Knowledge-Based Systems 99 (2016) 168–182

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records

Cheng Li*, Santu Rana, Dinh Phung, Svetha Venkatesh

Center of Pattern Recognition and Data Analytics, Deakin University, Australia

ARTICLE INFO

Article history:
Received 6 August 2015
Revised 12 November 2015
Accepted 6 February 2016
Available online 3 March 2016

Keywords:
Bayesian nonparametric models
Correspondence models
Word distances
Disease topics
Readmission prediction
Procedure codes prediction

ABSTRACT

Electronic Medical Record (EMR) has established itself as a valuable resource for large scale analysis of health data. A hospital EMR dataset typically consists of medical records of hospitalized patients. A medical record contains diagnostic information (diagnosis codes), procedures performed (procedure codes) and admission details. Traditional topic models, such as latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP), can be employed to discover disease topics from EMR data by treating patients as documents and diagnosis codes as words. This topic modeling helps to understand the constitution of patient diseases and offers a tool for better planning of treatment. In this paper, we propose a novel and flexible hierarchical Bayesian nonparametric model, the word distance dependent Chinese restaurant franchise (wddCRF), which incorporates word-to-word distances to discover semantically-coherent disease topics. We are motivated by the fact that diagnosis codes are connected in the form of ICD-10 tree structure which presents semantic relationships between codes. We exploit a decay function to incorporate distances between words at the bottom level of wddCRF. Efficient inference is derived for the wddCRF by using MCMC technique. Furthermore, since procedure codes are often correlated with diagnosis codes, we develop the correspondence wddCRF (Corr-wddCRF) to explore conditional relationships of procedure codes for a given disease pattern. Efficient collapsed Gibbs sampling is derived for the Corr-wddCRF. We evaluate the proposed models on two real-world medical datasets – PolyVascular disease and Acute Myocardial Infarction disease. We demonstrate that the Corr-wddCRF model discovers more coherent topics than the Corr-HDP. We also use disease topic proportions as new features and show that using features from the Corr-wddCRF outperforms the baselines on 14-days readmission prediction. Beside these, the prediction for procedure codes based on the Corr-wddCRF also shows considerable accuracy.

1. Introduction

Healthcare data analysis forms the backbone of evidence-based medicine and clinical practices. Medical records in the form of electronic medical records, clinical notes, medical imaging or genomic data can be analyzed to produce evidence-based decision support tools. In recent years, many hospitals and care points have begun to implement electronic recording of health records for patients. Data such as diagnosis information, procedures performed and medications prescribed are recorded for each visit of a patient. Other meta-data such as demographic information and social-economic conditions are also recorded in a typical Electronic Medical Record (EMR) system. Altogether, EMR has established itself as a useful resource for clinical knowledge discovery. It has been used successfully in knowledge discovery, such as intervention-driven prediction models [1], personalized care [2] and medical information retrieval [3].

A related problem, of great interest to clinical communities, aiming to deliver more targeted care, is understanding disease co-morbidties for different cohorts of patients. We use topic modeling for this purpose. A topic model extracts topics from a corpus; as such documents have representations of topic proportions in the new latent semantic space. For EMRs, all the admission records for one patient are collapsed together, thereby allowing the construction of a single document. The diagnosis codes contained in admission records are treated as words. Topic models can be applied naturally over a cohort of patients to discover latent disease topics, i.e. the probability distributions over diagnosis codes. Latent Dirichlet allocation (LDA) [4] is one of the most widely-used topic models. To avoid model selection, its nonparametric version, hierarchical Dirichlet process (HDP) has been proposed in [5]. The HDP has been promising in many applications [6].

* Corresponding author. Tel.: +61 420714881.
E-mail addresses: cheng.l@deakin.edu.au (C. Li), santu.rana@deakin.edu.au (S. Rana), dinh.phung@deakin.edu.au (D. Phung), svetha.venkatesh@deakin.edu.au (S. Venkatesh).

http://dx.doi.org/10.1016/j.knosys.2016.02.005
0950-7051/

A fundamental assumption underlying the HDP is that words among one document are exchangeable [7]. This implies that topic assignments of words are conditionally independent, and are not related to the sequence of words. However, words may depend on each other in a much complex manner. For example, words in one sentence have syntactic or semantic relations; superpixels in one image are correlated with each other spatially. In the case of EMRs, diagnosis codes are assigned based on the ICD-10 [8] hierarchy, which is a hierarchical structure presenting semantic grouping of diseases. Diagnosis codes are represented as tree forms; thus, codes are related to each other through connected edges in a tree (this detail will be discussed Section 2). This kind of tree-structured side information can be used to measure the semantic proximity of diagnosis codes.

Recently, some research has considered word relations to improve the performance of topic models. Andrzejewski et al. [9] took the mixture of the Dirichlet tree distribution as the prior of LDA to encode complex domain knowledge on words. Hu et al. [10] developed interactive topic modeling by introducing constraints between words. Boyd-Graber et al. [11] encoded correlations between synonyms into topic models. The above work employs similarities or constraints between words to discover latent topics. None of them, however, can exploit the tree-structured side information as presented in EMRs. Furthermore, until now, topic models have not seen significant use in knowledge discovery for EMRs.

