Journal of Biomedical Informatics 74 (2017) 92–103

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Selecting relevant features from the electronic health record for clinical code prediction

Elyne Scheurwegs [a,b,*], Boris Cule [a], Kim Luyckx [c], Léon Luyten [d], Walter Daelemans [b]

[a] University of Antwerp, Advanced Database Research and Modelling Research Group (ADReM), Middelheimlaan 1, B-2020 Antwerp, Belgium
[b] University of Antwerp, Computational Linguistics and Psycholinguistics (CLiPS) Research Center, Lange Winkelstraat 40-42, B-2000 Antwerp, Belgium
[c] Antwerp University Hospital, ICT Department, Wilrijkstraat 10, B-2650 Edegem, Belgium
[d] Antwerp University Hospital, Medical Information Department, Wilrijkstraat 10, B-2650 Edegem, Belgium

ARTICLE INFO

Article history:
Received 17 May 2017
Revised 11 September 2017
Accepted 12 September 2017
Available online 14 September 2017

Keywords:
Feature selection
Data integration
EHR mining
Clinical coding
Data representation

ABSTRACT

A multitude of information sources is present in the electronic health record (EHR), each of which can contain clues to automatically assign diagnosis and procedure codes. These sources however show information overlap and quality differences, which complicates the retrieval of these clues. Through feature selection, a denser representation with a consistent quality and less information overlap can be obtained. We introduce and compare coverage-based feature selection methods, based on confidence and information gain. These approaches were evaluated over a range of medical specialties, with seven different medical specialties for ICD-9-CM code prediction (six at the Antwerp University Hospital and one in the MIMIC-III dataset) and two different medical specialties for ICD-10-CM code prediction. Using confidence coverage to integrate all sources in an EHR shows a consistent improvement in F-measure (49.83% for diagnosis codes on average), both compared with the baseline (44.25% for diagnosis codes on average) and with using the best standalone source (44.41% for diagnosis codes on average). Confidence coverage creates a concise patient stay representation independent of a rigid framework such as UMLS, and contains easily interpretable features. Confidence coverage has several advantages to a baseline setup. In our baseline setup, feature selection was limited to a filter removing features with less than five total occurrences in the trainingset. Prediction results improved consistently when using multiple heterogeneous sources to predict clinical codes, while reducing the number of features and the processing time.

1. Introduction

Electronic health records (EHRs) contain different types of information about patients and their stays in health facilities [1]. Clinical codes reflect diagnoses and procedures related to a patient stay and are primarily assigned for reporting and reimbursement purposes. Their widespread adoption in hospitals makes them a viable information source in research and monitoring applications [2,3].

Clinical codes are predefined in a classification system (e.g., ICD [4], ICPC [5]) and are assigned to a patient stay by clinical coders who analyse the patient's medical information. With the recent transition to ICD-10-CM/PCS, the coding complexity grew since there are up to nineteen times as many procedure codes and five times as many diagnosis codes. This increased complexity requires better techniques to assist clinical coders. Existing (proprietary) applications either focus on making the code system easily browsable or offer computer-assisted coding. Most of those applications operate on English data.

Computer-assisted coding helps clinical coders by pointing out relevant information, suggesting codes, or in simple cases automatically assigning codes without manual review [6]. Most systems described by Stanfill et al. are used in controlled settings (e.g., assuming a very specific notation format in medical data), predict only a limited set of codes, and are mainly based on information from discharge summaries or radiology reports [7]. While some of these approaches perform well, they are difficult to scale to larger datasets or port to different environments. Recent research moved on to using real-world data such as the MIMIC-III dataset [8]. In this paper, we propose several feature selection methods to assist an automatic coding algorithm, trained on multiple data sources. These methods are used to reduce redundancy between features extracted from different information sources

* Corresponding author at: University of Antwerp, Advanced Database Research and Modelling Research Group (ADReM), Middelheimlaan 1, B-2020 Antwerp, Belgium.
E-mail address: elyne.scheurwegs@uantwerpen.be (E. Scheurwegs).

http://dx.doi.org/10.1016/j.jbi.2017.09.004
1532-0464/

and thus create a denser representation of the information with minimal loss of quality.

1.1. Background

Recently, two techniques were proposed to leverage the sparsity of the output codes with layered prediction and to make classification on a complex, real-world dataset more effective [9,10]. For evaluation of new techniques, the MIMIC dataset is often used as a benchmark [8]. Perotte et al. predict diagnosis codes for MIMIC-II by using the code hierarchy to learn codes incrementally: they initially predict higher-level diagnosis categories (the first n digits of the code) and then predict the complete codes [9]. Subotin et al. predict ICD-10-PCS codes from a large set of discharge summaries by predicting parts of the code, since digits in ICD-10-PCS point to different properties of the procedure [10].

