

Project Milestone - Healthcare

Zenobia Liendo (Gp 1), Guillaume De Roo (Gp 2), Amit Karmakar (Gp 1)

<https://github.com/letslego/W266-ProjectMIMIC>

Abstract: *In this report, we investigate the ability to automatically label patient diseases based on hospital discharge notes. To date we are able to reproduce past results, and suggest some avenues for improvement.*

Introduction

Electronic Health Records (EHRs) have grown significantly over the years and now include an unprecedented amount and variety of patient information, including demographics, medical history, vital signs, prescriptions issued, procedures performed, etc. They usually contain both structured data (e.g. admission dates or test results) as well as unstructured data (e.g. notes written by doctors and nurses)

Provided it can be processed, the information in these records holds the promise of new medical insights and improved medical care, such as faster detection of epidemics, identification of symptoms, personalized treatment, or a more detailed understanding of treatment outcomes.

One such gains is a more automated and accurate way to report diseases. Since 1967, the World Health Organization (WHO) has developed an International Classification of Diseases (ICD) to “monitor the incidence and prevalence of diseases, observe reimbursements and resource allocation trends, and keep track of safety and quality guidelines”.¹

Currently this ICD labelling is done manually by administrative personnel based on definitions. In this paper we try to build an automatic labelling based on the discharge notes written by doctors.

Background

The problem of assigning ICD codes automatically to discharge summaries has previously been studied, in particular ICD-9 codes which is the ICD version mostly used nowadays. Here we mention four papers that used data from the [MIMIC clinical database](#) [8].

Luke Lefebure[3] investigated a neural network model for multi-label classification for assigning ICD-9 codes to patient discharge summaries. An early procedure for multi-label classification using neural networks was BP-MLL which uses a novel pairwise ranking loss function for training [5], but recent research found that cross entropy based loss produces better results[6].

Priyanka Nigam[2] applies deep learning models to the multi-label classification task of assigning ICD-9 labels to medical notes. They find that a Recurrent Neural Network (RNN) and a RNN with Long Short-term Memory (LSTM) units show an improvement over the Binary Relevance Logistic Regression

¹ <http://www.who.int/classifications/icd/en/>

model. The RNN models not only used the discharge summary note but also all the previous notes related to that hospital admission in chronological order. The LSTM model got a F1 score of 0.4168.

A group of researchers from MIT, Harvard, Tufts and other institutions[1] compared convolutional neural networks (CNNs), n-gram models, and approaches based on cTAKES [9] that extract predefined medical concepts from clinical notes and use them to predict patient phenotypes. They used data from the MIMIC-III database but didn't use the ICD-9 codes, they annotated their own phenotypes labels instead. The average F1-score for the CNN model was 76 and outperformed all other models.

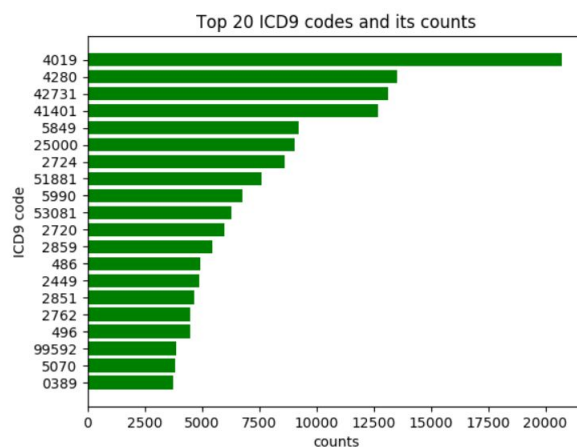
A Study from Columbia University and University of Oxford [7] experimented with two coding approaches: one that treats each ICD-9 code independently of each other (flat classifier), and the other leverages the hierarchical nature of ICD-9 codes into its modeling (hierarchy-based classifier). The hierarchy-based classifier outperforms the flat classifier with F-measures of 39.5% and 27.6%, respectively.

Dataset Exploration

The observations below are based on the MIMIC III database described in Appendix

ICD-9 Codes

ICD-9 is a system of about 15000 numerical codes representing diagnoses and procedures. ICD-9 has been superseded by ICD-10, which has around 70,000 codes, but many medical records, and in particularly those available through public datasets, still use the ICD-9 classification system. The classification is hierarchical, with categories for larger sets of similar health conditions that encompass labels for more specific classifications that take into account causes, specific locations in the body, etc. [Reference \[11\]](#).

