Applying an Ontology-guided Machine Learning Methodology to SEER-MHOS Dataset

Hua Min, Hedyeh Mobahi, Sava Vukomanovic, Katherine Irvin, Ilirjeta Krasniqi, Sanja Avramovic, and Janusz Wojtusiak*

Department of Health Administration and Policy, George Mason University, Fairfax, VA 22030, USA

ABSTRACT

Motivation: Bio-ontologies are becoming increasingly important in both knowledge representation and machine learning fields. Bio-ontologies are used to model the healthcare knowledge using a set of concepts and relationships among those concepts. Meanwhile, mainstream machine learning focuses on the ability to process massive amounts of data and build accurate models, but few methods provide intelligent assistance to address challenges in the biomedical data analysis such as the complexity, heterogeneity, and semantics of healthcare data. In this paper, an ontology-guided machine learning method is described and applied to discovery of patterns of patients' characteristics affecting their ability to perform activities of daily living. Bio-ontologies are used to provide computable knowledge for machine learning methods to "understand" biomedical data. SEER-MHOS data mapped to UMLS are analyzed to discover the patterns.

1 INTRODUCTION

1.1 Bio-ontologies

An ontology formally represents domain knowledge as a set of concepts and relationships between those concepts. In Artificial Intelligence (AI), ontologies have been applied as artifacts to represent human knowledge and as critical components in knowledge management such as the Semantic Web and business-to-business applications. In the biomedicine field, ontologies have been widely adopted and used in knowledge management, data integration, and decision support & reasoning (Bodenreider 2008, Madsen 2010). Currently bio-ontologies are used in the emerging data-driven science including data mining and machine learning (Hoehndorf 2015).

There are many existing bio-ontologies but each one has a scope, purpose and role of its own. Therefore, there are communication barriers between various information systems or applications if different vocabularies are used in different information systems and users. In order to solve these barriers, the Unified Medical Language System (UMLS) was developed by the National library of Medicine (NLM) in 1986 (Lindberg 1990) and it's being constattly updated since. It is a repository of medical vocabularies and has three major components: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. The 2015AB version of the UMLS contains more than 3 million concepts (CUIs) and 12 million unique concept names (AUIs) from over 150 source vocabularies. The UMLS establishes mappings between those bio-ontologies by assigning a concept unique identifier (CUI) to names from various vocabularies that have the same meaning. The mappings among these vocabularies allow computer systems to translate data among the various information systems. The rich relationships (22 million) among concepts in the UMLS also provide a solid foundation for reasoning the medical knowledge.

1.2 Machine Learning

Precision Medicine is an emerging approach for disease prevention and treatment that takes into account people's individual information including genomics, environment, and lifestyle (https://www.nih.gov/precision-medicine-initiative-cohort-program). This new era requires advanced methodologies for analyzing, synthesizing, and disseminating heterogeneous data and knowledge in order to discover relationships and create computational models for improving care and wellbeing. The focus on the Big Data analysis in the biomedical field creates an even greater need of advanced computational methodologies for turning data into computer-interpretable forms and using them to promote patient-centric healthcare. Machine Learning (ML) is widely used in creating predictive models on Big Data analysis and it is gaining popularity in medical and health applications.

One major challenge in machine learning is to communicate the meaning of attributes and their values to the learning algorithm. While many machine learning researchers focus on the ability to process massive amounts of data and build accurate models, the complexity, heterogeneity and semantics of biomedical data, along with transparency of created evidence, are often outside of the mainstream research. Even fewer methods allow for aiding data analysis by using ontologies. Two important existing machine learning areas that deal with complex data are statistical relational learning (De Raedt 2008, Getoor and Taskar, eds. 2007) and inductive logic programming (Lavrac and Dzeroski 1994). Both areas

[*] To whom correspondence should be addressed.

are concerned with the more general problem of learning from datasets with complicated structures (relational databases or predicates). However, the majority of ML methods (including the most popular Support Vector Machines, Random Forests, Logistic Regression, etc.) work with data stored in flat tables (or extracted into flat tables), and almost exclusively focus on numeric data, while ignoring semantic relationships (meaning) of data elements. Few methods allow for additional ad-hoc encoding of ontologies, hierarchies and other coding systems. Machine learning methods should be able to take advantage of the known complex relationships between attributes and values given by bio-ontologies rather than solely rely on simple structured data. Healthcare is particularly rich in knowledge, but few methods can use it.

1.3 AQ21 Rule Learning and Applications

AQ21 is a multi-task machine learning and data mining system for attributional rule learning, rule testing, and application to a wide range of classification problems (Wojtusiak 2004). It was developed by the Machine Learning and Inference Laboratory (MLI) at the George Mason University [http://www.mli.gmu.edu/software]. The program has been recently extended to include features specific for processing biomedical data (Wojtusiak 2012).

