International Journal of Medical Informatics 112 (2018) 59–67

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Federated learning of predictive models from federated Electronic Health Records

Theodora S. Brisimi[a], Ruidi Chen[a], Theofanie Mela[c], Alex Olshevsky[a], Ioannis Ch. Paschalidis[a,b,*], Wei Shi[a,d]

[a] Department of Electrical & Computer Engineering, and Division of Systems Engineering, Boston University, 8 Saint Mary's St., Boston, MA 02215, United States
[b] Department of Biomedical Engineering, Boston University, 44 Cummington Mall, Boston, MA 02215, United States
[c] Electrophysiology Lab/Arrhythmia Service, Massachusetts General Hospital, 55 Fruit St., Boston, MA 02114, United States
[d] School of Electrical & Computer Engineering, Arizona State University, Tempe, AZ, United States

ARTICLE INFO

Keywords:
Predictive models
Hospitalization
Heart diseases
Distributed learning
Electronic Health Records (EHRs)
Federated databases

ABSTRACT

Background: In an era of "big data," computationally efficient and privacy-aware solutions for large-scale machine learning problems become crucial, especially in the healthcare domain, where large amounts of data are stored in different locations and owned by different entities. Past research has been focused on centralized algorithms, which assume the existence of a central data repository (database) which stores and can process the data from all participants. Such an architecture, however, can be impractical when data are not centrally located, it does not scale well to very large datasets, and introduces single-point of failure risks which could compromise the integrity and privacy of the data. Given scores of data widely spread across hospitals/individuals, a decentralized computationally scalable methodology is very much in need.

Objective: We aim at solving a binary supervised classification problem to predict hospitalizations for cardiac events using a distributed algorithm. We seek to develop a general decentralized optimization framework enabling multiple data holders to collaborate and converge to a common predictive model, without explicitly exchanging raw data.

Methods: We focus on the soft-margin l_1-regularized sparse Support Vector Machine (sSVM) classifier. We develop an iterative cluster Primal Dual Splitting (cPDS) algorithm for solving the large-scale sSVM problem in a decentralized fashion. Such a distributed learning scheme is relevant for multi-institutional collaborations or peer-to-peer applications, allowing the data holders to collaborate, while keeping every participant's data private.

Results: We test cPDS on the problem of predicting hospitalizations due to heart diseases within a calendar year based on information in the patients Electronic Health Records prior to that year. cPDS converges faster than centralized methods at the cost of some communication between agents. It also converges faster and with less communication overhead compared to an alternative distributed algorithm. In both cases, it achieves similar prediction accuracy measured by the Area Under the Receiver Operating Characteristic Curve (AUC) of the classifier. We extract important features discovered by the algorithm that are predictive of future hospitalizations, thus providing a way to interpret the classification results and inform prevention efforts.

1. Introduction

1.1. Motivation

As the volume, variety, velocity and veracity (the four V's) of the clinical data grow, there is greater need for efficient computational models to mine these data. Insights from these techniques could help design efficient healthcare policies, detect disease causes, provide medical solutions that are personalized and less costly, and finally, improve the quality of care for the patients. We are motivated by problems in the medical domain that can be formulated as binary supervised classification problems and solved using Support Vector Machines; the applications range from prediction of the onset of diabetes [1,2], prediction of hospitalizations for cardiac events [3],

* Corresponding author at: Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02215, United States.
E-mail address: yannisp@bu.edu (I.C. Paschalidis).

https://doi.org/10.1016/j.ijmedinf.2018.01.007
Received 30 September 2017; Received in revised form 26 November 2017; Accepted 8 January 2018

prediction of medication adherence in heart failure patients [4], and cancer diagnosis [5], to automated recognition of the obstructive sleep apnea syndrome [6].

Results in the literature suggest that sparse classifiers (i.e., those that rely on few features), have strong predictive power and generalize well out-of-sample [7,8], providing at the same time interpretability in both models and results. Interpretability is crucial for healthcare practitioners to trust the algorithmic outcomes. Another major concern, especially in the medical domain, is the privacy of the data, attracting recent research efforts [9–11]. Two well-known examples of privacy breaches are the Netflix Prize and the Massachusetts Group Insurance Commission (GIC) medical records database. In both cases, individuals were identified even though the data had been through a de-identification process. This demonstrated that one's identity and other sensitive information could be compromised once a single center has access and processes all the data. Especially under the Precision Medicine Initiative [12], in the near future, these data could include individuals' genome information, which is too sensitive to be shared.

