ClinicalKnockoffs: Identification of significant features for Hospital Readmission

Abed Qaddoumi AMQ259@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Seong Woo Han SWH324@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Editor: Rajesh Ranganath

Abstract

Clinical notes contain patients' information and is generated in a massive amount everyday. Clinical notes are difficult to use effectively because they are unstructured. This paper aims to extract crucial information that predicts the rate of readmission of patients in 30 days based on the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset, using the Deep Knockoffs Model (ClinicalKnockoffs). In our study, we show that applying Deep Knockoffs can reveal the important features in clinical notes used by the machine learning model.

1. Introduction

With the rise of using Electronic Health Record (EHR), there has been a rapid increase of machine learning research on the data provided by these records. Due to the improved availability of EHR data, machine learning technique is bringing substantial contribution such as uncovering patterns and improving predictions. Unlike other fields, the medical field needs to understand the underlying reasons for the improvement of prediction. Even though the unstructured, high-dimensional, and sparse information of clinical notes make the data hard to use, it can provide us with insights on how machine learning models are making decisions.

Clinical notes have shown highly predictive powers in tasks such as predicting hospital readmission (Huang et al., 2019) and extracting basic information prediction (Boag et al., 2018). In order to better understand the predictive power of clinical notes, we can create new prediction tasks with these medical values.

Since clinical notes contain knowledge about patients, it can save money and lives (Pedersen et al., 2017). We show in this paper what information in clinical notes is crucial to decide readmissions using the Deep Knockoffs model (Romano et al., 2018). With this model, we extract crucial information inside clinical notes that predict readmission.

2. Related Work

There are numerous work in clinical NLP. Using clinical notes in machine learning can be broadly categorized into text representation and clinical prediction.

Text Representation: We specifically focus on text representation using Bidirectional Transformer (Devlin et al., 2018) that creates an efficient embedding for clinical notes. (Alsentzer et al. 2019) apply BERT on clinical notes and discharge summaries, demonstrating that BERT yields performance improvements on clinical notes. (Lee et al. 2019) apply BERT to biomedical literature and (Huang et al. 2019) apply BERT on clinical notes to find global contextual representation, showing the predictive power of unstructured text. Transformer is a strong fit because each term in the note interacts with every other term, achieving global contextual representation (Devlin et al., 2018).

Clinical Prediction: In this section, we focus on the literature that predicts readmission. (Huang et al. 2019) uses 30-day hospital readmission prediction with discharge summaries and the first few days of notes in the intensive care unit. (Futoma et al. 2015) use various machine learning methods such as random forests and neural networks on hospital readmission tasks. (Xiao et al. 2018b) use topic recurrent neural networks via learning interpretable patient representation and clinical concept embeddings for the readmission task. Similarly, (Rajkomar et al. 2018) predict readmission with Fast Healthcare Interoperability Resources codes from notes, integrated with structured information.

3. Method

We use the Deep Knockoffs model (Romano et al., 2018) because it is able to discover which predictors are important in a response variable together with a large number of potential explanatory variables. Deep Knockoffs is based on Model-X Knockoffs (Candes et al., 2018) which solves the controlled variable selection problem by constructing knockoff variables probabilistically while requiring the covariates to be random with a non-distribution. Our task is to extract the most important information inside the clinical notes that determines readmission. An overview of Deep Knockoffs approach is presented in (figure 1).

The machine receives n samples of the vector X from the clinical note distribution P_X. The machine is defined as a random mapping f_θ that takes as input a random $X \in \mathbb{R}^p$, an independent vector $V \sim \mathcal{N}(0, I) \in \mathbb{R}^p$ and returns an approximate knockoff copy $\tilde{X} =$ f_θ(X,V) $\in \mathbb{R}^p$. A family of machine generating exact knockoffs is given by

$$f_\theta(X, V) = X - X \sum^{-1} \text{diag}\{s\} + (2\text{diag}\{s\} - \text{diag}\{s\} \sum^{-1} \text{diag}\{s\})^{1/2} V \tag{1}$$

for any choice of vectors that keeps the matrix V positive-definite (Candes et al., 2018). Then, the scoring function J examines the empirical distribution of (X,$\tilde{X}$) and quantifies its compliance with the exchangeability properties in the Model-X Knockoffs (Candes et al., 2018). After this process, the machine generates approximate knockoff copies $\tilde{X}$ for new observations of X drawn from the same P_X. Deep Knockoffs model extends the applicability of the knockoffs framework to make it model independent. Compared to the original method in (Candes et al., 2018), the additional computational burden of fitting a neural network is important. We refer readers to (Romano et al., 2018) for a more detailed description.

