Journal of Biomedical Informatics 58 (2015) 28–36

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A probabilistic topic model for clinical risk stratification from electronic health records

Zhengxing Huang [a,*], Wei Dong [b], Huilong Duan [a]

[a] College of Biomedical Engineering and Instrument Science, Zhejiang University, China
[b] Department of Cardiology, Chinese PLA General Hospital, China

ARTICLE INFO

Article history:
Received 16 May 2015
Revised 4 September 2015
Accepted 5 September 2015
Available online 11 September 2015

Keywords:
Risk stratification
Probabilistic topic model
Latent Dirichlet Allocation
Electronic health record
Patient sub-profile
Joint patient sub-profile-risk modeling

ABSTRACT

Background and objective: Risk stratification aims to provide physicians with the accurate assessment of a patient's clinical risk such that an individualized prevention or management strategy can be developed and delivered. Existing risk stratification techniques mainly focus on predicting the overall risk of an individual patient in a supervised manner, and, at the cohort level, often offer little insight beyond a flat score-based segmentation from the labeled clinical dataset. To this end, in this paper, we propose a new approach for risk stratification by exploring a large volume of electronic health records (EHRs) in an unsupervised fashion.

Methods: Along this line, this paper proposes a novel probabilistic topic modeling framework called probabilistic risk stratification model (PRSM) based on Latent Dirichlet Allocation (LDA). The proposed PRSM recognizes a patient clinical state as a probabilistic combination of latent sub-profiles, and generates sub-profile-specific risk tiers of patients from their EHRs in a fully unsupervised fashion. The achieved stratification results can be easily recognized as high-, medium- and low-risk, respectively. In addition, we present an extension of PRSM, called weakly supervised PRSM (WS-PRSM) by incorporating minimum prior information into the model, in order to improve the risk stratification accuracy, and to make our models highly portable to risk stratification tasks of various diseases.

Results: We verify the effectiveness of the proposed approach on a clinical dataset containing 3463 coronary heart disease (CHD) patient instances. Both PRSM and WS-PRSM were compared with two established supervised risk stratification algorithms, i.e., logistic regression and support vector machine, and showed the effectiveness of our models in risk stratification of CHD in terms of the Area Under the receiver operating characteristic Curve (AUC) analysis. As well, in comparison with PRSM, WS-PRSM has over 2% performance gain, on the experimental dataset, demonstrating that incorporating risk scoring knowledge as prior information can improve the performance in risk stratification.

Conclusions: Experimental results reveal that our models achieve competitive performance in risk stratification in comparison with existing supervised approaches. In addition, the unsupervised nature of our models makes them highly portable to the risk stratification tasks of various diseases. Moreover, patient sub-profiles and sub-profile-specific risk tiers generated by our models are coherent and informative, and provide significant potential to be explored for the further tasks, such as patient cohort analysis. We hypothesize that the proposed framework can readily meet the demand for risk stratification from a large volume of EHRs in an open-ended fashion.

1. Introduction

Risk stratification, as the starting point for developing prevention or management strategies, plays a key role in personalized medicine, care plan management, drug development, and cost estimation [1,2,19,30]. For health professionals, once they estimate the risk of a patient, they may be able to apply measures to decrease the risk for the patient and improve the outcome.

Traditional risk stratification models are established on a small hand-picked subset of patient features (also called risk factors) from highly stratified patient cohorts [3,4,13,23] such that statistical analysis techniques (e.g., logistic regression, Cox regression, etc.) can be employed to train a risk stratification model in which the contribution of each individual patient feature to the overall risk of patients can be estimated [7]. The trained model is then

[*] Corresponding author.
E-mail address: zhengxinghuang@zju.edu.cn (Z. Huang).

http://dx.doi.org/10.1016/j.jbi.2015.09.005
1532-0464/© 2015 Elsevier Inc. All rights reserved.

applied to test patients to compute their overall risk scores, based on which, patient cohorts can be stratified into several tiers, e.g., *low*-, *medium*-, and *high*-risk [3,10]. These models are valuable and have been widely studied in clinical settings, however, they are fragmented where the conclusion only holds under well-controlled conditions [13].

Recently, with the rapid development of healthcare information systems, a large collection of electronic health records (EHRs) has become available, which provides the opportunity to study medical cases, evidence and knowledge for various medical applications [5,24]. In particular, heterogeneous clinical information for patients (e.g., the patient demographics, laboratory tests results, radiological examination reports, etc.) are recorded in EHRs, which suggests data-centric and hypothesis-free approach from which machine learning techniques can be utilized for risk stratification [13]. Considerable work has been done in this line in order to achieve automatic feature selection from a large volume of EHRs, and significantly improve the accuracy of risk stratification [6,13,21,23].

