Journal of Biomedical Informatics 69 (2017) 259–266

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository

Liqin Wang [a,b,*], Peter J. Haug [a,b], Guilherme Del Fiol [a]

[a] Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA
[b] Homer Warner Research Center, Intermountain Healthcare, 5121 South Cottonwood Street, Murray, UT 84107, USA

ARTICLE INFO

Article history:
Received 2 January 2017
Revised 3 April 2017
Accepted 19 April 2017
Available online 20 April 2017

Keywords:
Machine learning
Classification
Biomedical literature
Clinical data repository
Ontology
Disease-specific vocabulary

ABSTRACT

Objective: Mining disease-specific associations from existing knowledge resources can be useful for building disease-specific ontologies and supporting knowledge-based applications. Many association mining techniques have been exploited. However, the challenge remains when those extracted associations contained much noise. It is unreliable to determine the relevance of the association by simply setting up arbitrary cut-off points on multiple scores of relevance; and it would be expensive to ask human experts to manually review a large number of associations. We propose that machine-learning-based classification can be used to separate the signal from the noise, and to provide a feasible approach to create and maintain disease-specific vocabularies.

Method: We initially focused on disease-medication associations for the purpose of simplicity. For a disease of interest, we extracted potentially treatment-related drug concepts from biomedical literature citations and from a local clinical data repository. Each concept was associated with multiple measures of relevance (i.e., features) such as frequency of occurrence. For the machine purpose of learning, we formed nine datasets for three diseases with each disease having two single-source datasets and one from the combination of previous two datasets. All the datasets were labeled using existing reference standards. Thereafter, we conducted two experiments: (1) to test if adding features from the clinical data repository would improve the performance of classification achieved using features from the biomedical literature only, and (2) to determine if classifier(s) trained with known medication-disease data sets would be generalizable to new disease(s).

Results: Simple logistic regression and LogitBoost were two classifiers identified as the preferred models separately for the biomedical-literature datasets and combined datasets. The performance of the classification using combined features provided significant improvement beyond that using biomedical-literature features alone (p-value < 0.001). The performance of the classifier built from known diseases to predict associated concepts for new diseases showed no significant difference from the performance of the classifier built and tested using the new disease's dataset.

Conclusion: It is feasible to use classification approaches to automatically predict the relevance of a concept to a disease of interest. It is useful to combine features from disparate sources for the task of classification. Classifiers built from known diseases were generalizable to new diseases.

1. Introduction

The biomedical literature and electronic medical records offer great opportunities for acquiring disease-specific medical knowledge. Automated extraction of disease-*concept* associations from these knowledge sources can speed the process of building disease-specific concept vocabularies which could be further used for various clinical applications, such as automated annotation of biomedical text [1,2], identification of disease cohorts [3], and development of diagnostic models [4]. In the present study, we propose an approach for automated extraction of disease-*medication* associations from the biomedical literature and a clinical data repository (CDR). The approach uses machine learning classification models to predict the relevance of concepts to the disease of interest. The approach is developed based on former studies [5–8]; and it overcomes a common challenge faced in these studies, which is to use the metrics of relevance of the disease-*concept* associations to effectively decrease the manual effort necessary

* Corresponding author at: Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Suite 140, Salt Lake City, UT 84108, USA.
E-mail address: liqin.wang@utah.edu (L. Wang).

http://dx.doi.org/10.1016/j.jbi.2017.04.014
1532-0464/

to review noisy collections of associations in order to build disease-specific concept vocabularies. To build classification models, we evaluated the proposition that combining features derived from a clinical data repository with those from the biomedical literature would result in better performance than using features from a single source. We also conducted an exploratory assessment of the model's generalizability in predicting the disease-*concept* associations extracted for other diseases.