These approaches still depend on the availability of discharge summaries keeping a uniform, descriptive, and complete notation. However, information is often missing from discharge summaries and most hospitals (and clinicians) have their own formatting and writing style. By supplementing the information found in discharge summaries with information found in other data sources, the missing information can be compensated for. This technique is already being used for clinical tasks such as identifying patient cohorts [11] and high-throughput phenotyping for patient cohort identification [3,12].

Pathak et al. mapped structured and unstructured data onto standardized thesauri (e.g., UMLS), resulting in a unified data view [12,13]. While this approach can be powerful, it also creates a strong dependency on the ontologies used and the completeness of the mapping. Mapping local ontologies (e.g., RIZIV [14]) or terms found in (Dutch) clinical notes, requires substantial research effort. Scheurwegs et al. performed early and late data integration with structured sources directly represented as features and unstructured sources converted into features through a bag-of-words representation [15]. Due to the difference in feature information density between sources - a bag-of-words representation has more and weaker features than a representation of structured sources - an early data integration approach did not work. Feeding the predictions per data source into a meta classifier proved to be a better approach. However, due to the compression of each source to a single data point for each class, a lot of information was lost.

In a recent article, a deep neural network, consisting of stacked autoencoders, was created to represent patients as a dense vector (called 'deep patient') [16]. The structured and unstructured sources used were generalized during preprocessing (e.g., notes are represented in 300 generalized categories, using topic modeling [17]). The resulting representation is claimed to be applicable to a variety of medical applications. Rather than transforming all data into a dense representation, this paper reports on research where the most interesting features are selected. This influences the usability of the algorithm: we optimize for the clinical coding task, whereas 'deep patient' will generate a general-purpose representation. We expect our approach to focus on specific information with strong correlations to a particular clinical code, while capturing the information that may have been removed from the dense vector representation.

Apart from getting complementary information from different sources and balancing the informativeness, a dense representation, created by either a feature selection algorithm or an algorithm that creates an entirely new feature space, can make it easier to deal with sources that are more noisy for some specialties when compared to a late data integration approach [15]. The latter would disregard the source entirely, while a dense representation would still be able to retain a few features originating from that source. Since the quality and amount of information found in certain sources differs for all specialties [15], an approach that learns from interesting features from each source without knowing the most informative sources beforehand makes the algorithm generalisable over multiple datasets.

Feature selection algorithms often improve algorithm performance, particularly in situations where training data is limited [18,19]. In classification, successful feature selection approaches range from techniques optimizing the goodness-of-fit towards a class by ranking features to techniques that specifically look for a minimal set of features by reducing inner-feature redundancies [20]. In a multi-label problem, we have a multitude of labels for which we want to optimize goodness-of-fit, while we also suffer from redundancies in our feature set. Ideally, we want to optimize both goodness-of-fit and inner-feature redundancy, as the goodness-of-fit can help us determine which of the strongly correlated features optimally represents the task at hand, while solely using goodness-of-fit leads to the selection of features that might be strongly correlated and do not contribute to the final prediction [21].

To tackle both, techniques such as mRMR [22] have been introduced, but computationally they do not perform well on datasets with a large number of features and a large number of classes, due to a pairwise feature-based redundancy calculation, and a separate calculation of these metrics for each label. Other techniques used for feature selection, such as markov-blanket algorithms [23,24], yield good results for feature selection, at the cost of computational efficiency, as they require the induction of a (partial) Bayesian network.

Database coverage-based algorithms gain a computational advantage over pairwise feature-based methods that compare redundancy, since the former only require one comparison per feature with the joint coverage of all previously selected features. This does require the usage of a ranking of features, which is not always required for algorithms that identify redundant features in a pairwise fashion.

Using the entropy of a feature on the entire dataset on a scoring mechanism for feature relevance can also be problematic due to the influence of both sparse features and classes in EHRs. We investigate these issues by comparing techniques using different measures for the informativeness of a feature and by using a technique that considers redundancy of features based on the dataset coverage of all previously selected features for a certain class instead of doing a pairwise comparison between features. We monitored the influence this has on process time as well.

1.2. Significance

In this paper, we show that complementary information can be extracted from multiple sources efficiently by selecting the most reliable information for predicting diagnosis and procedure codes in an early data integration approach. We introduce confidence-coverage as a feature selection method that uses a co-occurrence based scoring mechanism, combined with an instance-oriented selection designed to reduce information overlap between different extracted features. We evaluate performance of four feature selection methods on several datasets of the Antwerp University Hospital (UZA), six of which have ICD-9-CM (procedural and diagnosis) codes and two of which have ICD-10-CM and ICD-10-PCS codes assigned. For benchmarking purposes, we also evaluate on the MIMIC-III dataset.

2. Materials and methods

To prepare the raw data, present in both textual and structured sources, a representation of each source is created. These