AQ21 implements the Quasi-optimal (A^q) algorithm for constructing rules, and includes a number of features for handling multi-type data& ontologies, optimizing rules, and presenting results in transparent forms. AQ21 is a natural induction system that seeks patterns represented as attributional rules (Michalski 2004). The basic form of an attributional rule is: **CONSEQUENT <= PREMISE** where both CONSEQUENT and PREMISE are conjunctions of attributional conditions. Additionally AQ21 can learn rules with exceptions given by the formula **CONSEQUENT <= PREMISE |_ EXCEPTION**. Here, **EXCEPTION** can be either an attributional conjunctive description or a list of examples constituting exceptions to the rule. In the medical datasets, the exceptions are always negative examples such as recurrence and disease progression.

This paper aims at describing ontology-guided machine learning method that involves the use of knowledge embedded ontologies to effectively analyze the complex and heterogeneous biomedical data. The AQ21 is extended by adding ontologies (i.e., UMLS) that enables it to interpret the semantic meaning of data attributes. The purpose of the UMLS is to provide medical domain knowledge for ML methods to "understand" the meaning of biomedical data. The combination of the UMLS and the AQ21 provides an advanced computational and quantitative methodology to analyze biomedical data with the aid of the UMLS.

1.4 SEER-MHOS Dataset

This dataset links two large population-based data that provide detailed information about elderly persons with cancer (Clauser 2008). The SEER contains clinical, demographic and cause of death information for persons with cancer while the MHOS provides information about the health-related quality of life (HRQOL) of Medicare Advantage Organization (MAO) enrollees. In this paper, predictive models of QoL (especially for activities of daily living) are created for individual patients with different cancer diagnosis including prostate, breast, colorectal, lung and bronchus, uterus, bladder, head and neck, melanomas – skin, stomach, and pancreas.

2 METHODS

2.1 Source of Data

SEER-MHOS data was used to extract comorbidities and activities of daily living (ADLs) (self-reported), as well as cancer characteristics (SEER registry). The total number of patients in the SEER-MHOS is 1,849,311. First, patients with multiple cancers were excluded, then patients who completed surveys 3-years before and 2-years after the cancer diagnosis were extracted. If a patient had multiple surveys, one survey before and one survey after were used. The final dataset contains 4,583 cancer patients.

Dependent Variables: the primary outcomes were six ADLs (walking, dressing, bathing, moving in/out chair, toileting, and eating) after the cancer diagnosis.

Independent Variables: the potential predicators were selected based on the literature (Vissers 2013, Taneja 2013, Agborsangaya 2013, Amemiya 2007) as follows:

(1) Patient demographic such as age, race and marital status
(2) Six ADLs before cancer diagnosis
(3) Twelve comorbidities including Angina Pectoris/Coronary Artery Disease (ANGCAD), Arthritis of Hand/Wrist, Arthritis of Hip/Knee, Congestive heart failure (CHF), Emphysema/Asthma/Chronic obstructive pulmonary disease (COPD_E), Diabetes, Crohn's Disease/Ulcerative Colitis/Inflammatory Bowel Disease (GI_ETC), Hypertension, Myocardial Infarction (AMI), Other Heart Conditions, Sciatica, and Stroke,
(4) Six cancer characteristics such as grade, staging, tumor size, histology, tumor extension, and behavior
(5) Cancer radiation and surgery treatment indicators

2.2 Analyze the SEER-MHOS data with AQ21 and AQ21 extension

The rule-based machine learning system was applied to derive rules (or create models) from the SEER-MHOS dataset. We randomly divided the dataset into training (80%) and validation (20%) sets and used the training set to create prediction models and the validation set to assess model discrimination. Models were created in order to find the predictor or a set of predicators that can be used to predict the outcome (i.e., ADLs). Two machine learning methods were used to create ADL models: (1) AQ21 and (2) ontology-guided AQ21.

The ontology-guided AQ21 can optimize ADL rules by using UMLS knowledge (Min 2012). For example, a rule indicates that cancer diagnosis is a predictor for ADL impairment. The output of the AQ21 is as follows: ADL **<= [**Cancer Diagnosis = CUI1, CUI2], where CUI1 and CUI2 are cancer codes from the UMLS. The extended AQ21 can generalize the rule into their direct common parents (CUI3). Then the rule will be simply to ADL **<= [**Cancer Diagnosis = CUI3]. The program continuous to generalize rules until we find a negative data or block by medical knowledge. The advantage of the generalization include (1) Simplified the rules and 2. Cover more data in the model.

