterprise data warehouse is represented in the form of an ARFF file. This standard format is compatible with a number of data mining tool kits, notably Weka, an extensible data-mining framework from New Zealand. These files are easily converted to other standard data analysis formats. We have found this capability to be useful for testing new components prior to adding them to the ODMS toolkit.

**Results**

The ODMS is capable of providing valuable assistance in automating data mining processes focused on the creation of diagnostic systems. We are currently using the system in modeling efforts for pneumonia, sepsis, and pulmonary embolism. The ODMS’s major drawback is that our disease ontology is incomplete. Initial development efforts were directed toward demonstrating the value of this knowledge source as a tool in predictive data mining. The initial focus was in pulmonary diseases. We are now extending the ontology to encompass other diagnostic categories of interest to local research communities.
The use of ontologies to expedite clinical research is attractive for several reasons. First, a great deal of fundamental medical knowledge can be encoded in ontologies. A variety of knowledge assemblies have been created that capture important taxonomic relationships. These relationships are readily available in terminologies like ICD-9, SNOMED, LOINC, etc. To extend these structured terminologies we have been working to introduce meaningful links between the concepts represented (disease leads to clinical observation, medication treats disease, etc.). Many of these links are available in some form already; bringing them together into a unified ontology is the logical next step.

But an ontology that captures the relationships relevant to clinical research is insufficient. The key to model building is to link the concepts in the ontology to data in a clinical data warehouse. These links allow the ontology to direct the collection of data relevant to a particular diagnostic modeling problem.

In future versions of the system described here, we hope to both extend the ontology and to improve linkage to our enterprise data warehouse. Toward this end, we have been constructing a special collection of data derived from our EDW. We refer to this as the Analytic Health Repository and it has the expressed goal of tying all data possible to national or international medical terminologies. These are the same terminologies used in the disease ontology, and we anticipate that this approach will make the ODMS easier to enhance and maintain. We also hope that integrating knowledge stored in ontologies with the clinical experience represented by large data warehouses will allow us to provide medical researchers with an efficient and effective environment in which to ask important medical questions.

Supported in part by Grant LM010482 from the National Library of Medicine.

References

1. Haug PJ, Ferraro JP, Holmen J, et al. An Ontology-Driven, Diagnostic Modeling System. J Am Med Inform Assoc Published Online First: [March 24, 2013] doi:10.1136/amiajnl-2012-001376.
2. Dean NC, Jones BE, Ferraro JP, et al. Performance and utilization of an emergency department electronic screening tool for pneumonia. JAMA Intern Med. Published Online first: 18 Mar 2013.
3. Protégé-OWL: http://protege.stanford.edu/overview/protege-owl.html (accessed 21 Dec 2012).
4. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning 1997;29:131–63.
5. Cooper GF, Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9, 309-347 (1992).
6. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
7. Fayyad UM, Irani KB. Multi-interval discretization of continuous valued attributes for classification learning. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93); Chamberry, France. 1993:1022–7.
8. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009;11:10–8.
9. Netica Software. Norsys Software Corporation; [March 16, 2011]. http://www.norsys.com (accessed 21 Dec 2012).
10. ARFF: http://weka.wikispaces.com/ARFF (accessed 21 Dec 2012).
11. Weka: http://www.cs.waikato.ac.nz/ml/weka/ (accessed 21 Dec 2012).