2. Background and significance

Dozens of studies have investigated techniques for extracting disease-*concept* associations from the biomedical literature and electronic medical records. The *concepts* studied have included associated genes [9], signs and symptoms [10], findings [11], medications [7,8], and lab tests [7]. Numerous knowledge acquisition techniques have been proposed to extract relational information, including co-occurrence-based statistics [7,8,11], natural language processing (NLP) [12,13], graph theory [9,14], and others [15,16]. Zeng and Cimino retrieved disease-chemical relationships from the UMLS co-occurrence table (MRCOC) simply based on the co-occurrence of MeSH terms assigned to published articles [17]. Cao et al. used NLP and co-occurrence statistics (i.e., chi-square statistics and the proportion confidence interval) to extract disease-finding associations [11]. Chen et al. applied similar techniques to extract disease-drug pairs from PubMed® citations and clinical documents [8]. In those studies, NLP techniques had been used mainly for named entity recognition when the sources of the data were in "free-text" form. In addition, Rindflesch et al. developed a rule-based system called SemRep that extracts the semantic relations between the concepts identified in a particular sentence in the biomedical literature [12,18]. For example, given the sentence "a randomized trial of etanercept as monotherapy for psoriasis", a semantic predication was generated: *etanercept TREATS psoriasis*. Bundschus et al. explored using conditional random fields to identify the semantic relations between disease and medications and between disease and genes in biomedical text [15]. Xu and Wang used a pattern-learning approach to extract disease-drug and disease-disease risk pairs from biomedical abstracts [16,19]. In addition, the authors of the present study have developed a pipeline-based system which combines multiple techniques (i.e., document retrieval, SemRep, UMLS semantic network, and co-occurrence-based statistics) to extract disease-specific treatments (including medications, surgical procedures, medical devices, and activities) from biomedical titles and abstracts [6]. More details about this work can be found in section 3.1.

Existing statistically-based automated extraction techniques score the disease-*concept* candidate set allowing some reduction in noise, but leaving behind a large number of "bad" concept-disease pairs. The precision can be very low when focusing on high recall. For example, in a previous study, when counting all retrieved treatment concepts, we achieved a precision of less than 0.3 on two test diseases when comparing to manually-created reference vocabularies [6]. The challenge escalates when facing hundreds or thousands of concepts extracted for each disease in light of low precision. Ultimately, filtering out false-positives requires manual expert review, which is costly and time-intensive.

Disease-*concept* associations extracted by automated techniques have been assigned statistical scores, such as frequency of occurrence, which may provide some sort of indication for the strength of the relationship between the disease of interest and extracted *concepts*. Researchers previously investigated potential approaches to set proper thresholds based upon those statistical scores to identify a subset of important associations for further investigation. For example, Cao et al. explored using the volume test of Diaconis and Efron to identify thresholds using the chi-square score [20]. However, choosing cut-off points on these statistical scores is either empirical or arbitrary, and it would not generally apply well to a situation where extracted concepts are assigned multiple scores.

To determine the relevance of extracted concepts to the disease of interest is a binary classification issue. To address the above challenge, machine-learning-based classification techniques can possibly be used to predict the relevance of extracted disease-*concept* associations based upon the multiple statistical scores. This would eliminate a significant number of irrelevant concepts and keep a subset of "interesting" concepts for further investigation.

To develop an appropriate classification model, we considered two important questions: (1) what features should be used to build the model; and (2) how generalizable is the model?

Disease-specific associations could be extracted from different sources by multiple techniques, which generate different kinds of measures of relevance (i.e., features). For example, in a prior study, we used four scoring strategies (i.e., frequency of occurrence, interest, degree centrality, and weighted degree centrality) to extract disease-treatment associations from the biomedical literature [6]. Wright et al. applied five co-occurrence-based statistics (i.e., support, confidence, chi square, interest, and conviction) to extract disease-medication and disease-lab test associations from the electronic medical records [7]. Studies have shown that combining the results of extraction by different techniques/queries from a single source led to progressively improving retrieval performance [21–23]. Other studies also show that the results of extraction from the different sources are somewhat complementary [5,8]. With these findings in mind, we assumed that by combining the measures of relevance generated by different techniques from different sources (i.e., the biomedical literature and a CDR) as features within a classification system, the performance of the classifiers may be improved compared to using a single feature or features from a single source.

The generalizability of the classification model is important because it is difficult and expensive to build a classifier for each disease. However, for different diseases, the range and distribution of the values of the relevance measures may be different. This could affect the performance of a classifier when trained and tested on different disease datasets. We measure the generalizability of the classifier by determining if a classifier trained and tested on a different disease's datasets achieved as good performance as the classifier trained and tested on the same disease's dataset.

The ultimate goal of this study is to develop machine learning classifiers that could reduce the manual effort necessary to review noisy collections of disease-specific concepts. To achieve this goal, in the present study, we initially focused on disease-medication associations, and searched for classification models appropriate to predict the relevance of groups of medications to a specific disease. The models were designed to incorporate multiple statistical scores. We assessed two research questions that (1) would adding the features from the CDR improve the performance of models that used features from biomedical literature only; (2) would models built from known disease-medication associations be effective in predicting disease-medication associations for new diseases.

3. Materials and methods

The study methods consisted of the following steps (see Fig. 1): (1) extraction of disease-specific medications from the biomedical literature; (2) extraction of disease-specific medications from a local CDR; (3) preparation of datasets for classification, including merging the datasets from the disparate sources and validating disease-medication associations using reference standards;