Using Knowledge Sources to Improve Classification of Medical Text Reports

Adam Wilcox
Department of Medical Informatics
Columbia University
622 West 168th Street, VC-5
New York, NY 10032
adam.wilcox@columbia.edu

George Hripcsak
Department of Medical Informatics
Columbia University
622 West 168th Street, VC-5
New York, NY 10032
hripcsak@columbia.edu

Carol Friedman
Department of Medical Informatics
Columbia University
622 West 168th Street, VC-5
New York, NY 10032
carol.friedman@columbia.edu

ABSTRACT
Domain knowledge has been shown to be an important component of machine learning. However, the cost of obtaining domain knowledge to improve classifier generation can exceed the cost of manually creating classifiers. An alternative approach is to use existing knowledge sources to collect relevant domain knowledge, and improve machine learning. We investigated the use of two existing knowledge sources (a natural language processor and controlled vocabulary metathesaurus) to improve machine learning algorithm performance in building classifiers for medical text reports. Both knowledge sources were found to significantly improve classifier performance. This demonstrates that existing knowledge sources can easily be used to improve machine learning performance.

1. INTRODUCTION
An important consideration in data mining or machine learning is how domain knowledge can be used to improve the knowledge discovery process. Domain knowledge has been shown to improve algorithm performance, and is one of the most important components of successful data [1]. The cost of obtaining domain knowledge to improve learning algorithms is rarely considered, and can actually make inductive knowledge discovery from available data more expensive than knowledge extraction from available experts [2]. However, knowledge discovery may be improved by using available knowledge sources rather than extracting knowledge from experts. In this work, we investigated using available knowledge sources to improve feature selection in text mining. Specifically, we examined the effect of knowledge sources on the classification of medical text reports.

2. BACKGROUND
Valuable clinical data are collected as narrative text. To be useful to automated systems, the information contained in these narrative reports must be converted to a structured form [3]. Typically, the form is in standardized codes, representing not only general information directly retrieved from the narrative text, but complex conclusions drawn from it.

Inference methods, such as decision rules that classify text reports, can be used to convert the narrative text data into meaningful clinical interpretations [3]. But manual inference rule development is a difficult and time-intensive task, and thus expensive [4]. Researchers have investigated the use of inductive learning algorithms to automatically generate classifiers for medical documents. However, this research has involved extensive use of domain knowledge, which can be as difficult and expensive to collect as inference rules [5].

Some domain knowledge that may be useful to classifiers is available in the form of existing knowledge sources. The knowledge from these sources can be used to improve classifier performance at minimal incremental cost. The goal of this research is to demonstrate the effect of using knowledge from existing sources to improve the performance of medical text report classifiers.

3. METHODS
We examined the effect of two knowledge sources on the performance of medical text classifiers created by machine learning algorithms. These knowledge sources were a general-purpose medical language processor, and a thesaurus of controlled medical vocabularies. The classifiers were created to classify chest radiograph reports according to six clinical conditions as indicated in the reports. In this section, we describe the different components of this evaluation.

The Medical Language Extraction and Encoding System (MedLEE) was developed at Columbia University, and is in real use at New York Presbyterian Hospital [6]. The processor attempts to encode all clinical information available in reports. It converts the narrative text to a set of observations or findings, each with associated descriptive modifiers and values. MedLEE as a knowledge source can be used to convert raw text to a more structured form.

The thesaurus studied was the Unified Medical Language System (UMLS) Metathesaurus, developed and distributed by the National Library of Medicine [7]. The UMLS Metathesaurus contains information about biomedical terms from many controlled vocabularies, with relationships between terms from different vocabularies. The organization of the Metathesaurus is by concept or meaning, and alternate names for the same concept (such as synonyms and lexical variants) are linked together. While the Metathesaurus represents many relationships between

different concepts, the synonymy relationship was the only one used in this research.