3 RESULTS

3.1 Patient Cohort

This retrospective SEER-MHOS study included 4,583 cancer patients. The average age was 74.6 (Standard deviation =6.76) years. The summary of our dataset is shown in Table 1. Table 2 shows the number of patients who reported ADL problems before and after cancer diagnosis. Cancer diagnosis has an impact on patient reported ADLs since the percentage of patients increased from 38.1 to 64.1 after cancer diagnosis. The walking was the most affected ADL among older cancer patients.

Table 1. Characters of Patients in the SEER-MHOS dataset (n=4,583)

	Number	%		Number	%
Age			Cancer Type		
<65	169	3.7	Bladder	310	6.8
65-74	2,269	49.5	Breast	1,018	22.2
75-84	1,814	39.6	Colorectal	659	14.4
>=85	311	7.2	Head Neck	134	2.9
Top 5 Comorbidities			Lung	542	11.8
Arthritis Hip	1,745	38.1	Melanoma	330	7.2
Sciatica	1,040	22.7	Pancreas	60	1.3
Other Heart	971	21.2	Prostate	1,311	28.6
Diabetes	953	20.8	Stomach	61	1.3
ANGCAD	654	14.3	Uterus	158	3.4

Table 2. Number of patients with ADL disabilities before and after cancer diagnosis

ADLs	No. of patients before Cancer Diagnosis	%	No. of patients after Cancer Diagnosis	%
Bathing	256	5.6	477	10.4
Dressing	200	4.4	380	8.3
Eating	87	1.9	217	4.7
Chairing	417	9.1	642	14.0
Walking	637	13.9	933	20.4
Toileting	150	3.3	290	6.3
Total	**1,747**	**38.1**	**2,939**	**64.1**

3.2 Rule Induction from the SEER-MHOS

Application of the AQ21 software to SEER-MHOS data mapped to UMLS resulted in a number of models (rulesets) for predicting patients' deficiencies in performing activities of daily living. The class-weighted accuracy of the models ranged between 57.1% and 69.19%. Similarly, class-weighted accuracy of models derived from the same data without the use of UMLS ranged from 54.18% to 57.10%. The class-weighted accuracy was calculated by the formula w = p/(p+n), p is number of positive examples and n is the number of negative examples (Wojtusiak 2004). On average, the class-weighted accuracy improvement was 2.44% when using UMLS.

The program derived multiple rules that constitute ADL prediction models. These rules are highly transparent and easy to understand. Below is an example rule that was part of model for predicting decline in ability to bathe independently.

[Bathing Impairment] <=
[Marital status = 1,4,5,3: 183, 585,23%]
[Stage = 2,7,9: 156, 472, 24%]
[Cancer site = C0346647, C0153490, C0153492, C0024623, C0153437, C0007102, C0153350, C0496779, C0153434, C0242787, C0949022, C0153550, C0153553, C0153555, C0153435, C0153491, C0153433, C0153436, C0024624: 175, 437, 28%]
: p = 58, n = 72, q = 0.262, cx = 67

The rule has three conditions (the overall rule needs to be interpreted as a conjunction of these conditions). The first one indicated that among the patients whose marital status were single (Code 1), divorced (Code 4), widowed (Code 5), or separated (Code 3), 185 of them have bathing disability while 585 patients have not. The consistency of this prediction is 23%. The second line showed that patients' staging condition was a predictor for bathing disability (Code 2, 7, 9 stands for Regional by Direct Extension, Distant Sites and/or Distant Nodes and Unknown respectively). There were 19 detailed cancer sites (UMLS CUIs) in this rule. The quality of this rule set is 26.2% and complexity is 67. For description of the parameters, please refer to AQ21 User's Guide (Wojtusiak 2004).

4 DISCUSSIONS

The ontology-guided AQ21 is highly configurable and robust system with features especially useful in learning from biomedical data such as individual patient data, learning from aggregated data, and using medical knowledge. One major advantage is that it can optimize attributional rules with the assistance of medical knowledge from the UMLS. For example, rule generalization based on the hierarchical relationships. The advantages of the generalization include: (1) simplified rules and (2) more covered data. The rule generalization procedure (travel along the path) continues until we found a negative data or data against medical knowledge. For example, radiation therapy can be applied to prostate cancer and lung cancer, but if the system generalizes to other types of cancer (i.e., leukemia) the therapy does not apply anymore.

