Journal of Biomedical Informatics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Symptom severity classification with gradient tree boosting

Yang Liu [a,*], Yu Gu [b,1], John Chu Nguyen [a], Haodan Li [a], Jiawei Zhang [a], Yuan Gao [a], Yang Huang [a]

[a] Med Data Quest, Inc., 505 Coast Blvd S Ste 300, La Jolla, CA 92037, United States
[b] Department of Electrical and Computer Engineering, UCSD, 9500 Gilman Drive, La Jolla, CA 92093, United States

ARTICLE INFO

Article history:
Received 2 February 2017
Revised 3 May 2017
Accepted 17 May 2017
Available online xxxx

Keywords:
Text classification
Gradient tree boosting
Severity prediction
Bootstrap
Psychiatric evaluation
NLP

ABSTRACT

In this paper, we present our system as submitted in the CEGS N-GRID 2016 task 2 RDoC classification competition. The task was to determine symptom severity (0–3) in a domain for a patient based on the text provided in his/her initial psychiatric evaluation. We first preprocessed the psychiatry notes into a semi-structured questionnaire and transformed the short answers into either numerical, binary, or categorical features. We further trained weak Support Vector Regressors (SVR) for each verbose answer and combined regressors' output with other features to feed into the final gradient tree boosting classifier with resampling of individual notes. Our best submission achieved a macro-averaged Mean Absolute Error of 0.439, which translates to a normalized score of 81.75%.

1. Introduction

The psychiatric clinical evaluation is one of the most challenging types of documentation within the field of medicine. Contributing to the difficulty in understanding psychiatric notes is the sometimes haphazard combination of narrative styles and structured styles (i.e. templates or standardized questionnaires). Additionally, in comparison to other medical specialties, psychiatry emphasizes patient-derived subjective communication that may lead to a disorganized psychiatric interview due to conveying the history in a non-linear, superfluous, confusing, and/or redundant manner. Some common surveys employed for psychiatry include questionnaires listing DSM-5 [1] criteria for mental disorders such as generalized anxiety disorder, major depression disorder and at tention-deficit/hyperactivity disorder.

An ideal psychiatric note should demonstrate internal consistency — the history, physical examination, assessment, and treatment plan should all support each other. An ideal psychiatric note strives for a firm diagnosis for the patient, utilizing all sources of information available, including laboratory results and other medical consults. An ideal psychiatric note should be as specific as possible, such as using "Major depressive disorder, recurrent, severe, currently in partial remission" versus using a vague description of "Major depressive disorder." In practice, however, the rarity of such an ideal note makes it difficult for clinicians to interpret and share psychiatric notes. Therefore, efforts to computationally parse and assess psychiatric notes should aid in properly stratifying psychiatric patients in terms of disorder and severity. Severity classification allows for triaging of patients in order to identify those at acute risk so that prompt medical care can be provided. The implications of not having a severity classifier could prove to be costly in terms of morbidity and mortality, as those with moderate to severe illness may be inadvertently delayed treatment. The high prevalence of psychiatric disorders in a given population inevitably leads to some patients with incomplete or missing diagnoses. Advances in clinical natural language processing (cNLP) and machine learning could efficiently help alleviate adverse outcomes by flagging and noting documentation deficiencies. Furthermore, once patients are properly identified and classified by disorder and severity, subsequent data analysis on patient subgroups could be adequately performed in order to discover optimal treatment strategies.

The Research Domain Criteria (RDoC) is a research framework for new ways of studying mental disorders. It focuses on five psychiatric domains — positive valence, negative valence, cognitive, social processes, and arousal and regulatory systems. Track 2 of the 2016 CEGS N-GRID Shared Task in the Clinical Natural Language focuses on one domain: positive valence [2]. The organizers provided a corpus composed of initial psychiatric evaluation records of patients along with a general severity score (0–3) of

* Corresponding author.
E-mail addresses: yangliu@meddataquest.com (Y. Liu), patrickgu001@gmail.com (Y. Gu), john.nguyen@singleragenomics.com (J.C. Nguyen), haodanli@meddataquest.com (H. Li), jiaweizhang@meddataquest.com (J. Zhang), gary@meddataquest.com (Y. Gao), yanghuang@meddataquest.com (Y. Huang).

[1] This work was done when the second author was on an internship at Med Data Quest, Inc.

http://dx.doi.org/10.1016/j.jbi.2017.05.015
1532-0464/

Please cite this article in press as: Y. Liu et al., Symptom severity classification with gradient tree boosting, J Biomed Inform (2017), http://dx.doi.org/10.1016/j.jbi.2017.05.015

positive valence domain for each note annotated by expert clinicians. The task was to build a system that can automatically predict an overall positive valence severity score based on the patients' psychiatric notes.

A straightforward approach to this challenge is to treat it as a traditional text classification task [3]. However, psychiatric evaluation records are significantly different from common text documents. Psychiatric documents typically contain validated surveys that consist of templates with evaluation questions and answers. The answer to the question can vary from very short to a yes/no response, or it can be verbose in describing illness history. The simple "bag of words" model can neither associate questions with the corresponding answers, nor handle different answers appropriately based on its property. Another disadvantage of the traditional text classifier is that the model is a "black box" that takes the input of tens of thousands of features, which is very hard for clinicians to interpret and validate [4].

