Journal of Biomedical Informatics 57 (2015) 377–385

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Scaling and contextualizing personalized healthcare: A case study of disease prediction algorithm integration

Keith Feldman[a], Darcy Davis[b], Nitesh V. Chawla[a,*]

[a] University of Notre Dame, Notre Dame, IN, USA
[b] Advocate Healthcare Downers Grove, IL, USA

ARTICLE INFO

Article history:
Received 1 April 2015
Revised 21 July 2015
Accepted 26 July 2015
Available online 5 August 2015

Keywords:
Personalized healthcare
Big Data
Clinical informatics
Data mining

ABSTRACT

Today, advances in medical informatics brought on by the increasing availability of electronic medical records (EMR) have allowed for the proliferation of data-centric tools, especially in the context of personalized healthcare. While these tools have the potential to greatly improve the quality of patient care, the effective utilization of their techniques within clinical practice may encounter two significant challenges. First, the increasing amount of electronic data generated by clinical processes can impose scalability challenges for current computational tools, requiring parallel or distributed implementations of such tools to scale. Secondly, as technology becomes increasingly intertwined in clinical workflows these tools must not only operate efficiently, but also in an interpretable manner. Failure to identity areas of uncertainty or provide appropriate context creates a potentially complex situation for both physicians and patients. This paper will present a case study investigating the issues associated with first scaling a disease prediction algorithm to accommodate dataset sizes expected in large medical practices. It will then provide an analysis on the diagnoses predictions, attempting to provide contextual information to convey the certainty of the results to a physician. Finally it will investigate latent demographic features of the patient's themselves, which may have an impact on the accuracy of the diagnosis predictions.

1. Introduction

Over the past decade the digitization of healthcare records has provided a foundation for data scientists and clinicians alike to employ data mining and machine learning techniques on medical datasets [1]. These techniques have allowed for not only substantial improvements to existing clinical decision support systems, but also a platform for improved patient-centered outcomes through the development of personalized prediction models tailored to a patient's medical history and current condition [2–5]. While powerful, the integration of such tools into clinical workflows is a challenging endeavor. This paper will address two major components integral for the successful integration of analytical tools into a clinical workflow.

Of first concern is incorporating these tools within a clinical time frame and context. Due to the time sensitive nature of clinical scenarios, the machine learning models on which these tools are built must allow for execution within a relevant timeframe, which is often only the duration of a patient visit. As the data becomes increasingly available, drawn from multiple sources and encompasses multiple modalities, it also becomes critical for machine learning methods to process them both accurately and quickly. If a patient's current visit is used during a physician's office visit to suggest a series of personalized recommendations, then it is warranted that the back-end prediction engine is able to deliver within this timeframe. However the quantity of data necessary to build these models presents a significant issue for successful long-term utilization. It is important to remember that each medical encounter will result in additional data added to a patient's electronic health record.

Although for now this limitation can be overcome with algorithmic ingenuity, healthcare's "Big Data" may soon exceed the ability of standard data processing techniques, given its variety, veracity and volume. As a result we must look to other approaches, such as distributed computation, in order to scale personalized healthcare models. Failure to do so may result in the need to artificially restrict the data on which the models are built. This would typically be accomplished either through the process of feature or instance selection, the difficulties of which have already been well documented [6–8].

The ability to utilize these tools within a constrained time window is not the only obstacle to their deployment in clinical

[*] Corresponding author.
E-mail addresses: kfeldman@nd.edu (K. Feldman), davis.darcy.a@gmail.com (D. Davis), nchawla@nd.edu (N.V. Chawla).

http://dx.doi.org/10.1016/j.jbi.2015.07.017
1532-0464/© 2015 Elsevier Inc. All rights reserved.

settings. The second concern stems from the realization that with the implementation of predictive models, patients are no longer the only individuals receiving treatment recommendations. Physicians will now begin to receive treatment recommendations personalized to their current patient. Although these tools and techniques are designed to augment the existing skills of the physician, expanding their clinical knowledge beyond their prior experience and education, they also introduce a new challenge of providing an appropriate narrative with the predictions or the analytics.

It is important to remember that while these predictions may be the result of advanced machine learning models, they have to be assessed and communicated within a clinical context. While physicians will likely be equipped to understand the clinical aspects of the recommendations they receive, as well as the risks associated with them, there is currently no process in place to ensure the algorithmic results are clinically interpretable. To date a substantial set of prior work has been done tuning the performance of these algorithms, and although these evaluations help to create functional and effective models, many fail to perform any medically focused evaluation of the predicted instances [9,10].

