Journal of Biomedical Informatics 84 (2018) 103–113

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Patient representation learning and interpretable evaluation using clinical notes

Madhumita Sushil[a,b,*], Simon Šuster[b], Kim Luyckx[a], Walter Daelemans[b]

[a] Antwerp University Hospital, ICT Department, Wilrijkstraat 10, Edegem 2650, Belgium
[b] Computational Linguistics and Psycholinguistics (CLiPS) Research Center, University of Antwerp, Prinsstraat 13, Antwerp 2000, Belgium

ARTICLE INFO

Keywords:
Representation learning
Patient representations
Model interpretability
Natural language processing
Unsupervised learning

ABSTRACT

We have three contributions in this work: 1. We explore the utility of a stacked denoising autoencoder and a paragraph vector model to learn task-independent dense patient representations directly from clinical notes. To analyze if these representations are transferable across tasks, we evaluate them in multiple supervised setups to predict patient mortality, primary diagnostic and procedural category, and gender. We compare their performance with sparse representations obtained from a bag-of-words model. We observe that the learned generalized representations significantly outperform the sparse representations when we have few positive instances to learn from, and there is an absence of strong lexical features. 2. We compare the model performance of the feature set constructed from a bag of words to that obtained from medical concepts. In the latter case, concepts represent problems, treatments, and tests. We find that concept identification does not improve the classification performance. 3. We propose novel techniques to facilitate model interpretability. To understand and interpret the representations, we explore the best encoded features within the patient representations obtained from the autoencoder model. Further, we calculate feature sensitivity across two networks to identify the most significant input features for different classification tasks when we use these pretrained representations as the supervised input. We successfully extract the most influential features for the pipeline using this technique.

1. Introduction

Representation learning refers to learning features of data that can be used by machine learning algorithms for different tasks. Sparse representations, such as a bag of words from textual documents, treat every dimension independently. For example, in one-hot sparse representations, the terms ‘pain’ and ‘ache’ correspond to separate dimensions despite being synonyms of each other. Several techniques exist to model such dependence and reduce sparsity. The generalized or distributed representations learned using these techniques are referred to as low dimensional, or dense data representations. Unsupervised techniques for representation learning have become popular due to their ability to transfer the knowledge from large unlabeled corpora to the tasks with smaller labeled datasets, which can help circumvent the problem of overfitting [1].

Representation learning techniques have been used extensively within and outside the clinical domain to learn the semantics of words, phrases, and documents [2,3]. We apply such techniques to create a patient semantic space by learning dense vector representations at the patient level. In a patient semantic space, “similar” patients should have similar vectors. Patient similarity metrics are widely used in several applications to assist clinical staff. Some examples are finding similar patients for rare diseases [4], identification of patient cohorts for disease subgroups [5], providing personalized treatments [6,7], and predictive modeling tasks such as patient prognosis [8,9] and risk factor identification [10]. The notion of patient similarity is defined differently for different use cases. When it is defined as an ontology-guided distance between specific structured properties of patients such as diseases and treatments, it represents patient relationships corresponding to those properties. For example, if patient similarity is calculated as a hierarchical distance between the primary diagnostic codes of patients in the UMLS®metathesaurus [11], the value represents a diagnostic similarity. When it is defined as an intersection between the sets of blood tests performed on patients, patient similarity maps to blood test similarity. If patient similarity value is 1 for the patients of the same gender and 0 otherwise, groups of similar patients are gender-specific patient cohorts. However, when we calculate similarity between distributed patient representations, the different properties that influence the similarity value are unknown. Within the learned patient representations, we aim to capture similarity on multiple dimensions,

* Corresponding author at: Computational Linguistics and Psycholinguistics (CLiPS) Research Center, University of Antwerp, Prinsstraat 13, Antwerp 2000, Belgium.
E-mail address: madhumita.sushil@outlook.com (M. Sushil).

https://doi.org/10.1016/j.jbi.2018.06.016
Received 16 February 2018; Received in revised form 7 June 2018; Accepted 28 June 2018
Available online 03 July 2018
1532-0464/

such as complaints, diagnoses, procedures performed, etc., which would encapsulate a holistic view of the patients.

In this work, we create unsupervised dense patient representations from clinical notes in the freely available MIMIC-III database [12]. We aim to learn patient representations that can later be used to identify sets of similar patients based on representation similarity. We focus on different techniques to learn patient representations using only textual data. We explore the usage of two neural representation learning architectures—a stacked denoising autoencoder [13], and a paragraph vector architecture [14]—for unsupervised learning. We then transfer the representations learned from the complete patient space to different supervised tasks, with an aim to generalize better on the tasks for which we have limited labeled data.

Dense representations can capture semantics, but at a loss of interpretability. Yet, it is critical to understand model behavior when statistical outputs influence clinical decisions [15]. We take a step towards bridging this gap by proposing different techniques to interpret the information encoded in the patient vectors, and to extract the features that most influence the classification output when these representations are used as the input.

