Journal of Biomedical Informatics 58 (2015) S92–S102

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases

Christopher Kotfila [a,*], Özlem Uzuner [b]

[a] Informatics Department, University at Albany, State University of New York, Albany, NY, USA
[b] Department of Information Studies, University at Albany, State University of New York, NY, USA

ARTICLE INFO

Article history:
Received 26 March 2015
Revised 20 July 2015
Accepted 22 July 2015
Available online 1 August 2015

Keywords:
Phenotyping
Classification
Natural language processing

ABSTRACT

Automated phenotype identification plays a critical role in cohort selection and bioinformatics data mining. Natural Language Processing (NLP)-informed classification techniques can robustly identify phenotypes in unstructured medical notes. In this paper, we systematically assess the effect of naive, lexically normalized, and semantic feature spaces on classifier performance for obesity, atherosclerotic cardiovascular disease (CAD), hyperlipidemia, hypertension, and diabetes. We train support vector machines (SVMs) using individual feature spaces as well as combinations of these feature spaces on two small training corpora (730 and 790 documents) and a combined (1520 documents) training corpus. We assess the importance of feature spaces and training data size on SVM model performance. We show that inclusion of semantically-informed features does not statistically improve performance for these models. The addition of training data has weak effects of mixed statistical significance across disease classes suggesting larger corpora are not necessary to achieve relatively high performance with these models.

1. Introduction

With the proliferation of electronic health records (EHR) in recent years, automated phenotyping for cohort selection has become an area of growing interest for the biomedical informatics community [31]. Despite a wide array of focused work, many challenges still persist for delivering practical phenotyping technologies including high-throughput generalized algorithms that are applicable across different diseases without the need for local- or domain-specific rules [20]. Commonly structured EHR-related information such as ICD-9 codes have been shown to be insufficient for producing state-of-the-art performance [23] which is why many phenotyping systems employ semantically-informed Natural Language Processing (NLP) methodologies to unlock the unstructured data contained in clinical narratives [14]. Currently, 34 out of 47 published phenotyping algorithms on the eMERGE [28] Phenotype KnowledgBase include NLP components (https://phekb.org/ Accessed: February 13th 2015).

1.1. Problem definition

In this paper, we systematically assess the effect of feature spaces, feature weights, and support vector machine (SVM) kernels on model performance on a phenotyping task as the training data is roughly doubled. To do this, we cast the document-level multi-label classification task set out in the 2014 i2b2/UTHealth shared task as a series of five document-level binary classification tasks (one per disease) consistent with the phenotype identification literature. Concretely, given a document from our test set, our goal is to identify the presence (or not) of five different diseases in each patient for each medical record. The five diseases we identify are: obesity, atherosclerotic cardiovascular disease (CAD), hypertension, hyperlipidemia, and diabetes. Using this task as our test bed, we assess the effect of minimally normalized, lexically normalized, and semantically-informed feature spaces on phenotype identification, with various weighting schemes, kernels, and as the training data is doubled. Our primary motivation is not to create the highest performing system possible but to implement reasonable systems and assess the impact of common feature spaces, feature weights, kernels and their combinations on overall system performance. Because of the high cost of annotating data for supervised NLP and machine learning tasks, we also investigate the effect of doubling training data on model performance. To do this, we exploit overlapping annotations from the 2008 i2b2 Obesity Challenge shared task [37] and the 2014 i2b2/UTHealth shared task. We evaluate a broad array of models on the 2014 test corpus using the 2008 and 2014 training corpora, and a combined 2008/2014 training corpus.

* Corresponding author at: Informatics Department, University at Albany, 1400 Washington Ave., Albany, NY 12222, USA. Tel.: +1 (518) 526 8964.
E-mail address: ckotfila@albany.edu (C. Kotfila).

http://dx.doi.org/10.1016/j.jbi.2015.07.016
1532-0464/

2. Related work

NLP-informed machine learning algorithms have been shown to be successful in identifying patients with rheumatoid arthritis [6,10], diabetes [40], colorectal cancer, and venous thromboembolism [10], in risk adjustment for ICU patients [27] and for smoking history detection [22]. In several instances, these methods have been successfully ported across institutions, demonstrating the robustness of the NLP-informed machine learning approach to patient phenotyping [7,40]. A broad array of tools, techniques and ontologies have been developed for incorporating biomedically-relevant semantic information into machine learning techniques.

One approach for semantically-informed machine learning-based phenotype identification leverages the fixed vocabulary of the Unified Medical Language System's [4,24] (UMLS) concept unique identifiers (CUIs). The UMLS is a metathesaurus that knits together a wide array of medical vocabularies and provides, among other things, lexical and conceptual crosswalks between constituent terminologies. Many machine-learning approaches to phenotype identification preprocess patient clinical narratives to extract CUIs for use as features. These CUIs are then used, either alone or in concert with other structured EHR information (e.g., ICD-9 codes), for predicting patient membership in a particular phenotype [3,10,39]. The process of identifying medical concepts and resolving them to a fixed vocabulary from arbitrary text is not a trivial problem [5]. Luckily, several mature tools exist for expediting the process. MetaMap (formerly MMtx) is a commonly used tool for extracting medical concepts from free-form text and mapping them to the controlled vocabulary of UMLS CUIs [1]. It has a proven track record for high-throughput indexing of medical documents based on semantic content [2].