One big challenge for the ontology-guided machine learning method is the performance issue. This issue comes from two sources: (1) UMLS as well as other medical ontologies are extremely large and complex, thus standard search methods cannot be applied and (2) size of the SEER-Medicare. In this paper, we limited our study to patients whose diagnosis was with only one cancer. The restriction reduced the sample size from 1,849,311 to 4,583. The majority of SEER-MHOS patients have more than one cancer diagnoses. The associations between multiple cancer diagnoses, comorbidities, and ADLs are complicated.

Although we worked on a small set of SEER-MHOS, the hierarchical structure from the UMLS was already large and complicated. For example, our SEER-MHOS data contains 572 distinct ICD-O codes in 4,583 individual patient records. All those codes are mapped to UMLS CUIs and 15,983 hierarchical records (path form those leaves to their roots) are extracted from the UMLS.

In the future, we will optimize machine learning algorithms, utilize supercomputers, and implement parallel processing.

5 CONCLUSIONS

In this paper, we apply an ontology-guided AQ21 method to promote the effectiveness and efficiency of machine learning in healthcare. The UMLS is used to provide the medical knowledge for the machine learning algorithms. It has been proved that the ontology-guided AQ21 can be applied to analyze the SEER-MHOS dataset.

ACKNOWLEDGMENTS

The study was supported by the Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee.

REFERENCES

Agborsangaya, C.B. et al. (2013). Health-related quality of life and healthcare utilization in multimorbidity: results of a cross-sectional survey. *Qual Life Res*, **22**(4): p. 791-9.

Amemiya, T., et al. (2007). Activities of daily living and quality of life of elderly patients after elective surgery for gastric and colorectal cancers. *Ann Surg*, 2007. **246**(2): p. 222-8.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*: 67-79

Clauser, S.B. and Haffer S.C. (2008). SEER-MHOS: a new federal collaboration on cancer outcomes research. *Health Care Financ Rev*, **29**(4): p. 1-4.

de Coronado, S., et al. (2004). NCI Thesaurus: using science-based terminology to integrate cancer research results. *Stud Health Technol Inform*, **107**(Pt 1): p. 33-7.

De Raedt, L. (2008). Logical and Relational Learning. Springer-Verlag.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*, **121**: p. 279-90.

Getoor, L. and Taskar B., eds. (2007). Introduction to Statistical Relational Learing, MIT Press.

Hoehndorf, R. et al. (2015). The role of ontologies in biological and biomedical research: a functional perspective. *Brief Bioinform*, **16**(6): p. 1069-80.

Lavrac N. and Dzeroski S. (1994). Inductive Logic Programming: Techniques and Applications. New York: Ellis Horwood.

Lindberg, C. (1990). The Unified Medical Language System (UMLS) of the National Library of Medicine. *J Am Med Rec Assoc*, **61**(5): p. 40-2.

Loy, P. (1978). International Classification of Diseases--9th revision. *Med Rec Health Care Inf J*, **19**(2): p. 390-6.

Madsen, M. (2010) Health care ontologies: knowledge models for record sharing and decision support. *Stud Health Technol Inform*, **151**: p. 104-14.Min, H. and J. Wojtusiak (2012). Clinical data analysis using ontology-guided rule learning. Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems. Maui, Hawaii, USA, ACM**:** 17-22.

Michalski, R.S. (2004). ATTRIBUTIONAL CALCULUS*:* A Logic and Representation Language for Natural Induction, in Reports of the MLI Laboratory MLI 04-2. George Mason University: Fairfax, VA.

Nelson, S.J., et al. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc*, **18**(4): p. 441-8.

Taneja, S.S. (2013) Re: impact of age and comorbidities on long-term survival of patients with high-risk prostate cancer treated with radical prostatectomy: a multi-institutional competing-risks analysis. *J Urol*, **189**(3): p. 901.

The Gene Ontology Consortium (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, **38**(Database issue): p. D331-5.

Vissers, P.A., et al. (2013) The impact of comorbidity on Health-Related Quality of Life among cancer survivors: analyses of data from the PROFILES registry. *J Cancer Surviv*, **7**(4): p. 602-13.

Warren, J.L., et al. (2002). Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med Care*. **40**(8 Suppl): p. IV-3-18.

Wojtusiak, J. (2004) AQ21 USER'S GUIDE, Reports of the Machine Learning and Inference Laboratory, MLI 04-3, George Mason University, Fairfax, VA

Wojtusiak, J., et al. (2006). The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and its Novel Features. in Tools with Artificial Intelligence, *ICTAI '06. 18th IEEE International Conference on*. 2006.

Wojtusiak, J., (2012) "Recent Advances in AQ21 Rule Learning System for Healthcare Data," American Medical Informatics Annual Symposium, Chicago, November.