2. Related work

Dense representations of words [16–19] and documents [14,20] have become popular because they are learned using unsupervised techniques, they capture the semantics in the content, and they generalize well across multiple tasks and domains. An *autoencoder* learns the data distribution and the corresponding dense representations in the process of first encoding data into an intermediate form and then decoding it. Miotto et al. [21] first proposed the use of a stacked denoising autoencoder to learn patient representations. They have shown promising results when patient vectors are first learned by a stacked denoising autoencoder from structured data combined with 300 topics from unstructured data, and are then used with Random Forests classifiers to identify future disease categories of patients. Following their work, Dubois et al. [22] have proposed two techniques to obtain patient representations from clinical notes. The first technique is unsupervised and performs an aggregation of concept embeddings into note and patient level representations, known as 'embed-and-aggregate'. The second technique uses a recurrent neural network (RNN) with a bag-of-concepts representation of patient notes as time steps. The RNN is trained to predict disease categories of patients. The representations learned in this supervised setup are then transferred to other tasks. Apart from these works, Suresh et al. [23] have performed a preliminary exploration of the use of sequence-to-sequence autoencoders to induce patient phenotypes using structured time-series data. They have compared different autoencoder architectures based on their reconstruction error when they are trained to encode patient phenotypes. An application of these phenotypes to different clinical prediction tasks has been reserved for future work. In the same vein as these previous works, we investigate the applicability of a stacked denoising autoencoder to learn patient representations *directly from unstructured data*, and analyze the tasks that these representations can be successfully applied to.

One of the evaluation tasks for us is *patient mortality prediction*. Johnson et al. [24] provide a good overview of the previous approaches for mortality prediction on the MIMIC datasets with an aim of replicating the experiments. Following the work by Ghassemi et al. [25], Grnarova et al. [26] have shown significant improvements for mortality prediction tasks on using a two-level convolutional neural network (CNN) architecture, as compared to the use of topic models and doc2vec representations as inputs to linear support vector machines (SVMs). Besides these works, Jo et al. [27] have recently used long short term memory networks (LSTMs) and topic modeling for mortality prediction. They treat topics for patient notes as time steps for LSTMs. These topics are learned jointly using an encoder network. They have shown performance gains when the topics are jointly learned, compared to those pretrained using LDA [28].

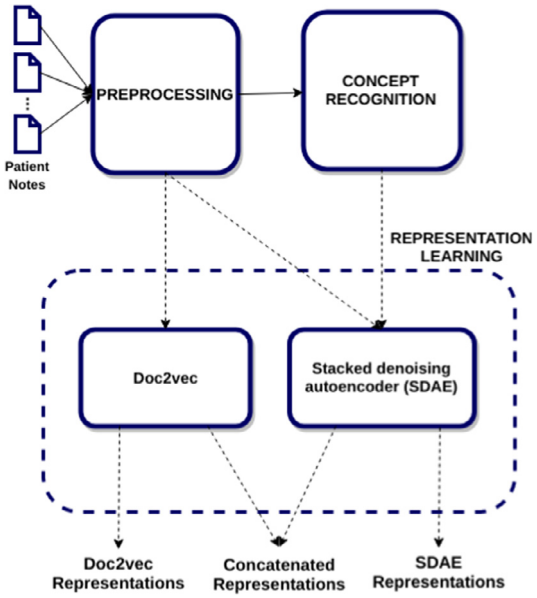

Fig. 1. An overview of the patient representation pipeline. The dashed lines indicate one of several operations, and are not performed in parallel.

3. Methods

3.1. Learning patient representations

In this section, we describe a stacked denoising autoencoder and a paragraph vector architecture doc2vec, in the context of learning task-independent dense patient representations in an unsupervised manner. The corresponding methodology for learning these dense representations is illustrated in Fig. 1.

3.1.1. Stacked denoising autoencoder

Given the previous success of autoencoders for representation learning using structured data with or without topic models learned from unstructured data, we explore the use of a stacked denoising autoencoder (SDAE) [13] to learn task-independent patient representations from raw clinical text, forgoing the use of intermediate techniques like topic modeling. Although the premise of learning patient representations using an SDAE is not novel in itself, our contribution lies in analyzing if such a model is also successful when used only with clinical notes, and if the learned representations can be successfully applied for a range of tasks that are different from patient prognosis. This analysis gives us insight into successful and transferable patient representation architectures for unstructured data.

During the **pretraining** phase, every layer of an SDAE is sequentially trained as an independent denoising autoencoder. An autoencoder learns to first encode the input data I into an intermediate representation R, and then decode R into I. Denoising refers to the process of first adding noise to corrupt the input I into $\tilde{I}$, and then training an autoencoder to reconstruct I using $\tilde{I}$ as the input. We use the dropout noise [29], where a random proportion of the input nodes are set to 0. In the process of denoising, the model also learns the data distribution. In an SDAE, the intermediate representations obtained from the autoencoder at layer $n-1$ are used as the uncorrupted input to the autoencoder at layer n, for all the layers in the SDAE. To pretrain patient representations using an SDAE, high-dimensional (sparse) patient data are used as the input to the autoencoder at the first layer of the SDAE. The intermediate representations obtained from the autoencoder at the final layer are treated as the low-dimensional (dense)