Computers in Biology and Medicine 76 (2016) 30–38

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records

Reinier Kop [a,*], Mark Hoogendoorn [a], Annette ten Teije [a], Frederike L. Büchner [b], Pauline Slottje [c], Leon M.G. Moons [d], Mattijs E. Numans [b,c,e]

[a] VU University Amsterdam, Department of Computer Science, Amsterdam, The Netherlands
[b] Leiden University Medical Center, Department of Public Health and Primary Care, Leiden, The Netherlands
[c] VU University Medical Center, Academic Network of General Practice, Department of General Practice and Elderly Care Medicine, Amsterdam, The Netherlands
[d] Utrecht University Medical Center, Department of Gastroenterology and Hepatology, Utrecht, The Netherlands
[e] Utrecht University Medical Center, Julius Center of Health Sciences and Primary Care, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 15 April 2016
Received in revised form
14 June 2016
Accepted 20 June 2016

Keywords:
Colorectal cancer
Data mining
Data processing
Electronic medical records
Machine learning

ABSTRACT

Over the past years, research utilizing routine care data extracted from Electronic Medical Records (EMRs) has increased tremendously. Yet there are no straightforward, standardized strategies for pre-processing these data. We propose a dedicated medical pre-processing pipeline aimed at taking on many problems and opportunities contained within EMR data, such as their temporal, inaccurate and incomplete nature. The pipeline is demonstrated on a dataset of routinely recorded data in general practice EMRs of over 260,000 patients, in which the occurrence of colorectal cancer (CRC) is predicted using various machine learning techniques (i.e., CART, LR, RF) and subsets of the data. CRC is a common type of cancer, of which early detection has proven to be important yet challenging.

The results are threefold. First, the predictive models generated using our pipeline reconfirmed known predictors and identified new, medically plausible, predictors derived from the cardiovascular and metabolic disease domain, validating the pipeline's effectiveness. Second, the difference between the best model generated by the data-driven subset (AUC 0.891) and the best model generated by the current state of the art hypothesis-driven subset (AUC 0.864) is statistically significant at the 95% confidence interval level. Third, the pipeline itself is highly generic and independent of the specific disease targeted and the EMR used. In conclusion, the application of established machine learning techniques in combination with the proposed pipeline on EMRs has great potential to enhance disease prediction, and hence early detection and intervention in medical practice.

1. Introduction

Predictive models for diseases can greatly contribute to the domain of health. They can help to identify high risk groups and diseases in an early stage, be a facilitator for proactive care, and assist in selecting the most effective treatment. In the current era, more and more medical data are stored electronically allowing more accurate predictive models to be generated. Amongst others, these data include detailed medication prescriptions, laboratory results, coded and free-text consultation visits. Traditional, more hypothesis-driven, predictive model approaches from the medical and epidemiological domain are still applicable and valuable. However, they no longer necessarily lead to the best possible predictive models as they do not fully utilize the wealth of information contained within the EMRs.

This is where the domain of machine learning comes into play. Algorithms originating from that field are well-suited to maximize usage of the variety of information stored within EMRs and work in a data-driven way contrary to the aforementioned hypothesis-driven approaches. However, even for these sophisticated machine learning approaches, extracting useful predictors from EMRs is not a trivial task due to the very nature of the data. First of all, the data is of a highly temporal nature, whereby consecutive events are stored and linked to individual patient records such as consultations, prescribed medication, referrals and lab results. In addition, the data is typically incomplete caused by (1) the decision of the patient whether or not to present complaints, (2) a physician's observational competence, (3) a physician's registration routines and (4) the type of EMR system being used. Finally, certain values stored in the system cannot easily be interpreted if not seen in a

* Correspondence to: De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.
E-mail address: r.kop@vu.nl (R. Kop).

http://dx.doi.org/10.1016/j.compbiomed.2016.06.019
0010-4825/

context (e.g., a laboratory measurement value lacking a unit or reference values). All of these characteristics make it very difficult to apply off-the-shelf machine learning algorithms.

The aim of this work is to develop a dedicated pre-processing pipeline (consisting of a number of components) that is able to address all the aforementioned issues independent of the EMR used. The pipeline combines several pre-processing algorithms. In addition, we develop a novel approach to enrich the EMR data based on knowledge stored in medical ontologies. To evaluate the pipeline, we study the performance of various machine learning techniques in conjunction with our pipeline by applying them to the case of predicting colorectal cancer (CRC) based on a general practitioner (GP) EMR dataset with over 260,000 subjects. The task of predicting CRC occurrence is extremely relevant as it has the third highest incidence among all cancer types worldwide [29]. Early detection is important for survival and quality of life. However, it is also challenging as the symptoms are mostly non-specific and show considerable overlap with benign disease.

Various approaches have been proposed to cope with some of the aforementioned issues. For example, Patnaik et al. [31] and Batal et al. [3] developed approaches that exploit the temporal dimension. However, none of these approaches have tried to cover the entire range of problems in full at once. To the best of our knowledge, the techniques have not been applied to GP data in a very extensive way to show the benefit of the approaches in that context.

