Cancer Treatment Classification with Electronic Medical Health Records (Student Abstract)

Jiaming Zeng,[1] Imon Banerjee,[2] Michael Gensheimer,[3] Daniel Rubin[4]

[1]Department of Management Science and Engineering, Stanford University
[2]Department of Computer Science, Emory University
[3]Department of Biomedical Data Science, Stanford University
[4]Department of Radiation Oncology, Stanford University

Abstract

We built a natural language processing (NLP) language model that can be used to extract cancer treatment information using structured and unstructured electronic medical records (EMR). Our work appears to be the first that combines EMR and NLP for treatment identification.

Introduction

Knowing the sequence of treatments administered to a cancer patient is important for personalized medicine and sequential treatment planning. Our final goal is to leverage the full EMR, including the information available in the clinical notes, to build causal models for treatment effectiveness. For that purpose, we need a sufficiently large dataset with labeled treatment information. However, cancer registries only record the initial line of treatment, even that requires hours of expensive manual labour.

We aim to build a NLP language model that can extract longitudinal treatment information using a combination of structured and unstructured EMR data. Starting with a model trained only on the initial line of treatment, we explore if we can extend the model to extract unrecorded or longitudinal treatment information. The extracted sequential treatments can then be used for future analysis and treatment planning.

Related Works

Automated clinical text classification is an area of growing interest in NLP applications (Wang et al. 2019). Existing literature has shown success in classifying cancer stage information (Warner et al. 2015), disease characteristics (Zhu et al. 2012), and medical subdomains (Weng et al. 2017). Our work appears to be the first work that utilizes NLP for cancer treatment classification from EMRs.

Dataset

We built our dataset from the Stanford Cancer Institute Research Database (SCIRDB), where we selected for patients with localized prostate, esophageal, and oropharynx cancer and diagnosis date after 2008. From SCIRDB, we pulled a total of 4,420 patients with 483,782 clinical notes among the three cancer types. From the California Cancer Registry (CCR), we pulled the initial treatment information available for these selected patients.

Initial Line of Treatment	Entries	Total Notes	Mean Notes
surgery	1,530	28,917	18.9
chemotherapy + radiation	312	18,084	58.0
hormone + radiation	255	5,348	21.0
radiation	239	4,513	18.9
hormone	53	812	15.32

Table 1: Treatment types for classification and summary of available notes.

From the 4,420 patients, we filtered for patients with at least one recorded treatment in the CCR and at least one associated note/encounter in SCIRDB. We classify the initial line of treatment as all treatments performed within 6 months of the first administered treatment and performed classification for the 6 most common treatments found in the CCR. We randomly reserved 10% of the patients from the compiled dataset for testing. Notes associated with these patients were not used for training the embedding models.

After filtering, we compiled a dataset of 2,389 patients to train a treatment classification model. The treatment types selected for classification prediction are summarized in Table 1. For surgery, only patients with surgery at the primary site were selected. For the clinical notes associated with each treatment entry, we compiled the notes into one document by filtering for sentences that contained at least one treatment term. The treatment term dictionary was built from clinical terms in NCI metathesaurus vocabulary and medications curated by experts familiar with our treatments.

In addition to the clinical notes, we also extracted structured data from the information available in the SCIRDB. The features extracted are summarized in Table 2.

Methodology

We trained a machine learning model for treatment classification using combinations of structure and unstructured data and compared the performance. The methodology can be described in two major parts: 1) build a NLP language model using the entire vocabulary, 2) train a machine learn-

Feature	Description
gender	patient gender (M or F)
race	patient race (white, black, asian, other, unknown)
ethnicity	ethnicity of patient (hispanic/latino, non-hispanic/latino, unknown)
age	age range of patient. ($\leq$ 25, 25-50, every 10 years from 50-90, $\geq$ 90)
cancer site	cancer location (prostate, oropharynx, or esophagus)
cancer stage	clinical stage of cancer (stage 0-4, unknown)
hospitalization	count of patient hospitalization (0, 1, 2, $\geq$3)

Table 2: Summary of the structured data.

ing model for treatment classification.

