Journal of Biomedical Informatics 72 (2017) 23–32

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Automatic prediction of coronary artery disease from clinical narratives

Kevin Buchan [a,*], Michele Filannino [b], Özlem Uzuner [b]

[a] *Department of Information Science, State University of New York at Albany, NY, USA*
[b] *Department of Computer Science, State University of New York at Albany, NY, USA*

ARTICLE INFO

Article history:
Received 14 December 2016
Revised 19 June 2017
Accepted 22 June 2017
Available online 27 June 2017

Keywords:
Coronary artery disease
Prediction
Natural language processing
Machine learning
Dimensionality reduction
Ontology

ABSTRACT

Coronary Artery Disease (CAD) is not only the most common form of heart disease, but also the leading cause of death in both men and women (Coronary Artery Disease: MedlinePlus, 2015). We present a system that is able to automatically predict whether patients develop coronary artery disease based on their narrative medical histories, i.e., clinical free text. Although the free text in medical records has been used in several studies for identifying risk factors of coronary artery disease, to the best of our knowledge our work marks the first attempt at automatically predicting development of CAD. We tackle this task on a small corpus of diabetic patients. The size of this corpus makes it important to limit the number of features in order to avoid overfitting. We propose an ontology-guided approach to feature extraction, and compare it with two classic feature selection techniques. Our system achieves state-of-the-art performance of 77.4% F1 score.

1. Introduction

Coronary Artery Disease (CAD) is not only the most common form of heart disease, but also the leading cause of death in both men and women [1]. The second track of the 2014 i2b2/UTHealth challenge targeted the automatic identification of risk factors for CAD: a complex clinical NLP task which could benefit from concept extraction, assertion classification, diagnosis extraction, medication extraction, smoking history, and family history [2].

Free text is considered to be a rich source of information for purposes of health care operations and research [3]. Furthermore, there have been several recent studies demonstrating the effectiveness of natural language processing (NLP) and machine learning methods for disease detection using clinical free text, which are discussed as related works in this paper. Thus, our aim is to develop a model for automatically predicting *development of CAD* from clinical free text.

We study the 2014 i2b2 Heart Disease Risk Factors Challenge Data [24] from a different perspective. Our purpose is to develop a system that automatically predicts patients who develop CAD based on their narrative medical histories *before a diagnosis of CAD*. For this purpose, we examine common risk factors for CAD—these risk factors consist of many of the same known risk factors for type-2 diabetes. They include high cholesterol, high-blood pressure, obesity, lack of physical activity, unhealthy diet, and stress [25]. As all patients in the corpus have diabetes, they are all at high risk for CAD and carry a lot of the same overall risk factors. This makes it challenging to separate the patients who actually develop the disease from those who do not. Additionally, solving this task on a small corpus requires special attention to overfitting. Our hypothesis is that it is possible to predict whether patients will develop CAD using a domain ontology to reduce the high dimensional nature of free text medical records.

Our approach to CAD prediction is unique in that we examine unstructured data (i.e., clinical free text in patients' electronic medical records (EMRs)) to predict which patients will develop a CAD diagnosis in the future. This is a natural language processing (NLP) and machine learning task. We believe that our system can complement, supplement and even provide a second perspective to existing CAD models that use only non-textual, structured data for predicting the disease [23]. As part of the original 2014 i2b2/UTHealth challenge, several teams developed systems with the goal of identifying *risk factors* for heart disease [4–22]. However, to the best of our knowledge our work marks the first attempt at automatically predicting *development of CAD* using *free text* in medical records.

We approach the CAD prediction task as a document classification problem. This means that we treat each record as one sample, independent of any previous or future sample (i.e., we disregard the longitudinal nature of the data). We simply classify if given one patient record at a discrete point in time that patient will eventually develop (or not develop) a CAD diagnosis. To improve classifier performance, we propose an ontology-guided approach to

[*] Corresponding author.
E-mail address: kbuchan@albany.edu (K. Buchan).

http://dx.doi.org/10.1016/j.jbi.2017.06.019
1532-0464/© 2017 The Authors. Published by Elsevier Inc.

feature extraction and compare this with two standard feature selection techniques. Specifically, our novel feature extraction technique automatically filters out features based on domain knowledge in the Unified Medical Language System (UMLS).

The clinical application of our model is in classification of patients who will develop CAD in difficult-to-discriminate situations; e.g., when patients are all at high risk for CAD and carry many of the risk factors.

1.1. Related works

There is a well established volume of research outlining NLP and machine learning methods for disease classification in clinical free text. Pineda et al. applied the pipeline-based NLP tool Topaz to extract 31 UMLS concept unique identifier (CUI) features for classification of influenza in emergency department free-text reports [26]. The team compared seven different classifiers to an expert-built Bayesian classifier and achieved a 93% F1 score.