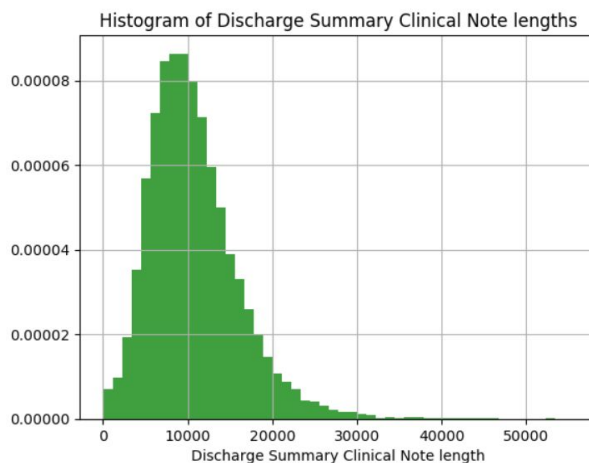


Example of ICD-9 hierarchy and descriptions:

005	Other food poisoning (bacterial)
Excludes: salmonella infections (003.0-003.9)	
toxic effect of:	
food contaminants (989.7)	
noxious foodstuffs (988.0-988.9)	
005.0	Staphylococcal food poisoning
	Staphylococcal toxemia specified as due to food
005.1	Botulism food poisoning
	Botulism NOS
	Food poisoning due to Clostridium botulinum
Excludes: infant botulism (040.41)	
wound botulism (040.42)	
005.2	Food poisoning due to Clostridium perfringens [C. welchii]
	Enteritis necroticans
005.3	Food poisoning due to other Clostridia
005.4	Food poisoning due to Vibrio parahaemolyticus
005.8	Other bacterial food poisoning
Excludes: salmonella food poisoning (003.0-003.9)	
005.81	Food poisoning due to Vibrio vulnificus
005.89	Other bacterial food poisoning
	Food poisoning due to Bacillus cereus
005.9	Food poisoning, unspecified

Discharge Summary

The database presents multiple clinical notes categories including things like “Radiology”, “Nutrition”, “Pharmacy”, or “Social Work”. Here, we focus on “Discharge Summaries” and the description of the ICD-9 codes. Discharge summaries normally already provide a synthesis of main aspects. In the baseline, we consider the original discharge summary “Reports” but will later add possible “Addendums” as well. Notes can contain dates, patient or doctor identifiers, together with text describing diagnosis and treatment, and can vary in size (see below). Unfortunately, the wording is not standardized (e.g. we can cite at least 13 ways to write hypercholesterolemia).



hypercholesteremia
hypercholesterinemia
hypercholestermia
hypercholesterioemia
hypercholesterolaemia
hypercholesterolinemia
hypercholestolemia
hypercholestremia
hypercholestreolemia
hypercholestrolemia
hypercholeterolemia
hypercholesterolemia
hypercholsterolemia

Methods

In our project, we use the discharge summary text to try and predict the associated ICD-9 codes. We can break the prediction pipeline into several steps on which we plan to do variations: preprocessing, error metric, cost minimization algorithm. For the latter, we list 4 models developed as “baseline”.

Input Preprocessing

Cleaning: The original text has some special encodings that we need to clean (e.g. remove “/n”, or “[**]**” for date formats and doctor identifiers).

Vectorization: Word / group of words need to be converted into a vector for further algorithms

- Traditionally, we can remove punctuations and casing depending on other steps. Use stemming to limit the vocabulary and then tokenize the text.
- We can also use Named-Entity Recognition (NER) to extract concept unique identifiers (CUIs) in UMLS (Unified Medical Language System). Since some NER softwares can be sensitive to casing or punctuations (e.g. “a” and “A” have different probabilities to be blood types), we need to feed them with the original text. Options include Bag-of-words, OboAnnotator, MetaMap, C-TAKES, NCBO, HiTEX, CRF. [4]
- Resulting sentence can be encoded, either as binary vectors, using the TF-IDF, or in some cases the NER using a “score” provided by the NER software.

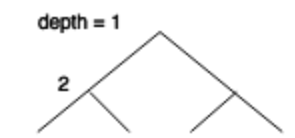
Vocabulary pruning: Even with NER the resulting vocabulary is too large to feed further steps without proceeding to a pruning of the vocabulary. We can use the most frequent words, for example.

Training Metric

Here we are performing a multi-label prediction: in the original set, the number of ICD codes associated to a given hospital admission can vary from 1 to 40.

Since there is a limited/known number of codes, we could theoretically predict a binary vector of length equal to the number of codes. However model complexity and runtime would increase, while rare codes would prove hard to predict. Hence, we will sometimes restrict ourselves to the top 20 codes.

There are multiple ways to measure the error of a given output vector to train the algorithm (cross-entropy, average difference, etc.). Since the ICD codes are related through a tree, we can also try to define a metric which penalizes more heavily if errors are “further” away in the tree hierarchy. For example, we can imagine a metric corresponding to the depth at which vectors are similar (shown in the rightmost column)



	ICD1	ICD2	ICD3	ICD4	mean $ x-y $	$1/\max(\text{depth } x=y)$
Real	0	0	0	1		
Prediction 1	0	1	0	0	1/2	1
Prediction 2	0	0	1	0	1/2	1/2

Algorithms

We have worked on the following 4 baseline models.