In this paper, we propose a novel and flexible model towards knowledge discovery for EMRs that exploits the tree-structured side information in a Bayesian nonparametric setting. This side information can be used to measure word distances. The proposed model, termed the *word distance dependent Chinese restaurant franchise* (wddCRF), incorporates word distances at the bottom level of this model. Specifically, we propose the shortest-path strategy to measure the distances between two diagnosis codes. A decay function is further introduced to measure the closeness between codes. We first sample its most possible link for each diagnosis code and then construct local groups of diagnosis codes from connected components of diagnosis codes. We next sample disease topics from groups of diagnosis codes.

We explain the differences between the HDP and the wddCRF by their metaphors. We know that the HDP uses a two-level Dirichlet process prior over observations. The Chinese restaurant franchise (CRF) is one alternative representation for the HDP. It is supposed that dishes (topics) are shared among many Chinese restaurants (documents). Each restaurant consists of an infinite number of tables. For each restaurant, when a new customer (word) arrives, he may pick an occupied table with a probability in proportion to the number of customers already at that table and may also pick an unoccupied table with a non-zero probability. This allocation of tables is performed by the bottom-level Dirichlet process [12]. Then, each table orders a dish (topic) with a probability proportional to the pan-franchise popularity of the dish. A table can also order a new dish with a non-zero probability. The allocation of dishes is performed by the top-level Dirichlet process. For the wddCRF, the allocation of local tables is determined by customer distances and the likelihood of particular customers assigned to tables. The metaphor is changed in the way that a new customer selects a table at which to sit. A new customer (word) now selects a table with a probability that is proportional to both the strength of the table and also the friendliness (distance) of the occupants (other words from the table). The change in the table partitions in turn affects the global dish distributions at the top level and makes the dish distributions be correlated with word distances.

The wddCRF model discovers disease topics over diagnosis codes. In EMRs, procedure codes are also recorded along with diagnosis codes. Diagnosis codes and procedure codes construct pairs of data streams, and by nature they are correlated. We can exploit this correlation to predict procedures for a given patient with a set of diagnosis codes. Potentially, this information can be used for better hospital resource planning. This can be achieved by finding conditional relationships between latent variables of diagnosis codes and procedure codes. The correspondence topic models [13,14] were initially proposed to match image content and captions. We further extend the correspondence version of wddCRF (Corr-wddCRF), where diagnosis codes are generated by following the wddCRF, and procedure codes are generated in correspondence with diagnosis codes. The Corr-wddCRF can discover latent disease topics and procedure topics simultaneously.

We evaluate our Corr-wddCRF on two real-world medical datasets: PolyVascular disease (PolyVD) and Acute Myocardial Infarction disease (AMI). The discovered disease topics are found to be coherent and consistent with the ICD-10 disease relations. Further, we evaluate the efficacy of Corr-wddCRF by using the disease topic compositions as dimensionality-reduced features, which are then used to predict 14-day readmission following hospital discharge. We find that the Corr-wddCRF-based features outperform the Corr-HDP-based features for this prediction task. The experiment results also show that our Corr-wddCRF reveals the conditional relations between diagnosis codes and procedure codes, and performs better than the Corr-HDP when predicting procedures.

Our *main contributions* of this work are:

- Proposal of a novel Bayesian nonparametric wddCRF that allows us to include word distances in the topic model. This is achieved by using a distance-dependent prior at the lower level of the model;
- Proposal of the correspondence wddCRF to explore the conditional relationship between diagnosis codes and procedure codes in an Electronic Medical Record;
- Derivation of the efficient posterior inference for the wddCRF and Corr-wddCRF;
- Demonstration of the superiority of the proposed Corr-wddCRF over the Corr-HDP in terms of the quality of the topics, readmission prediction and procedure prediction. Experiments are performed on two real-world medical datasets by exploiting side information available in terms of a semantic tree structure.

A preliminary version of this work was presented in [15]. In this paper, we have developed a new model the Corr-wddCRF based on the previous wddCRF. Gibbs sampling inference have been derived for the Corr-wddCRF. We have further used the Corr-wddCRF to predict procedure codes for a given diagnosis record. Diagnosis topics and procedure topics can be discovered in one unified framework. The Corr-wddCRF shows the considerable improvement of prediction accuracy over the unconstrained Corr-HDP.

The rest of the paper is organized as follows. Section 2 formulates the novel problem. Section 3 presents some relevant models. Section 4 elaborates the wddCRF and Corr-wddCRF and their inference approaches. Section 5 presents experiments and discusses the results, Section 6 describes related work, and finally, Section 7 concludes this paper.

2. Data description and task formulation

A electronic medical record often contains two primary sets of codes: diagnosis codes and procedure codes. Diagnosis codes are coded disease information assigned by physicians as per the WHO ICD-10AM coding guidelines. Procedure codes are related to the procedure performed to patients during hospital stay. This includes any surgery, pharmacotherapy, imaging performed during admissions. They are coded by following the ACHI[1]

[1] http://www.achi.org.au/