Similar methods have been applied to detect thromboembolic disease in free-text radiology reports [27]. Specifically, Pham et al. developed a system that pre-processed documents using a simple sentence segmenter and tokenizer. They created a lexicon to define concept types, which were incorporated into the feature space along with filtered unigrams and bigrams. They then experimented with Weka to train Support Vector Machine (SVM) and Maximum Entropy (MaxEnt) classifiers, of which the MaxEnt classifier achieved the highest F1 score of 98%.

Furthermore, Redd et al. developed a set of retrieval criteria for identifying patients at risk for scleroderma renal crisis in electronic medical records [28]. The team developed their NLP system using data from the Veterans Informatics and Computing Infrastructure (VINCI). Their concept extraction criteria included specific disease and symptom mentions related to systemic sclerosis (SSc). The group then trained an SVM classifier to detect documents that indicated a diagnosis of SSc and reported an F1 score of 87.3%.

Several teams have experimented with domain ontologies to guide feature extraction for text classification. Wang et al., e.g., established a concept hierarchy by mapping raw terms to medical concepts using the UMLS, which they then searched to obtain the optimal concept set [29]. This feature selection technique improved the overall accuracy of their text classification system as compared with Principal Component Analysis (PCA) for dimensionality reduction.

Additionally, Garla and Brand exploited the UMLS ontology during feature engineering to improve the performance of machine-learning-based classifiers trained on the 2008 i2b2 Obesity Challenge Data Set [30]. This data set includes 15 diseases, including CAD, and its classification based on one narrative record per patient. To enhance feature ranking for this task, Garla and Brand propagated contingency tables of concepts in UMLS to their hypernyms, which they refer to as the propagated information gain. They then assigned each concept the highest propagated information of any hypernym. The use of this technique yielded the greatest performance improvement for their system, however, it did not improve performance on the classification of CAD.

For predicting CAD before it develops, we experimented with Naive Bayes, SVM, and MaxEnt classifiers and tested dimensionality reduction techniques including PCA, mutual information, and domain ontology-guided feature extraction. Our hypothesis is that the medical concepts most relevant to predicting CAD have formally defined relationships as such in the UMLS Semantic Network that can be exploited to automatically predict the disease. By engineering features around these concepts, as opposed to constructing features for every possible concept in our documents, we focus our feature space on information that really matters for our task. The reduction in the feature, in turn, results in simpler and more robust models, that run without any significant loss of performance.

2. Data

The 2014 i2b2 Heart Disease Risk Factors Challenge data set consists of 1304 longitudinal records of a total of 296 diabetic patients. Each patient in the corpus belongs to one of three cohorts:

1. patients who had a CAD diagnosis in the first record of their patient profile
2. patients who developed a CAD diagnosis sometime later in their patient profile
3. patients who did not develop a CAD diagnosis

The criteria for classifying CAD and no-CAD patients in our study has been defined and validated in two earlier studies [31,3]. To create the corpus for the 2014 i2b2 Heart Disease Risk Factors Challenge, an expert cardiologist developed the definition for CAD. Specifically, the following search criteria were used against Partners HealthCare Electronic Medical Records (EMR) [31]:

- at least 3 CAD codes or 1 procedure code for a coronary revascularization
- at least 4 codified mentions of beta-adrenergic inhibitor medications
- at least 4 codified mentions of anti-platelet agents (such as aspirin)
- at least 4 codified mentions of statins (cholesterol lowering drugs)

For the purposes of our study, we focused on prediction of CAD before the patients were officially diagnosed, i.e., they had an annotated CAD diagnosis in the i2b2/UTHealth data. We therefore discarded from the data any records with a CAD diagnosis. This removed all patients who were diagnosed with CAD at the onset of their patient profile. Additionally, for patients who later received a CAD diagnosis, records were discarded beginning with the one in which the patient received the diagnosis. This left us with the records of the patients who did not develop CAD (referred to as no-CAD patients), and the records from those who do develop CAD before their diagnosis (referred to as CAD patients).

After discarding all records with diagnosis of CAD, we checked our CAD and no-CAD patients with respect to their level of sickness. To achieve this, we calculated normalized frequencies of the number of symptoms, diseases and medications extracted by cTAKES in each record and divided by the length of the document (i.e., the number of word tokens per record), as shown in Eq. (1). Intuitively, this calculation measures the disease density of the record.

Eq. (1) – Normalized frequency of sickness in patient records.

$$\frac{\textit{number of diseases} + \textit{number of symptoms} + \textit{number of medications}}{\textit{number of word tokens}} \tag{1}$$

We found that the no-CAD patients were depicted as being sicker in their records than the CAD patients as described by their pre-CAD diagnosis records. We decided to control this factor so that we could make the machine learning models agnostic with respect to the level of sickness in the two populations. Thus, we matched a random subsample of records in the no-CAD patient population to levels of sickness in CAD patients (see Appendix A Fig. A.1).