Learning patients embedding from EHR diagnosis data using topic modeling

Xiaoxi Zhao xz2740, Chirong Zhang cz2533

December 20, 2019

Abstract

In this project[1], we aim to learn patient representation in an unsupervised way according to patients previous diagnosis in electronic health records (EHR) data by implementing Latent Dirichlet Allocation (LDA) and Embedded Topic Model (ETM). Learning patient embedding from previous patient records is one of the major way to do personalized medication and can also assist doctors to make diagnosis. We evaluate the embeddings by a downstream task, predicting whether the patient will have Chronic kidney disease (CKD) in the future and compare the unsupervised embedding with supervised embedding obtained from Recurrent neural network (RNN).

1 Introduction

Patients' medical history including results of laboratory test, medications, diagnostic procedures and detailed summary about disease itself have been accumulated throughout the years. These records are documented as electronic health records (EHR). Taking advantage of this tremendous dataset can not only help doctor's diagnosis but also foresee some potential diseases based on the other patients with similar records.

Despite the huge benefit of utilizing these records, there's still some barrier in realizing it. The most important one is the issue about the complicated structure of the data. EHR record very detailed information about the admission, which makes it complicated for people to extract valuable information.

However, the recent NLP technology especially word2vec[6] greatly lower the barrier. If we regard each recorded information as one word, patient as a document, phenotype (the set of observable characteristics of an individual disease) as a topic, we can use a wide range of method from NLP to understanding complicated EHR data.

[1]Project Github Repository: `https://github.com/Trccc/6701Project`.

Our goal is to first learn representations of diagnosis concepts coded by ICD9, then high level representation of diagnosis (phenotypes) across a large set of patients based on topic modelling[1][3]. By combining the above presentation, we can embed the patients into a lower dimension in an unsupervised way. This embedding procedure can serve as an encoded representation of patients and be used in many downstream tasks, such as prediction on heart failure or re-admission, cluster of patients, and providing personalized medication advices to people.

2 Phenotype modeling

The phenotype model proposed in this project is based on the topic model literature[1][3]. This is a fully unsupervised model which learns computational representations of disease based on past patient diagnosis.

2.1 Latent Dirichlet Allocation

As discussed in class, LDA is a method of topic modeling in which each document can be viewed as a mixture of various topics who has a probability distribution on different words. In EHR dataset, we consider the phenotypes as latent topics, patients as documents, and clinical records as words. LDA model serves as a baseline model.

2.2 Embedded Topic Model

Embedded Topic Model [3] is a generative model that marries traditional topic models with word embeddings. In particular, it models each word with a categorical distribution whose natural parameter is the inner product between a word embedding and an embedding of its assigned topic. The application to EHR dataset is the same as what we do in LDA.

3 Results and experiment on patient embedding

3.1 Dataset

We conduct our experiment on MIMIC-III[5](Medical Information Mart for Intensive Care). MIMIC-III is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. As a widely used public EHR database, MIMIC-III includes clinical records such as medications, laboratory measurements, observations, procedure, diagnostic, etc. In this project, we will mainly focus on diagnosis codes in MIMIC-III to learn phenotypes and unsupervised patients embedding.

To evalutate the embedding, we will consider a classification task based on CKD data. This involves data consisting of 780 positive samples and 780 negative samples(no diagnosed CKD) randomly sampled from the whole dataset. Preprocessing detail and sample data (see table4) are available in Appendix.

3.2 Experimental Setup

We first obtain distributed representation of diagnosis by Skipgram with negative sampling[6] on the whole diagnosis dataset. We then marry the diagnosis embedding with three different models, two unsupervised ones: LDA, ETM and one supervised model to compare: RNN, to get the patient representations.

3.2.1 Patient Embedding

For the notation, we denote α as topic embedding, ρ as word embedding, θ as topic distribution within each document and β as word distribution within each topics.

For LDA and ETM, the method to get patient embedding is quite similar. Basically, it is the weighted average of topic embeddings. i.e. d-th patient embedding $p_d = \alpha^T \theta_d$.

In ETM, α is a learned parameter during model training while in LDA, the topic embedding is obtained by weighted average of word embedding in that topics. i.e. k-th topic embedding $\alpha_k = \rho^T \beta$.

In RNN, we used GRU to overcome long term dependency, where diagnosis code is the input. The final hidden layer is extracted as the patient embedding.

