Artificial Intelligence in Medicine 69 (2016) 53–61

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer

Mark Hoogendoorn[a,b,*], Peter Szolovits[b], Leon M.G. Moons[c], Mattijs E. Numans[d]

[a] Department of Computer Science, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
[b] Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, MA 02139, USA
[c] Department of Gastroenterology and Hepatology, Utrecht University Medical Center, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
[d] Department of Public Health and Primary Care, Leiden University Medical Center, Hippocratespad 21, 2333 ZD Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 6 November 2015
Accepted 23 March 2016

Keywords:
Natural language processing
Predictive modeling
Uncoded consultation notes
Colorectal cancer

ABSTRACT

Objective: Machine learning techniques can be used to extract predictive models for diseases from electronic medical records (EMRs). However, the nature of EMRs makes it difficult to apply off-the-shelf machine learning techniques while still exploiting the rich content of the EMRs. In this paper, we explore the usage of a range of natural language processing (NLP) techniques to extract valuable predictors from uncoded consultation notes and study whether they can help to improve predictive performance.
Methods: We study a number of existing techniques for the extraction of predictors from the consultation notes, namely a bag of words based approach and topic modeling. In addition, we develop a dedicated technique to match the uncoded consultation notes with a medical ontology. We apply these techniques as an extension to an existing pipeline to extract predictors from EMRs. We evaluate them in the context of predictive modeling for colorectal cancer (CRC), a disease known to be difficult to diagnose before performing an endoscopy.
Results: Our results show that we are able to extract useful information from the consultation notes. The predictive performance of the ontology-based extraction method moves significantly beyond the benchmark of age and gender alone (area under the receiver operating characteristic curve (AUC) of 0.870 versus 0.831). We also observe more accurate predictive models by adding features derived from processing the consultation notes compared to solely using coded data (AUC of 0.896 versus 0.882) although the difference is not significant. The extracted features from the notes are shown be equally predictive (i.e. there is no significant difference in performance) compared to the coded data of the consultations.
Conclusion: It is possible to extract useful predictors from uncoded consultation notes that improve predictive performance. Techniques linking text to concepts in medical ontologies to derive these predictors are shown to perform best for predicting CRC in our EMR dataset.

1. Introduction

Electronic medical records (EMRs) are a valuable resource in the development of predictive models for diseases. The increasing level of integration of information from different caretakers into single EMR systems increases the possibilities even more. The step from an EMR to a predictive model is, however, far from trivial and requires dedicated processing techniques for a number of reasons: First of all, EMRs are typically ambiguous because different caretakers use different coding conventions. Furthermore, some information stored in the system might require background knowledge or context to be sufficiently usable in the development of predictive models (e.g. a raw lab value). Third, information stored in EMRs is of a highly temporal nature, whereas traditional predictive modeling techniques are unable to take advantage of this temporal dimension. Finally, not all EMR data is always coded; uncoded notes written by a physician are frequently seen as part of EMRs.

In previous research (cf. [1]) we have developed a preprocessing pipeline that includes components to handle the first three characteristics of EMRs, allowing for the application of off-the-shelf machine learning algorithms while benefiting from the

* Corresponding author at: Department of Computer Science, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.
E-mail addresses: m.hoogendoorn@vu.nl (M. Hoogendoorn), psz@mit.edu (P. Szolovits), l.m.g.moons@umcutrecht.nl (L.M.G. Moons), m.e.numans@lumc.nl (M.E. Numans).

http://dx.doi.org/10.1016/j.artmed.2016.03.003
0933-3657/

rich content of the EMRs. However, the pipeline does not yet include a natural language processing (NLP) component which is able to distill useful information from uncoded notes. Research has shown (see e.g. [2]) that such notes can be beneficial when it comes to the development of predictive models, even when coded data are present. In this paper, we study three different NLP approaches and investigate their added value: (1) a simple bag-of-words approach (seen as a benchmark), (2) a topic modeling approach using both latent dirichlet allocation (LDA, cf. [3]) and hierarchical dirichlet processes (HDP, cf. [4]) where topics of text descriptions are identified in an unsupervised way using Bayesian learning, and (3) a dedicated approach (introduced in this paper) which matches the text with a medical ontology (UMLS [5] and an alternative coding scheme called ICPC [6]). Although there are numerous studies aimed at extracting knowledge from medical text, hardly any has tried to compare a range of techniques. In addition, the notes we study are brief, and more keyword oriented and not full blown reports, making the application of known NLP tools for processing medical uncoded text (e.g. [7,8]) less appropriate.