We tested the ability of machine learning algorithms to determine the presence of six clinical conditions in chest radiograph reports: congestive heart failure, chronic obstructive pulmonary disease, acute bacterial pneumonia, neoplasm, pleural effusion without congestive heart failure, and pneumothorax. The performance of the learning algorithms was compared to a set of classification rules written by a domain expert. The reports were also manually read and classified by a set of physicians. The data consisted of 200 chest radiograph reports.

Five different machine learning algorithms, representing different approaches to inductive learning, were used to generate classifiers: MC4, CN2, naïve-Bayes, IB, and decision tables. The algorithms were interfaced by the MLC++ machine learning library [8].

We studied the effect of each knowledge source (UMLS, NLP) on machine learning performance. There were four different report structures studied with machine learning algorithms: raw text, raw text with UMLS, NLP output, NLP output with UMLS. With raw text, reports were converted to document vectors of all words in the document collection, with a binary value indicating their presence in the specific report. With NLP, the document vectors were created from the parsed observations present in the reports. The UMLS was used to convert synonymous concepts in the raw text or NLP observations to UMLS identifying codes prior to the creation of the document vectors.

For training and test sets, we used the 200 physician-classified reports, with leave-one-out cross-validation to eliminate bias associated with training on the test set. For each report structure and each algorithm, we computed an average ROC area for the six clinical conditions. This value was averaged across the five learning algorithms to create a single performance value for each report type.

4. RESULTS

Figure 1 shows the classifier performance for each report structure, in terms of ROC area. The performance of expert rules and physicians is also shown. Both knowledge sources (UMLS and NLP) significantly improved performance from that using raw text ($p < 0.001$). In addition, when these knowledge sources were combined, there was significant improvement over methods using only one knowledge source ($p < 0.02$). No method using machine learning performed as well as expert-written rules or physicians, regardless of knowledge sources used.

5. CONCLUSIONS

This study demonstrated the feasibility of using available knowledge sources to improve knowledge discovery from text data. Two available resources both significantly improved classifiers built by machine learning algorithms. The freely-available UMLS Metathesaurus and the medical language processor MedLEE both significantly improved classifier performance when used separately, and improved performance even more when used in conjunction. Improvement was

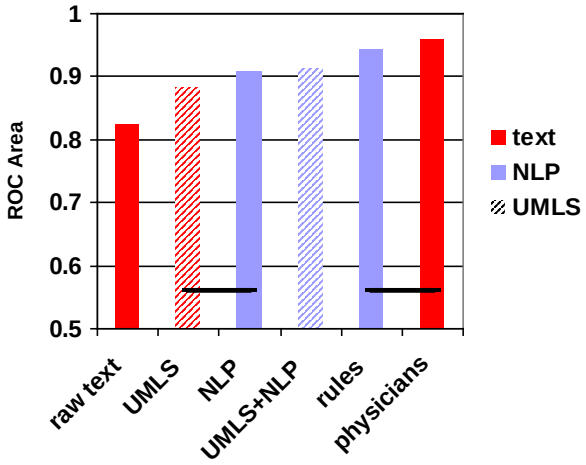

Figure 1: Machine learning algorithm performance for different report structures. Black bars connecting columns indicate those for which no significant difference was detected.

demonstrated in classifiers that detected indications of clinical conditions from radiograph reports. This result demonstrates that existing knowledge sources can be used automatically to improve knowledge discovery.

6. REFERENCES

[1] Pazzani M, Kibler D. The utility of knowledge in inductive learning. Machine Learning 1992; 9:57-94.

[2] Wilcox A. Automated Classification of Medical Text Reports [dissertation]. Columbia University, Department of Medical Informatics; 2000.

[3] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995; 122(9):681-8.

[4] Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res 1993 Oct; 26(5):467-81.

[5] Wilcox A, Hripcsak G. Medical text representations for inductive learning. Accepted to 2000 AMIA Fall Symposium.

[6] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994; 1(2):161-74.

[7] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 Aug; 32(4):281-91.

[8] Kohavi R, Sommerfield D, Dougherty J. Data mining using MLC++: a machine learning library in C++. In: Tools with Artificial Intelligence; IEEE Computer Society Press; 1996. p. 234-45.