Journal of Biomedical Informatics 53 (2015) 147–155

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Towards actionable risk stratification: A bilinear approach

Xiang Wang *, Fei Wang, Jianying Hu, Robert Sorrentino

IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

ARTICLE INFO

Article history:
Received 29 April 2014
Accepted 7 October 2014
Available online 14 October 2014

Keywords:
Risk stratification
Bilinear model
Logistic regression
Matrix factorization
Dimensionality reduction

ABSTRACT

Risk stratification is instrumental to modern clinical decision support systems. Comprehensive risk stratification should be able to provide the clinicians with not only the accurate assessment of a patient's risk but also the clinical context to be acted upon. However, existing risk stratification techniques mainly focus on predicting the risk score for individual patients; at the cohort level, they offer little insight beyond a flat score-based segmentation. This essentially reduces a patient to a score and thus removes him/her from his/her clinical context. To address this limitation, in this paper we propose a bilinear model for risk stratification that simultaneously captures the three key aspects of risk stratification: (1) it predicts the risk of each individual patient; (2) it stratifies the patient cohort based on not only the risk score but also the clinical characteristics; and (3) it embeds all patients into clinical contexts with clear interpretation. We apply our model to a cohort of 4977 patients, 1127 among which were diagnosed with Congestive Heart Failure (CHF). We demonstrate that our model cannot only accurately predict the onset risk of CHF but also provide rich and actionable clinical insights into the patient cohort.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Risk stratification is indispensable to modern clinical decision support systems. By providing the clinicians (or other healthcare practitioners) an assessment of an individual's risk against an adverse outcome, risk stratification plays a central role in personalized medicine, care plan management, and cost estimation [1]. While the application area of risk stratification is broad, generally speaking, a comprehensive risk stratification method has three fundamental goals:

1. **Risk Score Prediction**: Given the clinical features, predict an individual's risk against a certain adverse outcome, such as disease onset, hospitalization, and mortality.
2. **Patient Cohort Stratification**: Segment a patient cohort into coherent groups based on the patients' risk as well as clinical characteristics.
3. **Clinical Context Discovery**: Identify the clinical contexts that underpin the patients' risk assessment.

The vast majority of existing risk stratification techniques are based on multivariate regression analysis [2–6], especially linear regression and logistic regression. Given a set of training patients and their clinical features, the regression model is fit to the training data such that the contribution of each individual clinical feature (also called risk factors) to the overall risk can be estimated (the regression coefficient). The trained model is then applied to a group of test patients to compute their overall risk scores. Based on the their risk scores, the patient cohort can be stratified into several tiers, e.g. high, medium, and low risk.

Given sufficient amounts of training data [7], existing regression models can accurately predict the risk scores for individual patients (Goal 1 above). However, they offer limited insights when it comes to patient cohort stratification (Goal 2) and clinical context discovery (Goal 3). Imagine we have a cohort of patients who are at risk of Congestive Heart Failure (CHF). As illustrated in Fig. 1,[1] traditional regression model will be able to identify two high-risk individuals (red dots). Based on the fact that they have similar risk scores, these two patients will be stratified into the same group regardless of their clinical conditions. In fact, these two individuals may have high risks of CHF for very different reasons, e.g. one with Chronic Obstructive Pulmonary Disease (COPD) and the other with Chronic Kidney Disease (CKD). These clinical contexts are crucial when the clinicians wish to act upon the risk stratification results, e.g. to devise personalized treatment plan, yet they are not adequately addressed by existing regression models.

* Corresponding author.
E-mail addresses: wangxi@us.ibm.com (X. Wang), fwang@us.ibm.com (F. Wang), jyhu@us.ibm.com (J. Hu), sorrentino@us.ibm.com (R. Sorrentino).

[1] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.

http://dx.doi.org/10.1016/j.jbi.2014.10.004
1532-0464/© 2014 Elsevier Inc. All rights reserved.

In this work we would like to address the limitation of existing risk stratification techniques by proposing a novel bilinear risk stratification model. Our model aims to perform a comprehensive risk analysis of a given patient cohort, from which clinicians and healthcare practitioners can derive more actionable insights. Our model is designed to achieve all the three above stated goals in a principled and integrated fashion. Specifically, our model learns an embedding of the patients into a low-dimensional risk space such that: (1) the distance from a patient to the origin becomes a measure of his/her risk (risk prediction); (2) patients with similar risk scores and clinical characteristics are close together in the space (cohort stratification); and (3) each dimension of the space provides an interpretation of the clinical context (context discovery). Therefore our model is able to give clinicians a full picture of not only the individual patients' risk but also the distinct phenotypes associated with their risk; not only the contribution of individual clinical features but the clinical contexts they collectively define.