However due to the complexities of human disease, and the uniqueness of each patient, a deeper understanding of the algorithm producing the recommendation is critical for the successful integration of these tools into a clinical workflow. As an example the high probability of a frequently misdiagnosed disease may not be as diagnostically useful for a physician as would a slightly lower probability disease, that when predicted is almost always correct.

This paper will provide a case study addressing each of the integration challenges discussed above, walking through the process of bringing a disease prediction algorithm out of an academic setting and preparing it for the complexities of a clinical setting. For the study we will be utilizing the disease prediction algorithm CARE (Collaborative Assessment and Recommendation Engine) [11]. We have chosen CARE as the algorithm has already been shown to be effective, and as we will see CARE is a good proxy for an entire class of disease prediction algorithm utilizing *patient similarity techniques*. The scaling of CARE using distributed computing constructs can thus provide a possible template for integration with other existing disease prediction algorithms that leverage large-scale electronic health care records. Finally, as we will see throughout this paper, it is important to contextualize the outcome of any clinical decision making aid for patient as well as physician consumption in order to reach the goal of “patient empowerment and engagement”.

The paper is structured as follows. We will begin with a backend system-level investigation into the task of scaling the CARE algorithm to accommodate the patient datasets representative of true clinical databases. Next we include an in-depth analysis of the CARE algorithm from a clinical perspective, identifying those diagnoses that CARE can frequently predict correctly, and those that may present difficulty, and how these insights may translate to the patient. It will then evaluate patient demographic data, identifying latent features which may indicate the difficulty of correctly predicting an individual’s future diseases as well the distribution of diagnosis across the highest and lowest performing individuals.

2. Disease prediction algorithm

Over recent years a number of disease prediction algorithms have been developed to accomplish a multitude of tasks. While some algorithms focus on modeling an individual’s risk of developing specific diagnoses such as cardiac conditions, others can be utilized in a more general approach to identify individuals’ high-risk future conditions [5,12–14]. The past few years have witnessed further development of these predictive tasks, creating systems to model tasks such as the progression of degenerative diseases as well as extensions into the genomic field, identifying target sites utilized in biomarker and drug discovery [15–17].

2.1. The CARE algorithm

Amongst the earliest general disease prediction models for personalized healthcare that leverage patient similarity is the CARE Algorithm. CARE uses collaborative filtering of an individual’s medical history in order to identify high likelihood diagnoses in the patient’s future. Collaborative filtering is traditionally a technique by which similar individuals are identified through a set of known shared preferences or attributes. The intent of collaborative filtering is to identify new preferences for an individual based on the non-shared preferences identified between other similar individuals [18–20]. While these techniques have been utilized for many years in online applications such as movie, book and product recommendations, they have recently shown promise in the healthcare domain as well. Beyond CARE, a number of recent algorithms have utilized collaborative filtering for applications such as nursing decision support, medical context identification, and identification of sudden deterioration for a patient’s medical condition [21–23].

An architectural diagram of the standard CARE algorithm can be seen in Fig. 1 [11] and is comprised of three major steps. For a patient *p*, the algorithm begins with an initial filtering on all patients within the database, isolating only those patients who have at least one disease in common with *p*. This is done as totally disparate patients offer no potential similarity information, and will only serve to extend the computation time. Next utilizing this subset of patients the collaborative filtering step is performed. CARE’s collaborative filtering algorithm incorporates a binary coding of diagnoses codes, with 1 representing a present diagnosis, and 0 one which is absent or undiagnosed. In addition, the inverse frequency of each diagnosis is used in order to give higher weight to less common diagnosis. This is particularly important as some diagnosis, such as hypertension, are present in 33.64% of all patients [11]. CARE also incorporates a time component representing when in the patient’s medical history did they develop a disease. Next, an ensemble of such collaborative filtering models is generated for each similar set of patients identified for each disease of *p*. Finally the results are then aggregated, yielding a ranked list of high probability diseases for *p*.

3. Materials and methods

As mentioned prior the CARE algorithm has been previously shown to be accurate, and in an effort to maintain consistency with the published work the original dataset and source code were used in this case study as in the original CARE evaluation.

3.1. Data

The dataset utilized for this work contains approximately 32 million anonymized Medicare claims each representing a patient visit, accounting for just over 13 million unique patients. As per the original CARE work, in order to ensure sufficient diagnosis history during training only those patients with over 5 visits were considered for evaluation in this paper [11].

The Medicare claim itself contains 16 features as well as a unique identifier for patients with multiple visits. The claim is broken into two main sections, *patient demographics* and *diagnosis codes*. Patient demographics contains the date of the visit, the patient’s