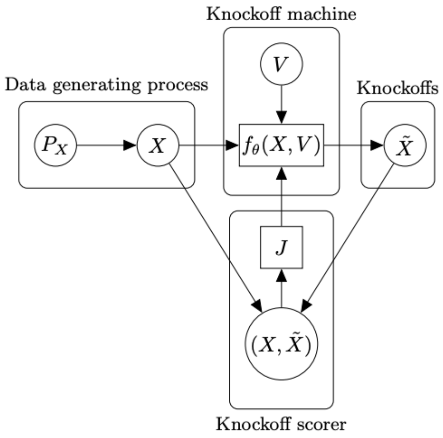

Figure 1. Illustration of Deep Knockoffs machine

4. Experiment Setup

4.1 Dataset

We utilize the MIMIC-III (Medical Information Mart for Intensive Care III) dataset (Johnson et al., 2016), a free hospital database which contains various electronic health records for 58,976 unique hospital admissions from 38,597 patients in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. In this project, we will make use of ADMISSIONS.csv file, a table containing admission and discharge dates, and NOTEEVENTS.csv file, a table containing all notes of each hospitalization.

4.2 Preprocess

We preprocess the data with lowercasing all the words and removing line break, carriage returns, and de-identified brackets. For the ADMISSIONS file, we convert strings to dates and combine the next admission dates with the unplanned re-admissions dates. Then, we calculate the days until the next admission. We merge the discharge notes from NOTEEVENTS file with the results from ADMISSIONS file. We remove the entries with the ADMISSION_TYPE of NEWBORN. Then, we add a column OUTPUT_LABEL to see which patients were readmitted in less than 30 days.

4.3 Training

There are two parts of training in this paper. The first part is to train a logistic regression. We use this model because of its interpretability and fast training time. The second part of the training is to generate a Deep Knockoffs machine. The logistic regression uses 4,184 samples of X with 100 features for its training.

The Deep Knockoffs is trained with 2,092 samples from X and using 100 features. We randomly select 2,092 samples from the 4,184 X training. The model uses the measure pairwise second-order knockoff correlations matrix, which limits the feature space of the training dataset to approximately 100 features. The model trains for 100 epochs with a 100 iteration of the full data per epoch and 0.01 learning rate. We use Multivariate Student distribution to train Deep Knockoffs as the other distribution result with errors for this task. Choosing a multivariate Gaussian or a Gaussian mixture distribution does not generate any results when training the Deep knockoffs machine on it.

5. Result

Logistic regression yields an accuracy of 64% for predicting readmission. This model is sufficient for our experiment when compared to Clinical-BERT, which has accuracy of 76% on the same task (Huang et al., 2019). The Deep Knockoffs model shows four features that reduce the accuracy by more than 0.38%. The four points are shown in (figure 2). The number and the content of features are [17: 'daily', 20: ‘discharge', 56: 'neg', 83: 'status']. When we compare the features discovered by the Deep Knockoffs with the important features used by the logistical regression model, we observe a correlation between the two results. (Figure 3) shows the same features of daily, discharge, neg, and status all included in the top features used for predicting readmission by the logistic regression model.