Our approach first carefully preprocessed the note into a structured format before applying classifiers. The questions were normalized into a standard template. The answers were handled in a manner based on the property of the questions. The formatted question-answer pairs were then directly used as feature-value pairs and processed by the final classifier. We specifically applied gradient tree boosting as our classifier because of its success in many recent data mining competitions [5]. The model output are decision trees which are much easier for physicians to interpret and validate. Our method also produced the formatted question-answer pairs as a side product, which can be stored into a structured database and could facilitate future investigation of the evaluation records.

Another distinction of our approach is to use bootstrapping to generate resampling of notes to accommodate the unbalanced and small size of training data. Annotated medical documents are found in much less quantity than other sources of annotated documents. The distribution of labels in the annotated corpus is often biased. Resampling the annotated documents with replacement is a simple yet effective method that can be applied in all related clinical natural language processing tasks.

The rest of the paper is organized as follows. Section 2 introduces related work in previous studies. Section 3 describes details of each step of our system. Section 4 presents the evaluation results and error analysis. Section 5 proposes several potential directions to improve our system. Section 6 concludes the paper.

2. Related work

Many previous works have applied natural language processing techniques to electronic health data to determine symptom severity of psychiatric diseases. Perils et al. [6] trained a logistic regression model to predict the probability of a patient being clinically depressed or not by analyzing words and phrases extracted from medical notes of patients with major depressive disorder. Howes et al. [7] applied topic modeling and sentiment analysis to texts of online therapy for depression. They found that using general features such as the discussion topic and sentiment can predict symptom severity with comparable accuracy to face-to-face data. Gorrell et al. [8] built a system to automatically extract the negative symptoms of schizophrenia from patient medical records. They applied the Support Vector Machine with unigrams and part of speech features and manually engineered rules to classify sentences of medical notes for each of the eleven negative symptoms of schizophrenia. Other than psychiatric diseases, Xia et al. [9] extracted narrative variables on symptoms, signs, and medications from notes using the clinical Text Analytics and Knowledge Extraction System (cTAKES) and mapped the concepts into either SNOMED-CT or RxNorm. They trained a logistic regression model with these variables to identify a cohort of multiple sclerosis (MS) patients.

3. Methods

In this section, we present our approach to symptom severity classification. We will first introduce the CEGS N-GRID 2016 dataset and the official evaluation metric, followed by three major steps of our classification system: (1) pre-processing, (2) feature extraction, and (3) classification using gradient tree boosting with resampling.

3.1. Corpus and evaluation metric

The corpus for this year's competition contains 1000 de-identified initial psychiatric evaluation records provided by Partners Healthcare and the Neuropsychiatric Genome-Scale and RDoC Individualized Domains (N-GRID) project of Harvard Medical School. Each note describes one patient. Some of these notes have been rated on an ordinal scale of 0–3 (absent to severe) with respect to the patient's symptom severity in the positive valence RDoC domain by expert clinicians. The distribution of the annotations in the training and test set is described in Table 1.

We used 325 notes with gold labels in the training set to train our classifiers. Each of these notes was annotated by two expert clinicians. A third expert clinician intervened in case of disagreements and acted as a tie-breaker. The 108 notes in the training set annotated by only one clinician were used as the hold-out data to evaluate our models' performances before the official test set was released. We did not use the notes lacking annotations in our submission.

In this competition, the submitted results are evaluated against the gold standard using the macro-averaged Mean Absolute Error:

$$MAE^M = \frac{1}{|C|}\sum_{j=1}^{|C|}\frac{1}{|D_j|}\sum_{x_i \in D_j}|h(x_i) - y_i|, \tag{1}$$

where C is the set of severity scores (0–3), D_j is the collection of records having severity score $j, h(x_i)$ and y_i are the predicted score and gold standard respectively. Note that this measure gives the same importance to every class, regardless of its relative frequency [2].

3.2. Pre-processing

We first preprocessed the raw notes into semi-structured question–answer pairs. Two steps of pre-processing are described below.

3.2.1. Text normalization

The text normalization step deals with the issue of concatenated words in this corpus (e.g. treatmentNeeds, husbandAxis). We first separated such strings at the position of the capital letter in the middle of the string. However, this approach cannot handle the erroneously concatenated uppercase abbreviations (e.g. "LSDADHD" should be "LSD ADHD"). For such cases, we constructed a dictionary of common psychiatric abbreviations and parsed the uppercase strings by simple dictionary matching.

Table 1
Distribution of annotations in training/test set.

	Training	Test
Total	600	400
Annotated with gold labels	325	216
Annotated by only one annotator	108	NA
Not annotated	167	184

Please cite this article in press as: Y. Liu et al., Symptom severity classification with gradient tree boosting, J Biomed Inform (2017), http://dx.doi.org/10.1016/j.jbi.2017.05.015