Information Sciences 330 (2016) 245–259

Contents lists available at ScienceDirect

Information Sciences

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

A distributed ensemble approach for mining healthcare data under privacy constraints

Yan Li, Changxin Bai, Chandan K. Reddy*

Department of Computer Science, Wayne state University, Detroit, MI 48202, United States

ARTICLE INFO

Article history:
Received 16 June 2015
Revised 14 September 2015
Accepted 6 October 2015
Available online 20 October 2015

Keywords:
Privacy-preserving data mining
Ensemble learning
Electronic health record
Boosting
Machine learning
Healthcare

ABSTRACT

In recent years, electronic health records (EHRs) have been widely adapted at many healthcare facilities in an attempt to improve the quality of patient care and increase the productivity and efficiency of healthcare delivery. These EHRs can accurately diagnose diseases if utilized appropriately. While the EHRs can potentially resolve many of the existing problems associated with disease diagnosis, one of the main obstacles in effectively using them is the patient privacy and sensitivity of the medical information available in the EHR. Due to these concerns, even if the EHRs are available for storage and retrieval purposes, sharing of the patient records between different healthcare facilities has become a major concern and has hampered some of the effective advantages of using EHRs. Due to this lack of data sharing, most of the facilities aim at building clinical decision support systems using limited amount of patient data from their own EHR systems to provide important diagnosis related decisions. It becomes quite infeasible for a newly established healthcare facility to build a robust decision making system due to the lack of sufficient patient records. However, to make effective decisions from clinical data, it is indispensable to have large amounts of data to train the decision models. In this regard, there are conflicting objectives of preserving patient privacy and having sufficient data for modeling and decision making. To handle such disparate goals, we develop two adaptive distributed privacy-preserving algorithms based on a distributed ensemble strategy. The basic idea of our approach is to build an elegant model for each participating facility to accurately learn the data distribution, and then transfer the useful healthcare knowledge acquired on their data from these participators in the form of their own decision models without revealing and sharing the patient-level sensitive data, thus protecting patient privacy. We demonstrate that our approach can successfully build accurate and robust prediction models, under privacy constraints, using the healthcare data collected from different geographical locations. We demonstrate the performance of our method using the type-2 diabetes EHRs accumulated from multiple sources from all fifty states in the U.S. Our method was evaluated on diagnosing diabetes in the presence of insufficient number of patient records from certain regions without revealing the actual patient data from other regions. Using the proposed approach, we also discovered the important biomarkers, both universal and region-specific, and validated the selected biomarkers using the biomedical literature.

* Corresponding author. Tel.: +13135779005.
E-mail address: reddy@cs.wayne.edu (C.K. Reddy).

http://dx.doi.org/10.1016/j.ins.2015.10.011
0020-0255/

1. Introduction

In recent years, electronic health records (EHRs) have been widely adapted at many healthcare facilities in an attempt to improve the quality of patient care and increase the productivity and efficiency of healthcare delivery. Typically, EHRs are generated and maintained within a particular healthcare facility, such as a hospital, clinic or physician's office. These EHRs can not only aid in various daily healthcare operations but also help in accurately diagnosing diseases, if utilized appropriately [32].

While the EHRs can resolve many of the existing problems associated with disease diagnosis, one of the main obstacles in effectively using them is the patient privacy and sensitivity of the medical information available in the EHR. EHRs generally contain patient information about demographics, diagnostics, medications, living habits and other health-related information. It is needless to say that most of these information is extremely sensitive [31]. EHR systems must abide the Health Insurance Portability and Accountability Act (HIPAA) [6] to preserve the privacy of patient information. Thus, the public EHR products must be certified by certain institutes such as Certification Commission for Healthcare Information Technology (CCHIT), and Office of the National Coordination for Health Information Technology (ONC) [38]. Due to these legal concerns and medical ethics [4,19], physicians are always keen about maintaining the highest possible standards while protecting patient privacy [36].

Due to these concerns, even if the EHRs are available for storage and retrieval purposes, sharing of patient records between different healthcare facilities has become a major concern and has hampered some of the effective advantages of using EHRs. Due to this lack of data sharing, most of the facilities do not have any other option but to build their own clinical decision support systems with limited amount of patient data available in their own EHRs for important diagnosis related decisions. In addition, it becomes quite infeasible for a newly established healthcare facility, small hospital, or rural hospital to build a robust decision making system due to lack of sufficient patient records. However, to make effective decisions from clinical data, it is indispensable to have large amounts of data to train the decision making models. As a result, these small and/or rural hospitals are less motivated, and in 2011, only 20.8% of them were using EHR systems [12]. In this regard, it is clear that there is a conflicting objective of maintaining patient privacy and having sufficient data for modeling and decision making.

The problem statement of the work that is being developed in this paper is as follows. Given healthcare data collected from multiple facilities, how do we obtain a decision model that leverages the knowledge from all the facilities without revealing any patient specific information from any of the individual facilities. In other words, the goal is to develop a knowledge based data integration mechanism in a privacy-preserving context.

To deal with these challenges, we propose a privacy-preserving data mining framework based on horizontally distributed healthcare data. The goal of our work is to build an effective decision making system that can utilize the knowledge from multiple facilities (or geographical locations) without revealing any patient-level information [24]. In our approach, an elegant model for each participating facility is accurately learnt for approximating the local data distribution, and the useful healthcare knowledge acquired on such local data is then transferred in the form of their own decision models without revealing and sharing the sensitive data, thus, protecting the patient privacy. Transferring knowledge between multiple locations is a common practice with the goal of improving the prediction power [22]. Our approach can successfully build accurate and robust prediction models, under privacy constraints, with healthcare data collected from different geographical locations. Each participator shares its own local model with others and builds a specific integrated model based on its own specifications.

Merely trying to acquire the entire knowledge available from all the participators without carefully accounting for the distribution differences can potentially degrade the performance of the classifier [33]. While the overall data distribution for a particular disease must be similar within different hospitals since we are dealing with the same disease, there will still be certain significant differences that arise due to the differences in the demographics of the patient population. Such distribution differences will play a vital role in building robust integrated decision making models that are specific to the population group. However, in the horizontal distribution privacy-preserving problem, there is no access to the original EHR data, one cannot directly measure the data distribution differences among participators but can only approximately say how large the difference might be by analyzing the difference between the models trained by each participator.

Most of the state-of-the-art privacy-preserving data mining methods for horizontally distributed data are built based on a star network, where an untrusted third-party is needed [11,40]. Each participator shares their statistical data distribution with a centralized agent, and this central agent is responsible for building the integrated decision model [24]. However, this framework suffers from two important issues: (1) participators need frequent information exchange with the central agent and hence the communication cost is high; (2) the decision model is built on the central agent whose aim is to provide a decision support for all participators, but it ignores the data distribution differences between the participators.

In addition, in the horizontal distributed privacy preserving data mining literature, there is no prior work on preventing "negative impact during integration". In our work, as each participator will build a specific integrated model based on their own specifications, they can selectively decide which of the models from other participators can be used to build the integrated model. We build local prediction models to represent data, detect the data distribution difference, and transfer knowledge. In our work, the AdaBoost algorithm is chosen to be the local learner for each participator because it is a simple yet effective classifier which is extensively studied in the literature.

The main contributions of our work are summarized as follows:

- Propose an adaptive distributed privacy-preserving data mining technique based on an ensemble strategy which can successfully acquire knowledge from multiple healthcare facilities without gaining access to the sensitive patient data.