However, most of the existing approaches rely on supervised learning models trained from labeled datasets where each patient's EHR instance has been labeled as high-, medium-, or low-risk prior to training. Such labeled datasets are not always easily obtained from EHRs in clinical applications [21]. Note that patient risk scores are seldom explicitly recorded in EHRs. And it is a very tedious and time-consuming process to label patient dataset, piece by piece, in a posteriori manner. It has thus motivated the problem of using unsupervised or weakly supervised approaches for risk stratification. Another common deficiency of the aforementioned work is that it only focuses on detecting the overall risk score of a patient, without performing an in-depth analysis to discover latent patient sub-profiles and associated risk tiers. In clinical practice, a patient clinical status may be very complex and dynamic, such as comorbidities, complications, infections, or poisonings. This in turn leads to a patient's EHR to be represented by a mixture of latent patient sub-profiles, and each sub-profile is described as a set of patient features with their values [5,25]. Although detecting patient sub-profiles is a useful step for retrieving more detailed medical information, the lack of risk analysis on the extracted patient sub-profiles often limits the effectiveness of the mining results, as the physicians are not only interested in the overall risk score of a patient, but also his/her clinical characteristics (i.e., sub-profiles) and the risk toward the specific sub-profiles discovered. For example, coronary heart disease (CHD) patients with high-risks may have different sub-profiles, e.g., CHD with diabetes, or CHD with renal insufficiency, etc. Note that different clinical sub-profiles may result in different treatment plans for patients. In this sense, detecting a patient's clinical sub-profiles and stratifying his/her risk in an integrated manner can provide meaningful values to the physicians with more informative risk scoring of the patient.

In this paper, we propose a novel probabilistic topic model, i.e., probabilistic risk stratification model (PRSM), which integrates patient sub-profile discovery and risk stratification from EHRs in an unsupervised manner. The proposed PRSM is an extension of the state-of-the-art topic model Latent Dirichlet Allocation (LDA) [26], by constructing an additional risk tier layer, assuming that risk tiers are generated dependent on patient sub-profile distribution, and patient features are generated dependent on the joint risk tier-patient sub-profile distributions. In this way, our model links both patient sub-profile discovery and risk stratification simultaneously. In addition, we incorporate prior knowledge into the PRSM, called weakly supervised PRSM (WS-PRSM), to improve its accuracy, and more importantly, makes it highly portable to risk stratification tasks of various diseases. To the best of our knowledge, no other existing approaches present the same merits as our models.

The proposed approach is evaluated on a collection containing 3463 EHRs of CHD patients collected from the Cardiology Department of the Chinese PLA General Hospital. Experimental results validate the feasibility of jointly modeling risk tiers and patient sub-profiles from EHR data, and show that our models achieve comparable performance compared to existing supervised algorithms. Aside from automatically stratify patient risks using EHR data, our model can also extract meaningful patient sub-profiles with risk associations as illustrated by some patient sub-profile examples extracted from the experimental dataset.

The rest of this article is structured as follows. Section 2 summarizes some related studies. Section 3 presents preliminary knowledge of the proposed approach. Section 4 describes the proposed approach for risk stratification associated with the discovery of latent patient sub-profiles from EHRs. Section 5 carefully presents our experimental results on a clinical dataset collected from a Chinese hospital. Finally, some conclusions are given in Section 6.

2. Related work

As a fundamental problem for medical informatics, risk stratification is indispensable to modern clinical decision support systems by providing healthcare practitioners an assessment of an individual's risk against an adverse outcome [6,9]. Great bulk of work has been focused on the problem of risk stratification. Taking coronary heart disease (CHD) as an example, recent systematic reviews found that there are over 100 CHD risk stratification models produced between 1999 and 2009 [17,18], and the vast majority of these models are based on statistical analysis techniques [7,15]. For example, Fonarow et al., developed a classification and regression tree for risk stratification of patients hospitalized with the Acute Decompensated Heart Failure (ADHF). Their experimental results revealed that ADHF patients at low-, medium-, and high-risk for in-hospital mortality can be easily identified using vital sign and laboratory data obtained on hospital admission [22]. Wang et al., extended the linear regression model for risk stratification that provides clinicians with not only the accurate assessment of a patient's risk but also the clinical context to be acted upon [7]. It must mention that nearly all studies along this line have been estimated using a small hand-picked subset of features from highly stratified patient cohorts. As a result, they merely account for a small number of predetermined risk factors and are fragmented where the conclusion only holds under well-controlled conditions [13].

With the widely adoption of electronic health record (EHR) in healthcare organizations, more advanced machine learning and data mining algorithms were introduced into risk stratification [6,13,19,20]. EHR typically contains a diverse set of information types, including patient demographics, symptoms, vital signs, laboratory tests and treatments, etc., which provides a comprehensive source for risk stratification. Many machines learning techniques, such as decision trees [12], Bayesian network [13], and fuzzy inference system [16], have been proposed to explore the huge potentials of EHR data for risk stratification applications. For example, Karaolis et al., carried out data mining analysis using the C4.5 decision tree algorithm to assess the risk factors of coronary heart events [12]. Bandyopadhyay et al., presented a machine learning approach based on Bayesian networks trained on EHR data to predict the probability of cardiovascular events [13]. Liu et al., present an intelligent scoring system for risk stratification of chest pain patients. In particular, they adopted a hybrid sampling-based ensemble learning strategy to handle EHR data imbalance problem [11]. Singh et al., evaluated three different approaches that use machine learning to build predictive models using temporal EHR data of patients with compromised kidney function. Their