We used a CHF patient cohort extracted from a real Electronic Health Record (EHR) database to test our model. The cohort consists of 1127 case patients who were confirmed with CHF and 3850 control patients. We extracted both diagnosis codes and medication as clinical features. We applied our model to risk stratify this cohort and predicted the risk of future CHF onset. We compared the results from our model to that of logistic regression and demonstrated that our model not only achieved better prediction accuracy but also provided rich and actionable clinical insights that were missing in traditional methods.

2. Background

Risk stratification is a fundamental technique for medical informatics [1]. Traditionally, the goal of risk stratification was to regress the risk of patients based on a pre-selected set of risk factors [5,8]. Once the risk score is predicted, the patient will be assigned to a certain tier based on the score. After EHR has been widely adopted, more advanced machine learning algorithms were introduced into risk stratification to deal with the high dimensionality and sparseness of EHR data [9]. These techniques were able to achieve automatic feature selection and significantly improved the accuracy of risk prediction, but their outputs remained a flat score-based stratification and offered little new insights into the clinical characteristics of the patient cohort.

In a separated line of research, a variety of techniques have been proposed for cluster analysis of the patient cohort or disease phenotyping [10–12]. These techniques were able to discover homogeneous patient groups who have similar medical conditions as well as strongly correlated medical features. This type of analysis provided meaningful clinical contexts which medical experts can act upon. However, these disease phenotyping techniques were not integrated into the risk stratification framework, i.e. there was no direct correspondence between the identified disease phenotypes and the predicted risk of each individual patient.

In this paper we propose a novel bilinear model that integrates risk prediction and patient cohort analysis. Our model can be interpreted both from a regression point of view and an embedding point of view. From the regression perspective, our work is related to Bilinear Logistic Regression [13,14], which was recently proposed and applied to brain imaging analysis. The key difference is that for Bilinear Logistic Regression, each data instance is naturally represented by a matrix. The bilinear regressor introduced actually consists of two vectors, whose goal is to reduce the number of coefficients to be estimated. As a contrast, in our model each data instance (a patient) is still represented by a vector whereas the bilinear regressor is actually a matrix that captures the correlation between the medical features.

From the embedding perspective, our work is related to supervised dimensionality reduction and discriminant analysis, such as Fisher Linear Discriminant Analysis [15] and its extensions [16], supervised Principal Component Analysis [17], and supervised metric learning [18]. What our model and these previous techniques have in common is that they aim to find a projection/embedding that maximally separates the training samples from different classes. The fundamental difference is that existing supervised dimensionality reduction methods are symmetric, i.e. their objective is to pull data points from different classes apart from each other, without making distinctions between the actually class labels. In contrast, our model deals with a binary outcome and the two classes are not interchangeable. Our objective is to project the positive class as far away from the origin as possible while projecting the negative class as close to the origin as possible.

3. Methods

In this section we formally introduce our objective function, the algorithm, and also discuss some implementation issues in practice.

Suppose we can use a d-dimensional vector, $\mathbf{x} \in \mathbb{R}^d$, to encode the clinical records of a patient. A variety of encoding schemes can be used here, for instance, $\mathbf{x}_i = 1 (i = 1, 2, \dots, d)$ means the patient has feature i (e.g. a diagnosis code or a drug), 0 otherwise; alternatively, $\mathbf{x}_i$ could be the frequency count or weighted frequency count of feature i on this particular patient. Let y be the binary label associated with the outcome we want to predict: $y = +1$ means the patient had a certain outcome (case) and -1 otherwise (control). In traditional logistic regression, the risk for the outcome and the input feature vector are associated in a (generalized) linear form [19]:

$$\log \frac{p(y = +1|\mathbf{x})}{p(y = -1|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + \omega_0, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the regression coefficient vector.

In our work, we extend the linear regression model in Eq. (1) to the following bilinear form:

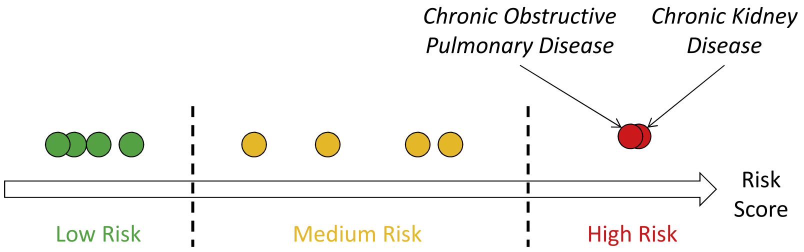

Fig. 1. The limitation of the existing regression-based risk stratification techniques. The two high-risk patients have very different clinical conditions but are regarded similar under the score-based stratification.

Download English Version:

https://daneshyari.com/en/article/6928255

Download Persian Version:

https://daneshyari.com/article/6928255

Daneshyari.com