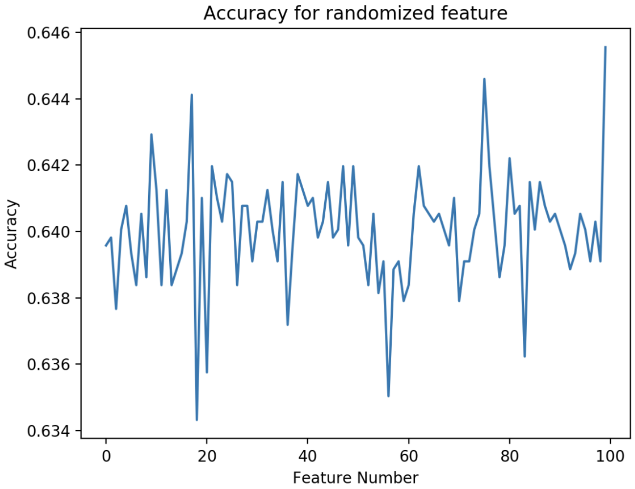

Figure 2. Randomnized feature Accuracy

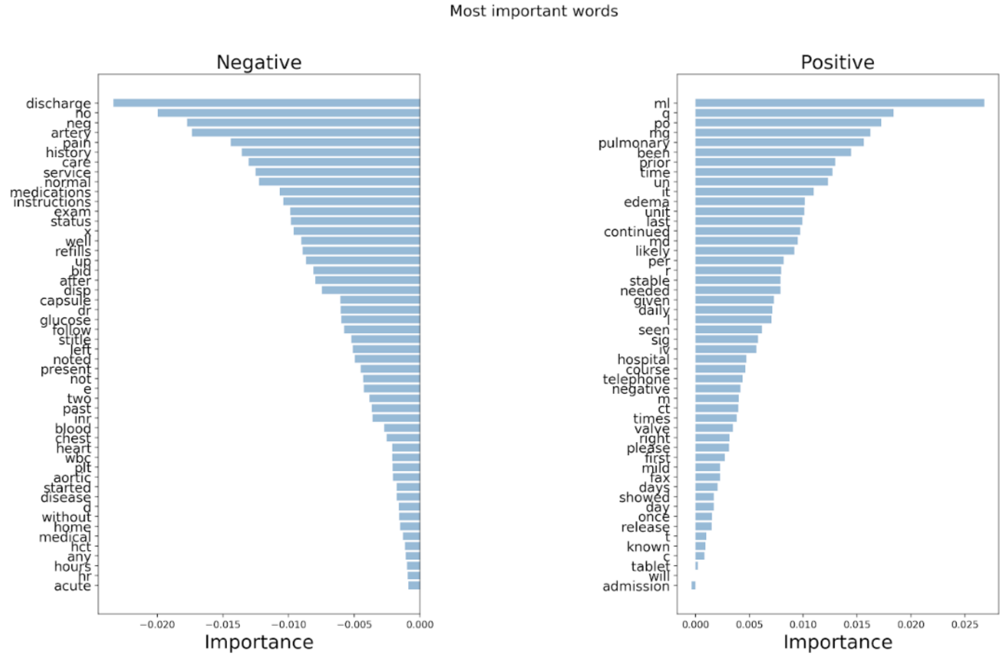

Figure 3. Important words for readmission

6. Discussion

As our model faces serious constraint on using large feature space, we would like to try the followings for future work. We will test the Deep Knockoffs model (Romano et al., 2018) with a larger feature space through batching features together. Then, we will compare our results with the results produced using the same system on the pre-trained model of ClinicalBert (Huang et al., 2019). Another experiment to test is to generate an individual Deep Knockoffs machine for each individual feature and compare the results with the previous results.

Acknowledgments

We would like to acknowledge support for this project from CSCI-GA 2565 Machine Learning class by Professor Rajesh Ranganath and multiple help from Mukund Sudarshan, Mark Goldstein, and Aahlad Manas Puli.

References

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? Unpacking predictive value in clinical note representations. AMIA Joint Summits on Translational Science, 2017:26–34, 05 2018.

E. J. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: model-X knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol., 80(3):551–577, 2018.

Y. Romano, M. Sesia, and E. J Candès. Deep knockoffs. arXiv preprint arXiv:1811.06687, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2018.

Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission." arXiv preprint arXiv:1904.05342 (2019).

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. arXiv:1904.03323, 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv:1901.08746, 2019.

Futoma, Joseph, Jonathan Morris, and Joseph Lucas. "A comparison of models for predicting early hospital readmissions." Journal of biomedical informatics 56 (2015): 229-238.

Cao Xiao, Tengfei Ma, Adji B Dieng, David M Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts. PLOS ONE, 13(4), 2018b.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1):18, 2018.

Rachael B Zuckerman, Steven H Sheingold, E John Orav, Joel Ruhter, and Arnold M Epstein. Readmissions, observation, and the hospital readmissions reduction program. New

England Journal of Medicine, 374(16):1543–1551, 2016.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, 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, 05 2016.

Craig A. Pedersen, Philip J. Schneider, and Douglas J. Scheckelhoff. ASHP national survey of pharmacy practice in hospital settings: Prescribing and transcribing—2016. American Journal of Health-System Pharmacy, 74(17):1336–1352, 2017.