Machine learning techniques to NLP often involve complex pipelines that include normalization, tokenization, sentence breaking, stopping, stemming, word sense disambiguation, part of speech tagging, and information extraction [26]. Overall system performance can be attributed to many different steps in that pipeline and the effect of a particular implementation choice on system performance is often unclear. From an engineering perspective this can be critically important. Not all steps in the pipeline are equally easy to implement or maintain in a production environment [33]. A systematic understanding of the tradeoffs in performance associated with individual implementation decisions can lead to better overall system design and user acceptance [13].

Feature extraction and selection is one area of pipeline design that is critically important to system success. For example, Bejan et al. [3] used a binary classifier for identifying pneumonia; comparing word n-grams, UMLS concepts and assertion values associated with pneumonia expressions. Using clinical notes from 426 patients, they showed statistical feature selection had a substantial improvement over a baseline system that used the complete set of features. Carroll et al. [6] saw significant improvement using a SVM over a rule-based system using an expert-defined feature set for phenotype identification of rheumatoid arthritis. Carroll et al. argued that with a curated feature set it should be possible to achieve state-of-the-art performance using 50–100 annotated documents. Using simple bag-of-words features, Wright et al. [40] employed SVMs to identify diabetes across different institutions using 2000 progress notes (1000 from each institution) achieving $F1$ measures of 0.934 and 0.935, respectively. They found that stop word filtering, feature selection, negation extraction, and named entity recognition did not substantially improve performance over a bag of words.

Training data size continues to be a key motivating factor in system development [22]. Annotation of medical data can be difficult and costly especially if private health care information must first be identified and removed. As a result, a wide and inconsistent array of training data sizes are reported in the literature and it is not always clear how increasing or decreasing training data size will affect reported model performance.

In general, literature has shown that machine learning algorithms with simple feature spaces and relatively straightforward applications of biomedical NLP tools for semantic feature extraction can perform well on particular phenotypes, and on records sourced from different hospitals. It is unclear how overall performance of these methods is affected by training corpus size, feature spaces, feature weighting schema, and SVM kernel choices. This paper addresses this gap by systematically comparing the effect of a number of well-established feature spaces and feature weighting schemes on classifier performance with various kernels as training data is roughly doubled.

3. Background

One of the purposes of the 2014 i2b2/UTHealth shared task was to create a NLP challenge that represented a culmination of the previous shared tasks [35]. Like previous i2b2 shared tasks, the 2014 challenge includes smoking history, identification of obesity and a selection of its comorbidities, medication identification, and temporal classification of medical events. By design, certain aspects of the 2014 i2b2/UTHealth shared task contain highly similar annotations with previous i2b2 shared tasks. For instance, the 2008 i2b2 Obesity Challenge includes annotations for diseases that overlap with the 2014 i2b2/UTHealth annotations.

In 2008, 30 teams participated in the Obesity Challenge which required teams to develop systems for identifying presence or absence in a patient of obesity and fifteen comorbidities based on information from unstructured narratives of medical discharge summaries. The Obesity Challenge task was defined by two experts who studied 50 pilot discharge summaries from the Partners HealthCare Research Patient Data Repository. The experts identified fifteen frequently-occurring comorbidities including CAD, diabetes mellitus, hypercholesterolemia, and hypertension. Obesity Challenge systems were required to make *textual* and *intuitive* judgments for each document and each disease. Textual judgements were based on direct references to the diseases in the discharge summary. Intuitive judgements were based on some amount of reasoning on the part of the expert annotators. Textual judgements for each disease fell into four classes: present, absent, questionable, or unmentioned. Intuitive judgements fell into three classes: present, absent, and questionable. For example, the statement "*the patient weighs 230 lbs and is 5 ft 2 inches*" would lead to a textual judgment of 'unmentioned' for obesity and an intuitive judgement of 'present' (i.e., obesity is not directly mentioned but can be inferred from the height and weight measurements). Intuitive judgements were primarily intended for the interpretation of textual judgements that fell into the unmentioned category.

In 2014, 27 teams participated in the i2b2/UTHealth shared task which required teams to develop systems for identifying diseases, medications, family history of CAD, and smoking status across a temporally-ordered series of unstructured medical notes for individual patients. Unlike the 2008 Obesity Challenge, disease annotations in the 2014 shared task marked only positive (e.g., present) instances of each disease. Neither directly-negated textual evidence nor inferred absence of a disease was marked. In addition to the presence of a disease, a temporal and indicator component were included with all 2014 disease annotations. The time component had acceptable values of "before document creation time" (before DCT), "during document creation time" (during DCT) and