We are particularly interested in addressing three challenges tied to healthcare data: (1) data reside in different locations (e.g., hospitals, doctors' offices, home-based devices, patients' smartphones); (2) there is a growing availability of data, which makes scalable frameworks important; and (3) aggregating data in a single database is infeasible or undesirable due to scale and/or data privacy concerns. In particular, even though maintaining all data in a central location enables the implementation of anonymization measures (e.g., k-anonymity [13]), it introduces a single point of attack or failure and makes it possible for a data breach to expose identifiable data for many individuals. Furthermore, establishing a central data repository requires significant infrastructure investments and overcoming information governance hurdles such as obtaining permissions for storing and processing data. Instead, a decentralized computational scheme that treats the available data as part of a federated (virtual) database, avoiding centralized data collection, processing, and raw data exchanges, may address the above challenges.

1.2. Aim

The focus of this paper is to develop a distributed (federated) method to predict hospitalizations during a target year for patients with heart diseases, based on their medical history as described in their Electronic Health Records (EHRs). The records of each patient may lie with them in the patient's smartphone, or may be stored in the EHR systems of different hospitals. In all cases, the collaboration of different parties (agents) is required to develop a global hospitalization prediction model. We will formulate the problem as a binary supervised classification problem and we will develop a distributed soft-margin ℓ_1-regularized (sparse) Support Vector Machines (sSVM) algorithm. We consider SVMs because they are effective classifiers [14] and perform well in predicting hospitalizations [3]. Furthermore, sparse classifiers can reveal relatively few predictive features and, thus, enable interpretation of the predictions [15].

1.3. Health application

We focus on cardiovascular conditions because they comprise a significant portion of morbidity and mortality, as well as, hospitalization in the U.S. and worldwide. In the fact, in the U.S. alone, more than 30% (equal to $9 billion) of hospitalizations deemed preventable are due to cardiovascular conditions [16]. For many decades, the research interest has been focused on understanding the pathophysiology of these conditions and treating them effectively. The efforts have now shifted to the understanding of the disease process and the early prevention. This goal has obvious public health implications, but also socioeconomic significance. It is well known that preventing the progression of the disease process by intensified follow up and treatment can result in long-term stability and improved survival of the patient. Hospitalization is a well-known negative prognostic factor for cardiovascular disease outcome. One critical step in the effort to halt the disease process is the understanding of the etiology and modifiable risk factors of hospitalization.

1.4. Main contributions

We summarize our main contributions below:

- We develop a federated optimization scheme (cPDS) for solving the sparse Support Vector Machine problem. Advantages include scalability and the fact that it avoids raw data exchanges, which is important in healthcare. We also demonstrate that cPDS has improved convergence rate and favorable communication cost compared to various centralized and distributed alternatives.
- We apply our new methodology to a dataset of de-identified Electronic Heart Records from the Boston Medical Center, containing patients with heart-related diseases. Each patient is described by a set of features, including demographics, diagnoses, prior admissions, and other relevant medical history.
- We use cPDS to differentiate between patients that are likely and not likely to be hospitalized within a target year and report and discuss the experimental results.
- The proposed cPDS framework is general and can be applied to any learning problem with a "nonsmooth + nonsmooth" loss function objective. Such problems can be found in machine learning, where we aim to minimize functions with non-smooth regularizers, or in distributed model predictive control.

2. Material and methods

2.1. Objective and background

We consider a dataset extracted from an EHR system, containing patients' demographic data such as age, gender, and race, physical characteristics such as weight, height, Body Mass Index (BMI), medical history captured by diagnoses, procedures, office visits, and a history of drug prescriptions, all captured by a feature vector $\varphi_i \in \mathbb{R}^d$, for each patient $i = 1, \ldots, n$. We are interested in predicting whether or not a patient will be hospitalized in a given year, for instance in the next calendar year from the time the record is being examined. We denote a hospitalization by a label $l_i = +1$, and a non-hospitalization by a label $l_i = -1$. Using machine learning terminology, this is a binary classification problem. Using the popular Support Vector Machine (SVM) classifier [14], we seek to find a hyperplane that maximizes the margin ("distance") between the two classes, while allowing a few points to be misclassified (as shown in Fig. 1). Further requiring that a few features are used, we end up with a sparse Support Vector Machine (sSVM) problem: