Computers in Biology and Medicine 73 (2016) 165–172

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A flexible data-driven comorbidity feature extraction framework

Costas Sideris*, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh

Department of Computer Science, University of California, Los Angeles, United States

ARTICLE INFO

Article history:
Received 15 December 2015
Received in revised form
26 March 2016
Accepted 19 April 2016

Keywords:
Knowledge discovery
Comorbidity
Clustering
Prediction

ABSTRACT

Disease and symptom diagnostic codes are a valuable resource for classifying and predicting patient outcomes. In this paper, we propose a novel methodology for utilizing disease diagnostic information in a predictive machine learning framework. Our methodology relies on a novel, clustering-based feature extraction framework using disease diagnostic information. To reduce the data dimensionality, we identify disease clusters using co-occurrence statistics. We optimize the number of generated clusters in the training set and then utilize these clusters as features to predict patient severity of condition and patient readmission risk. We build our clustering and feature extraction algorithm using the 2012 National Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP) which contains 7 million hospital discharge records and ICD-9-CM codes. The proposed framework is tested on Ronald Reagan UCLA Medical Center Electronic Health Records (EHR) from 3041 Congestive Heart Failure (CHF) patients and the UCI 130-US diabetes dataset that includes admissions from 69,980 diabetic patients. We compare our cluster-based feature set with the commonly used comorbidity frameworks including Charlson's index, Elixhauser's comorbidities and their variations. The proposed approach was shown to have significant gains between 10.7–22.1% in predictive accuracy for CHF severity of condition prediction and 4.65–5.75% in diabetes readmission prediction.

1. Introduction

Disease and symptom diagnostic codes are a valuable resource for classifying and predicting clinical outcomes. To provide a common framework for reporting, billing and research, the medical community has adopted the International Statistical Classification of Diseases (ICD) standard. ICD encodes diseases, symptoms, complaints as well as injuries and causes of accidents. ICD was designed with a hierarchical structure to map conditions to corresponding categories based on similarity, and is periodically revised to incorporate new medical findings.

The ICD-9-CM (Clinical Modification – Version 9) in use in the US health system contains more than 15,000 different disease and symptom diagnostic codes. This number is even higher in the newer ICD-10 coding scheme. Considering each of the diagnostic codes as a separate feature in a predictive framework can be computationally costly. It is also complicated due to the high variability in disease occurrence frequencies.

To allow doctors to quantify risk of patient mortality and other outcomes, several comorbidity schemes have been proposed. The two most commonly used comorbidity frameworks are Charlson's score and Elixhauser's comorbidities and their variations. They were designed to improve prediction of specific outcomes such as patient mortality and hospital charges/length of stay. Both of these schemes along with their variations have been extensively studied, but focus only on specific conditions and outcomes.

1.1. *Comorbidity frameworks*

1.1.1. *Charlson index*

Charlson's comorbidity index was first developed in 1987 [1] with the goal of providing a scoring system for patient mortality risk. It is used to predict ten-year mortality for patients based on a range of comorbid conditions such as AIDS, Cancer and Heart Failure. Twenty two conditions are considered in the Charlson Index and are weighted based on their severity with a score of 1, 2, 3, or 6. Charlson's index is nowadays most commonly calculated on the basis of the Quan revision of Deyo's ICD-9 mapping [2].

1.1.2. *Elixhauser comorbidities*

Elixhauser's comorbidity [3] measure was developed based on administrative data from the State of California. It takes into

Abbreviation: RMS, Remote health monitoring system; EHR, Electronic health records; ICD, International classification of diseases; ICD-9-CM, ICD version 9, clinical modification; LOS, Length of stay; HR, Heart rate; BP, Blood pressure; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; CHF, Congestive heart failure; NIS, National impatient sample; HCUP, Healthcare cost and utilization project; NLP, Natural language processing; UCLA, University of California, Los Angeles; UCI, University of California, Irvine; ROC, Receiver operating characteristic

* Corresponding author.
E-mail address: ksideris@ucla.edu (C. Sideris).

http://dx.doi.org/10.1016/j.compbiomed.2016.04.014
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

consideration a list of 30 comorbidities based on the ICD-9-CM codes. Since the comorbidities affect length of stay, hospital charges, and mortality differently, a unified index or score was not developed. The 30 comorbidities are also calculated on the basis of Quan's revision [2]. Another commonly used revision of Elixhauser's comorbidities was presented from the Agency for Healthcare Research and Quality (Elixhauser-AHRQ) [2]. Van Walraven et al. [4] presented a unified score for Elixhauser's comorbidities for predicting hospital death.

For a comparison of these measures and their variations in the context of predicting inpatient death, in-hospital adverse events, and readmission risk, we refer the reader to the review papers of Southern et al. [5], Farley et al. [6] and Sharabiani et al. [7].

In addition to these schemes, with the introduction and proliferation of Electronic Health Records (EHR) several methods have been proposed to identify risk factors by mining administrative data.

1.2. EHR mining methods

EHR data mining has been explored systematically to discover disease related information such as correlations, drug efficiency and other findings. Roque et al. [8] describe a framework to discover disease correlations through co-occurrence and map them to biological frameworks. Bauer-Mehren et al. [9] created a network from unstructured EHR to examine the efficiency of certain treatments and identify patient cohorts. Yao et al. [10] examined EHR mining in the context of drug efficiency/side-effect analysis. Baneyx et al. [11] built an ontology of pulmonary diseases from EHR using Natural Language Processing (NLP). In a prior effort [12], we described a framework for modeling the severity of condition for hospitalized Congestive Heart Failure (CHF) patients. Other notable efforts include Chen et al. [13], Rindflesch et al. [14] and Cao et al. [15,16] that focus on disease meta-information mining. Jensen et al. [17] discussed the current status and challenges in mining EHRs. A review of recent approaches is provided by [18].

To enable the effective utilization of disease and symptom information in a general classification scheme we propose a flexible, data-driven feature extraction scheme from ICD-9-CM diagnostic codes that can easily adapt to different classification tasks. We examine the efficiency of our scheme in the context of quantifying and predicting CHF patient risk. More specifically, this paper makes the following contributions:

- Designing and developing a flexible, data-driven approach for feature extraction from disease diagnostic information.
- Developing classification schemes for categorizing patients into high and low-risk cohorts.

Section 2 presents the different steps of our methodology and defines six methods for modeling severity of condition for CHF patients. Section 3 demonstrates the classification accuracy gains of the proposed methodology vs the commonly used Charlson Index, Elixhauser's Comorbidity and AHRQ schemes and Section 4 provides a summary of our effort.

A preliminary version of this work has been reported in [19].

2. Methodology

2.1. Feature extraction from ICD-9-CM codes

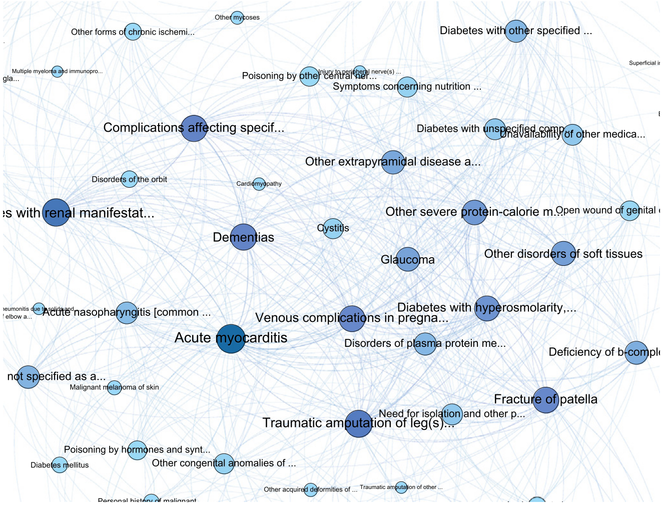
Other mycoses
Other forms of chronic ischemi...
Diabetes with other specified ...
Multiple myeloma and immunopro...
gla...
Superficial i
Poisoning by other central ner...
Injury to peripheral nerve(s) ...
Symptoms concerning nutrition ...
Complications affecting specif...
Diabetes with unspecified comp
Unavailability of other medica...
Other extrapyramidal disease a...
Disorders of the orbit
Cardiomyopathy
es with renal manifestat...
Other severe protein-calorie m...
Open wound of genital
Dementias
Cystitis
Glaucoma
Other disorders of soft tissues
neumonitis due to solids and
elbow a...
Acute nasopharyngitis [common ...
Diabetes with hyperosmolarity,...
Venous complications in pregna...
Acute myocarditis
Disorders of plasma protein me...
Deficiency of b-compl
not specified as a...
Malignant melanoma of skin
Fracture of patella
Need for isolation and other p...
Traumatic amputation of leg(s)...
Poisoning by hormones and synt...
Diabetes mellitus
Other congenital anomalies of ...
Other acquired deformities of ...
Traumatic amputation of other ...
Personal history of malignant

Fig. 1. Comorbidity co-occurrence network. The nodes are scaled and colored according to their degree in the network. Such networks can explain how certain disease groups are more likely than others. They can also be used to analyze similarities between patients, diseases and different treatments.

Our objective was to utilize all the information provided by recorded ICD-9-CM diagnostic codes while ensuring that the resulting dimensionality of the disease features is low enough to be handled efficiently by the classification system. To reduce the dimensionality of diagnostic codes, we first identified disease groups with high frequency of co-occurrence. These co-occurrence frequencies are modeled through the following Jaccard score:

$$S(A, B) = \frac{P(A \cap B)}{P(A \cup B)} \tag{1}$$

This score will be 0 when two codes never appear together (independent) and 1 when they always appear together. The co-occurrence frequencies along with the ICD-9-CM codes form a densely connected network with weighted edges (Fig. 1). Such networks can explain how certain disease groups are more likely than others. They can also be used to analyze similarities between patients, diseases and different treatments.

Using the calculated Jaccard score, we computed the distance matrix of the ICD-9-CM codes with the following distance function:

$$D(A, B) = 1 - S(A, B) \tag{2}$$

Finally, we clustered the ICD-9-CM codes according to their distance using hierarchical clustering based on the minimum variance method [20]. This clustering methodology ensures that the resulting clusters are as disjoint as possible. The clusters are converted into binary features by assigning a “1” if the patient presents a code within that cluster and “0” otherwise. The first 3 disease diagnoses are also encoded as categorical features based on the cluster number they belong.

Depending on the height at which the dendrogram is cut, a different number of disease clusters are generated. Using a large number of such features tends to improve classification accuracy but above a certain number of features over-fitting becomes a concern (Fig. 2). We initialize the search at the maximum height of the dendrogram and we reduce the height (thus increasing the number of features) at every step as long as the classification accuracy on the validation dataset increases. At each step we generate a new set of disease features and calculate the resulting classification accuracy on the validation dataset.

2.2. Data

We obtained EHR from the Ronald Reagan UCLA Medical Center between 2005 and 2009. The dataset consists of patients admitted primarily for CHF and related complications. The dataset includes