We study the performance of the different NLP techniques in the context of a large anonymized primary care dataset we have access to, covering around 90,000 patients in the region of Utrecht, the Netherlands. Specifically, we focus on predictive modeling of colorectal cancer (CRC), a disease known for its nonspecific symptoms. The dataset consists of coded data (on lab measurements, diagnoses during consultations, medications, and referrals) and includes uncoded doctor's notes associated with each consultation/visit of a patient. We aim to answer the following questions:

1. Can we distill information from the consultation notes that has predictive value with respect to CRC, and if so, what NLP technique results in the highest benefit?
2. Do the consultation notes have added predictive value in addition to the coded data in the dataset?
3. Can we obtain enough information from the consultation notes by themselves to obtain at least equivalent predictive performance for CRC as from just the coded data associated with the consultation?

This paper is organized as follows. Section 2 gives an overview of related work. Thereafter, the dataset is described in more detail in Section 3 followed by an explanation of the different algorithms we use for processing the notes in Section 4. The experimental setup to evaluate the algorithms and answer the research questions is expressed in Section 5 whereas the results are presented in Section 6. Finally, Section 7 is a discussion.

2. Related work

A variety of studies have been performed to explore how useful information can be extracted from medical texts and how valuable this information can be in predictive modeling.

First of all, a number of tools have been developed that allow for the identification of medical terms from text, thereby coupling it to a medical ontology (typically UMLS or a subset thereof). Good examples of such tools are MetaMap [9], health information text extraction system (HITEx, cf. [8]), and cTAKES [7]. All these tools perform basic pre-processing operations first (e.g. tokenization, stemming) and then use an algorithm to perform the best matching with a medical ontology. Research in this area is mostly focused on getting a high accuracy, i.e. attributing the right terms from the medical ontology to the text. The tools aim at exploiting properties of rich, full sentences written in the English language.

Studies that explore the benefit of distilling information from the text (i.e. uncoded data) for the purpose of predictive modeling are more limited in number. In [2] a study is performed in the area of rheumatoid arthritis showing that typed physician notes can complement coded data and result in a higher predictive value. Here, the HITEx system was used to extract relevant UMLS terms from the notes. Ref. [10] studies the usage of topic modeling to help in the classification of pediatric patients, specifically in the prediction of infant colic and shows interesting insights that can be gained from the application of such techniques. In [11] topic modeling is applied on an ICU progress notes dataset to identify mortality risk for ICU patients. The program first assigns UMLS terms to the notes, followed by the application of topic modeling over those terms. The predictive performance is significantly improved compared to the performance without utilizing the notes. Ref. [12] describes a topic modeling approach for mortality prediction based on free-text hospital notes. They have developed models for different time windows, namely in hospital, 30 day post-discharge, and 1 year post discharge. Their results show an improved predictive value in case the text notes are processed using topic modeling in all different time window settings. Finally, Luo et al. [13] explores the usage of more information from the structure of sentences by exploring them in the form of graphs where the nodes in the graph are medical terms found in the sentence and the edges involve the role of the word in the sentence. Subgraphs are identified that occur frequently. The system is shown to outperform all benchmarks. As can be seen, most predictive modeling approaches focus on one technique and try to show that the predictive power is improved; however none study a wide range of techniques and their individual contributions or benefits.

The studies described above focus on one specific technique to distill information from text. In very few works, a comparison of different approaches to extract knowledge from text is found. Tremblay et al. [14] is however an exception: the authors try to explore whether both supervised and unsupervised learning methods can be used to enhance coded data (that might be incomplete). The study focuses on fall injuries. Overall, they show that the two different types of approaches could complement the coded data.

There are also various studies that aim at predicting the specific disease that is the subject of our study: colorectal cancer. Two models in particular are worth mentioning: the Bristol Birmingham equation [15] and the model by Hippisley-Cox [16]. Both have been generated using primary care data. We have used the Bristol Birmingham equation as a benchmark before and have shown that we are able to move statistically significantly beyond that model (cf. [1]). In this paper, we will merely focus on the benefit of using the uncoded notes compared to the performance we have already obtained.

3. Dataset description and preparation

We analyzed an anonymized primary care dataset originating from a network of general practitioners (GPs) centered around the Utrecht University Medical Center, the Netherlands. It contains data of a total of just over 90,000 patients[1] for the period between July 1, 2006 and December 31, 2011. The number of positive CRC cases in the dataset is 588. The dataset covers the following information for each patient, all stored by the date at which the activity took place:

[1] Note that previously [1] we have reported on datasets covering more patients. The dataset we are using for this research is more limited in terms of number of patients, as we only have access to the consultation notes for a subset of the dataset we reported on earlier. It concerns a part of the previously reported dataset covering the practices working with the Promedico ASP system. Experiments using merely coded data do not show considerable differences in performance on this subset compared to the dataset covering more patients.