Training a NLP Language Model

For the NLP language model, we began by training with the entire cohort of 483,782 clinical notes (including notes from patients not considered in our training 2,389 dataset) to train an unsupervised doc2vec model (Le and Mikolov 2014). The doc2vec model trained on all available clinical notes are then used to generate embeddings for the unstructured data associated with each entry. Classification accuracy for best performing models are shown in Table 3.

We used gensim library to train the doc2vec models. For the doc2vec models, on the validation set, we tuned the vector size, $vs = [100, 200, 300, 400, 500, 600]$, the learning rate, $\alpha = [0.015, 0.025, 0.035, 0.045, 0.055, 0.065]$, and number of training epochs, $e = [10, 30, 50, 70, 90, 110]$. We trained with the distributed bag of words algorithm, set the number of “noise words” drawn to 5, and ignored all words with total frequency below 10.

Treatment Classification

For the treatment classification, we experimented with using only the structured data and then with both structured and word embeddings of the unstructured data. To combine the structured data and word embeddings, we normalized the structured data to the range of the word embeddings and use linear concatenation to generate a single vector.

Table 3 shows the best performing logistic regression model with the associated parameters. We used scikit-learn to implement the logistic regression model. For the model, we tuned for regularization strength $C \in [1\mathrm{e}{-4}, 1\mathrm{e}4]$, and optimization method, *solver* = [newton-cg, lbfgs, saga, sag] again on the validation.

Experimental Results

Results for the structured data and selected models for the structured and unstructured data are shown in Table 3. The structured data achieved a baseline accuracy of 87.23%. Combining with unstructured data boosted accuracy by as much as 3% to 90.59%.

For doc2vec, vector size did not seem to have a significant effect on the performance. α and epochs have the most effect, with epochs producing a more noticeable effect. For some parameter sets, the word embeddings actually lowered classification performance to as low as 84.10%. For logistic regression, newton-cg was the best performing optimizer.

Dataset	epochs	vector size	alpha	Accuracy
Structured only	-	-	-	87.24%
Structured + doc2vec	10	100	0.025	88.28%
Structured + doc2vec	30	300	0.065	90.59%
Structured + doc2vec	10	400	0.025	89.96%
Structured + doc2vec	10	500	0.025	90.59%
Structured + doc2vec	10	600	0.025	90.59%

Table 3: Testing accuracy from logistic regression.

Future Work

Our preliminary exploration for a treatment classification model shows promising results. In order to create a large treatment dataset, we plan to boost the model performance by evaluating non-linear ML methods, e.g. elastic net, random forest, and SGD boosting.

We plan to do more experimentation with structured and unstructured data separately, and explore more methods for combining. For the structured data, we will expand the features, e.g. ICD9, CPT codes. For the unstructured data, we will explore different ways of collapsing the documents. We also plan to implement a simple bag-of-words classifier as a baseline. For combining the structured and unstructured data, we will also explore other methods such as 1D convolution and other ensemble methods.

Furthermore, our current model aims to produce one model for prostate, oropharynx, and esophagus. We will build separate language models for each cancer type and explore model transfer effects.

References

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196.

Wang, Y.; Sohn, S.; Liu, S.; Shen, F.; Wang, L.; Atkinson, E. J.; Amin, S.; and Liu, H. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making* 19(1):1.

Warner, J. L.; Levy, M. A.; Neuss, M. N.; Warner, J. L.; Levy, M. A.; and Neuss, M. N. 2015. Recap: feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. *Journal of oncology practice* 12(2):157–158.

Weng, W.-H.; Wagholikar, K. B.; McCray, A. T.; Szolovits, P.; and Chueh, H. C. 2017. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making* 17(1):155.

Zhu, H.; Ni, Y.; Cai, P.; Qiu, Z.; and Cao, F. 2012. Automatic extracting of patient-related attributes: disease, age, gender and race. *Studies in health technology and informatics* 180:589–593.