3.2.2 Evaluation Criteria

We use perplexity score to determine whether the algorithm converges and calculate the topic coherence (TC)[7] which is a quantitative measure of the interpretability of a topic. We combine TC with topic diversity (TD) which is the percentage of unique words in the top 25 words of all topics. Diversity close to 0 indicates redundant topics while diversity close to 1 indicates more varied topics. We then report the topic quality (TQ), the product of TC and TD. We will also rely on domain knowledge about number of highest level of ICD9 categories (which is 19).

We then use this phenotype model to embed the patients and evaluate the embedding by the accuracy of prediction on whether the patient will have Chronic kidney disease (CKD) in the future using patient embedding by logistic regression.

3.2.3 Results

For LDA representation, we choose topic number as 5 which gives us the highest coherence score. For ETM, we choose topic number as 15, figure1 and figure2 show the coherence scores of different numbers of topics. The convergence plot for each algorithm is attached in Appendix, see figure3 and figure4.

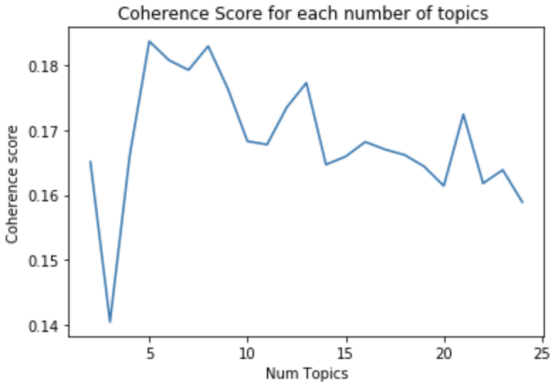

Figure 1: best topic number of LDA

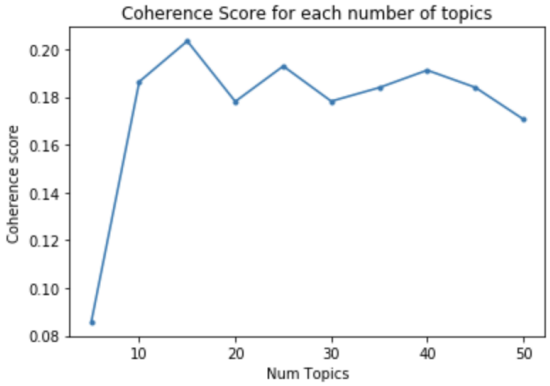

Figure 2: best topic number of ETM

We then compare the performance of LDA and ETM by comparing TC, TD and TQ, see table1. We can see that LDA has a higher topic quality, however, this is mainly because it gives a smaller number of topics thus has a higher topic diversity. On the other hand, ETM gives a topic number of 15, which is closer to our expert consensus (19) and yields a better TC, thus we choose ETM as our final model to phenotypes.

Model	Topic number	TC	TD	TQ
LDA	5	0.19	**0.696**	**0.13**
ETM	**15**	**0.20**	0.552	0.11

Table 1: Phenotype quality under different topic models

Phenotype 1	Phenotype 2	Phenotype 3
Tobacco use disorder	Hypertensive CKD	Severe sepsis
Alcohol cirrhosis liver	Neuropathy in diabetes	Septicemia
Chronic hepatitis C	Congestive heart failure	Septic shock
Thrombocytopenia	End stage renal disease	Acute kidney failure
Cirrhosis of liver	Diabetes with neurological manifestations	Urin tract infection

Table 2: Learned phenotypes and corresponding diagnosis

Table 2 shows three most used phenotypes as well as the most used diagnosis within each phenotype. We can see clearly that the first phenotype is related to liver, the second one to kidney, and the third one to sepsis. Our topic model successfully identifies phenotypes!

After getting the phenotypes and its distribution within each patient, we embed the patients by weighted the average of phenotype embeddings. This embedding is obtained in a purely unsupervised way. We then simply run logistic regression on the the different patient embeddings including the supervised embedding obtained from RNN. The results are shown in table3.

Model	Accuracy
LDA	0.66
ETM	0.73
RNN	0.77

Table 3: CKD prediction by logistic regression

From the logistic regression we can find that all of three methods provide some satisfactory results among which RNN has the best performance as expected since it is able to add bias to the embedding layers. However, RNN is trained in a supervised way and the embedding can only be used for one task, while LDA and ETM methods provide a unsupervised way for patient embedding which can be used for multiple tasks.