This paper is organized as follows. In Section 2 an overview of related work is presented. Section 3 presents the pipeline, including its individual components. The experimental setup to evaluate the approach is presented in Section 4, followed by the results in Section 5. Section 6 is a discussion, and we conclude in Section 7.

2. Related work

EMR data consist of various data types, and often contain missing or wrong data [15]. Processing the data so that samples of uniform length are created is an important first step if traditional predictive modeling techniques are to be used. Though EMR data are timestamped in nearly all cases, it has been shown that clinical prediction tasks have reasonable performance when ignoring the data's temporality (e.g., [8,20,12]), at times even outperforming hypothesis-driven approaches (e.g., [30]). An important limitation in some of these studies is the data-driven approach cannot be considered purely data-driven. Rather, the available data is often already tailored towards the disease or disorder under investigation ([20], and to some extent, [8,30]). For example, Kurt et al. [20] investigate the presence of coronary artery disease using features known to be good predictors for the disease. Such comparisons, though useful to showcase the potential of data-driven prediction in EMRs, are less interesting because new potential predictors will not be found. The present work allows for this as the available primary care data contains information not specific to any particular disease.

To further improve prediction quality, it is potentially useful to apply temporal pattern mining on EMR datasets. Temporal patterns are implicitly contained in EMR data, but to allow traditional prediction methods to use these patterns as features, patterns must be mined in advance. Temporal pattern mining can be viewed as a subtype of association learning [1] in the sense that we are interested in (temporal) relations between events rather than items (see e.g., [34,16,7,26,3]). Batal et al. [3] build progressively larger temporal patterns using a modified version of Agrawal and Srikant's apriori algorithm. However, this algorithm is applied on a dataset spanning days as opposed to months. Kop et al. [18] validate the effectiveness of their method for longer periods of time by reporting an increase in performance when applying their algorithm on EMR data spanning months.

Another way to find additional features is to look at the vast amount of semantic data available in ontologies on the web (e.g., SNOMED, the Systematized Nomenclature of Medicine). Melton et al. [24] use SNOMED to find patient similarity using semantic links between concepts. In La-Ongsri and Roddick [28], EMR concepts are mapped to ontology concepts with the goal to allow multiple levels of abstraction for efficient database usage. This work explores the usage of ontologies in another way: to generate additional features used in a prediction task.

This work is a continuation of previous work in which we explored different subsets of the current data. In Hoogendoorn et al. [12] the potential of EMR data was validated using established machine learning techniques and simple (non-temporal) pre-processing, already resulting in predictive performances regarding CRC better than solely relying on age/gender. This is often difficult within EMR data, as age and gender are known to be among the most obvious predictors in clinical prediction tasks. In Kop et al. [18], this research was expanded on by applying temporal pattern mining, further improving performance. The work currently described builds upon the above by (1) formalizing the entire pipeline, (2) adding lab measurement contextualization, (3) adding semantic enrichment of the data and (4) applying all this on a larger dataset. Furthermore, it attempts to validate the temporal pattern mining work of Batal et al. [3] and reinforces the potential of data-driven research.

Finally, worth mentioning are the analytical platforms Informatics for Integrating Biology and the Bedside (i2b2, [27]) and, by extension, Shared Health Research Information Network (SHRINE, [37]) that allow physicians and researchers to filter and analyze medical data. The important difference is that those platforms are built mostly for human users to better observe and understand their data, whereas our pipeline transforms medical data in order to automatically generate models. In theory, those platforms could be extended to incorporate a pre-processing pipeline such as the one described in this paper. This would allow for large-scale data mining on aggregated medical data sets, which is in line with our research goals.

3. Methodology

In this Section, we introduce our pre-processing pipeline. The code for the pipeline has been made available online.[1] The pipeline is composed of four steps, as shown in Fig. 1. The setup is highly generic and independent of the specific disease targeted and the used EMR. The data present in the EMR includes time stamped events in these categories: consultations, medication, referrals, and values of laboratory measurements. Formally, we specify a set of measurements $a_1, \ldots, a_m$ and a set of patients within our dataset as $p_1, \ldots, p_n$. The domain (i.e., possible values) of a measurement a_i is denoted by A_i. In addition we assume a number of time points $t_{start}, \ldots, t_{end}$ where time is considered in days and the duration is the same for all patients. The value of a measurement i as a specific time point t for patient p is denoted by $a_i(p, t)$. Finally, the type of measurement is specified by means of the type function: $type(a_i)$ which can take the values *consultation*, *medication*, *referral*, and *lab*. Each of the measurements has values set according to some coding scheme to classify the data (e.g., International Classification for Primary Care (ICPC; [4]), for symptoms and diagnosis or Anatomical Therapeutic Chemical (ATC) classification system for

[1] See https://github.com/ReinierKop/EMR-pre-processing-pipeline.