CS598 DL4H Project Proposal

March 27, 2022
Proj. Jimeng Sun, University of Illinois Urbana-Champaign
Project Team: Brayden Turner & Joshua Smith

Paper #1: Predicting of anaphylaxis in big data EMR by exploring machine learning approaches

Citation:
https://www.sciencedirect.com/science/article/pii/S1532046418301874?via%3Dihub

General Problem:
This paper is attempting to use a variety of machine learning approaches to predict the occurence of anaphylaxis, based on electronic medical records (EMR). Anaphylaxis is a life-threatening allergic reaction that occurs suddenly after contact with an allergen, but diagnosing it can require a manual review of a huge amount of information, which is a very costly and highly time consuming task.

Specific Approach:
The researchers studying this problem applied different machine learning approaches, using most widely used and efficient classifiers in text classification and comparing different document representations. Identification of anaphylaxis cases in EMR is a class-imbalanced problem, with less than 1% of the data describing a positive case. This is interesting because it requires a unique undersampling approach, which is based on the use of a k-means clustering algorithm. This technique had never been applied before to balance text datasets. The researchers applied classical machine learning algorithms (listed in ablations) as well as convolutional neural networks (CNNs) to classify the dataset.

Hypothesis:
The hypothesis put forward in the paper is that applying machine learning in this case can reduce the needed efforts for performing epidemiological studies about anaphylaxis.

Ablations:
One area we could perform an ablation would be with clustering word embeddings. In the paper, they used a k-means clustering algorithm in their bag-of-centroids approach. We could attempt to use another clustering algorithm such as DBSCAN. This may help in finding density-connected regions which would be highly applicable across this bag-of-words style approach. Another ablation possibility is tuning the hyperparameters for the different classifiers used. In the paper, Multinomial Naive Bayes, SVM, Logistic Regression, k-NN, MLP, and Random

Forest are used, chosen due to their large user base and good performance for text classification. They also compared the performance of these to a convolutional neural network.

Access to Data:
The dataset for this paper comes from computerized records of patients attended at the Emergency Department of the Hospital Universitario Fundación Alcorcón (HUFA), a major general hospital in Madrid. The records are in Spanish, and the paper also mentions that Spain has extremely strict data protection legislation due to EU Data Protection Directive 95/46 EC. In order to perform the work, they had to come to a signed agreement between the researchers and the hospital, and data access was provided directly. Thus, we can assume here that data access is incredibly unlikely.

Computational Feasibility:
Replicating the work in the paper does seem computationally feasible. The model was trained on an Nvidia Titan XP, which has 12 GB of VRAM. Both project partners have access to Nvidia RTX 3080s, which have 10 GB of VRAM. If this is not suitable, we can pay for Google CoLab and run the training on their Nvidia P100 GPUs with 16 GB of VRAM. The training took 60s per epoch on their GPU, compared to 2759s per epoch with a CPU, for a total training time of 3000s.

Existing Codebase:
We were unable to locate the codebase for this research paper.

Paper #2: Fine-grained Concept Linking using Neural Networks in Healthcare

Citation:
https://www.comp.nus.edu.sg/~dbsystem/download/dai-sigmod18-paper.pdf

General Problem:
Today, there is a gap between text snippets from the real world (which are entered by clinicians and contain abbreviations, synonyms, acronyms, and simplifications), and the complex medical concepts' canonical descriptions. An example of a concept might be an ICD-10-CM concept D50.0 with the description "iron deficiency anemia secondary to blood loss." Existing methods, such as simple machine learning models and dictionary-based approaches are unable to cross this gap of overlapping meanings between different snippets.

Specific Approach:
This research paper suggests a Neural Concept Linking (NCL) approach, which uses a novel architecture called COMposite AttentIonal encode-Decode neural network (COM-AID). This new approach uses an encoder-decoder which first encodes the concept (canonical description) into a vector, then decodes the vector into a text snippet similar to those produced by clinicians. By using an attention mechanism, COM-AID can enlarge minor concept differences, differentiating them.

Hypothesis:
The hypothesis is that this novel approach of an NCL can produce accurate linking between text snippets from clinicians and medical concepts' canonical descriptions. With the novel COM-AID architecture, they hypothesize NCL will outperform other techniques.

Ablations:
The paper uses seven methods for text comparison, NO-BLECoder (NC), pkduck, WMD, Doc2Vec, extended logistic regression, sequence-to-sequence, and machine translation. Specifically, in the area of extended logistic regression, the paper extends it to not only consider the text features mentioned in the original paper, but they add structural features devised by them. One interesting ablation would be to run the extended logistic regression without these additional structural features to determine the impact that these are making.

Access to Data:
Data is accessible from 2 sources: hospital-x and MIMIC-III. MIMIC-III is publicly available for use and consists of 58,976 diagnosis descriptions that are treated as queries and the ICD9-CM code is treated as the concept.

https://physionet.org/content/mimiciii-demo/1.4/

Computational Feasibility:
This work did not make use of any GPUs, per the implementation details: the network was implemented on a server with four E7540 CPUs and 503 GB of memory. We believe that we can meet the computational requirements either by using our local hardware or through Google CoLab.

Existing Codebase:
No links to existing codebase

Paper #3: INPREM: An Interpretable and Trustworthy Predictive Model for Healthcare

Citation:
https://www.kdd.org/kdd2020/accepted-papers/view/inprem-an-interpretable-and-trustworthy-predictive-model-for-healthcare

General Problem:
This paper attempts to solve the general problem of applying deep learning for clinical prediction tasks. In general, deep learning models lack interpretability by physicians and trustworthiness in results, so it is difficult to actually apply deep learning for meaningful outcomes. There are two major issues today: the first is that deep learning models are unable to tell physicians which medical events are most relevant to the output, and secondly, the models do not allow physicians to know how confidently the predicted probability can be trusted.

Specific Approach:
The researchers put forth a new model specifically for transparent use in healthcare, which is called an interpretable and trustworthy predictive model, or INPREM. INPREM is a linear model which provides a contribution matrix of the input variables which serve as evidence of results, and can help physicians understand the reasons behind a given prediction. In addition, INPREM contains a random gate over each weight of the model and a branch to estimate data noise. Using these two the model can provide details around the uncertainty of a given prediction. This also helps physicians to understand how trustworthy the model's results are.

Hypothesis:
The hypothesis set forth in the paper is that by using the INPREM model, the ability for physicians to understand and interpret deep learning results will be greatly improved. This in turn will lead to these physicians making more robust and accurate decisions. The researchers also predict that INPREM will outperform existing models due to its fundamentally improved design.

Ablations:
It would be interesting to see the impact of removing specific parts of INPREM that are designed to gate different weights and reduce data noise, to increase confidence in results. Perhaps one of these mechanisms is contributing to the result much more than the other, and using two ablations– one without the gating and one without the noise reduction branch– could help us to understand the impact of each.

Access to Data:
The dataset used for this research consists of medical records of 7499 patients from the Intensive Care Unit (ICU). Of this data, patients with at least two visits were chosen. The data for this research paper is available publicly through MIMIC-III..

https://physionet.org/content/mimiciii-demo/1.4/

Computational Feasibility:
Training for this model was done on PyTorch 1.0 using two Nvidia Titan XP GPUs. Nvidia Titan XP GPUs each contain 12 GB of VRAM for storing the model. Both project partners have access to Nvidia RTX 3080s, which have 10 GB of VRAM. If this is not suitable, we can pay for Google CoLab and run the training on their Nvidia P100 GPUs with 16 GB of VRAM.

Existing Codebase:
No links to existing codebase