4 Discussion

4.1 Conclusion

In this project, we get patient embedding and the phenotype by implementing topic modelling on the patients diagnosis data. We choose the phenotype number generated by ETM to be 15 and used as our final model. And this unsupervised result is used to embed the patients. We used this unsupervised embedding to predict the future CKD of the patient and the unsupervised embedding performed fairly well compared to the supervised embedding.

4.2 Future Work

As the patient health status will change based on the previous status, the temporality of EHR data and disease should be considered as well. At first, we try to model the phenotypes by using dynamic LDA [2] and dynamic ETM[4], but the most of the patients only have one visit, thus one record in MIMIC data, making the temporality hard to model. If we could have more data of patients with multiple visits, then the dynamic topic modelling will make our phenotype more meaningful.

Also, we can use this embedding to perform many other downstream tasks: prediction on temporal outcomes such as re-admission rate, prediction on static outcomes such as the risk of heart failure, cluster of patients, etc.

Finally, in this project, only diagnosis data is used which is only a small part of the whole dataset. In the future we can utilize more sturctured data like lab test, procedures as well as unstructured data like discharge summary(pure text) to accompany our understanding of EHR dataset.

References

[1] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. *23rd International Conference on Machine Learning*, 2006.

[3] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei. Topic modeling in embedding spaces. *arXiv:1907.04907*, 2019.

[4] A. B. Dieng, F. J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. *arXiv:1907.05545*, 2019.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. *Empirical Methods in Natural Language Processing*, page 262–272, 2011.

5 Appendix

5.1 Code

The code is available on https://github.com/Trccc/6701Project.

5.2 Preprocessing detail

To pre-process the dataset, we first find out the good CKD examples.

We define a good CKD example as follow. The patient must have multiple admissions and the CKD diagnosis must appear in the at least second admission so that we can make use of the information before CKD diagnosis to get the representation for downstream tasks.

To be noted, we only use the admission information that is longer one year before the admission with CKD diagnosis. For example, the patient has 4 admissions and has CKD diagnosis in the forth admission. We only take the first two admission's diagnosis as our document if the third admission happens within one year before the forth admission.

5.3 Dataset

Subject ID	Diagnosis	ckd
28	76516,4280,...	0
32	99673,40391,...	1

Table 4: Example Dataset. Subject ID represents unique patient ID. Diagnosis represents the ICD9 code. The length of the code varies from at least two to more than twenty. CKD is the target variable representing whether a patient will have CKD in the future.

5.4 Topic model

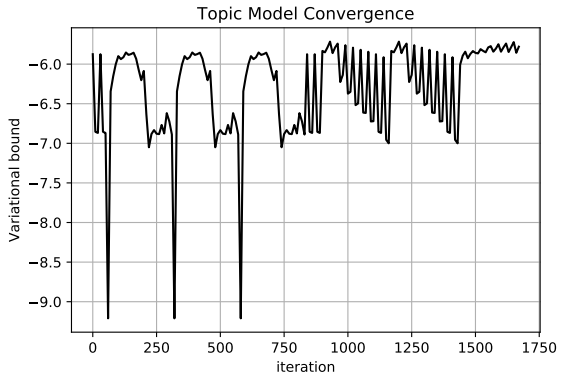

Figure 3: LDA Convergence plot

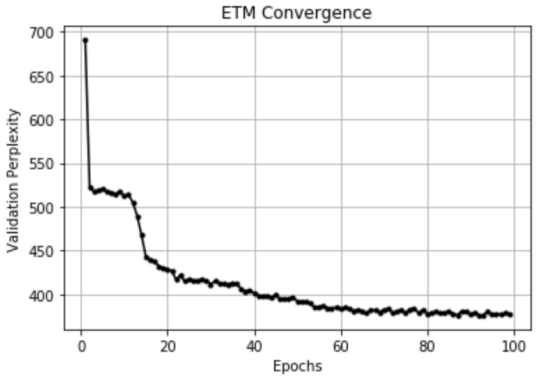

Figure 4: ETM Convergence plot

5.5 RNN structure

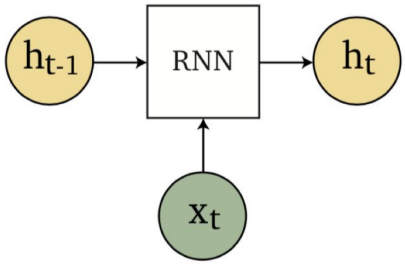

Figure 5: RNN structure