- Basic Baseline: Predicting top 4 ICD-9 codes for every discharge summary
- Neural Network Baseline: one hidden layer with relu activation, sigmoid activation on the output layer and cross entropy as the function loss
- Flat SVM Baseline
- Hierarchical SVM Baseline

In the future,

- We can also use Logistic Regression or Random Forest modeling.
- More advanced algorithms include a convolutional neural network (CNN) [1], or a RNN/LSTM with a single layer [2]
- Last, there is also a textual definition associated to each ICD code. Hence we can measure similarity between the discharge summary and the description, using techniques like Jaccard similarity or tf-idf to match notes with definitions and then only ICD codes.

Results and Discussion

Below we explain the algorithms developed to date and summarize their results in a side by side table to facilitate comparison

Preprocessing

In most cases, we simply tokenized the original text, and in the case of the basic baseline and the neural network, we used TF-IDF to proceed to a vectorization with pruning.

Efforts on NER were focused on using MetaMap. While first results and integration with Python were promising, processing the first records took between 30s and 1min. With 50,000 records, this is not scalable and hence was not used to date.

Basic Baseline

This Baseline is focusing on the top-20 ICD-9 codes, based on number of patients with that label. The resulting dataset has 45,293 records, each representing a discharge summary with its list of ICD-9 codes assigned (152,299 total). The true and predicted labels are converted to vectors to apply loss metrics for multilabel classification. In this case each vector has a length of 20 since there are only 20 possible ICD-9 codes in this baseline.

For the basic baseline, we make a fixed prediction corresponding to the top 4 ICD-9 codes for all records, and which can be seen below

```
Label with 4 ICD-9 codes: 4019 4280 42731 41401
Vector : [0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0]
```

Neural Network Baseline

This baseline is based on the Neural Network model implemented by Luke Lefebure [3]. It uses the same input dataset than the one used by the Basic Baseline and takes only 10,000 records to be able to compare results with the ones reported in [3]. It also converts the multi-labels to vectors.

The Clinical Notes are represented by bag-of-words vectors, including only 10,000 tokens with the largest tf and taking out words with high document frequencies.

```
TfidfVectorizer(max_features=10000, stop_words='english', max_df=0.9)
```

The Neural Network is implemented using Tensorflow, it has one hidden layer with relu activation, sigmoid activation on the output layer and cross entropy as the function loss [3].

The following parameters were used: 100 nodes in the hidden layer, 0.01 for the learning rate and batches of 50 records.

SVM Baselines

For both of the svm based models, there are two key elements here:

- Modelling ICD-9 Coding: We are assigning codes to discharge summaries without any constraint on the ICD-9 corpus. This becomes a multi-label classification task with over 7000 ICD-9 codes.
- Evaluation metrics specific to ICD-9 coding or in other words we compare the distances between the actual ICD-9 code (in diagnosis) to the predicted ICD-9 code in the ICD-9 hierarchical tree.

	Basic Baseline	SVM Flat	SVM Hierarchical	NN Baseline	NN from Stanford [3]
Discharge summaries	All 50,000 reports	22,815 documents	22,815 documents	Only report 10,000 notes	10,000 notes
Preprocessing		TF-IDF, remove stopwords	TF-IDF, remove stopwords	TF-IDF Pruning most frequent words	
ICD-9 domain	Top 20	7000	7000	Top 20	Top 20
Ranking loss metric	Train: 0.651 Dev: 0.648			Train: 0.3225 Dev: 0.3126	Dev: 0.317 Test: 0.391
F1 score	Train: 35.60% Dev: 35.94%	Dev: 14.676 %	Dev: 39.539 %	Train: 35.66% Dev: 35.96%	
Recall, precision, specificity*		Recall/Sensitivity 23.34 % Precision: 10.70 %	Recall/Sensitivity 30.075% Precision 57.695 %		

Next steps

For preprocessing, we plan to move to other NER softwares, such as MetaMap Lite and CTakes to increase performance and scale the processing to the full database.

For algorithms, we plan to investigate CNN models looks promising for classifying into ICD-9 codes.

- In the study done by MIT, Harvard, etc [1], had an average F1-score of 0.76 for classifying MIMIC discharge notes into a list of phenotypes, this is a flat list. The CNN model they used outperformed up to 37 points other alternative approaches like n-grams and NER.
- In comparison, Priyanka Nigam[2] applied a LSTM model to classify notes into ICD-9 codes with only a F1-score of 0.42

We also plan to explore “Hierarchical” algorithms

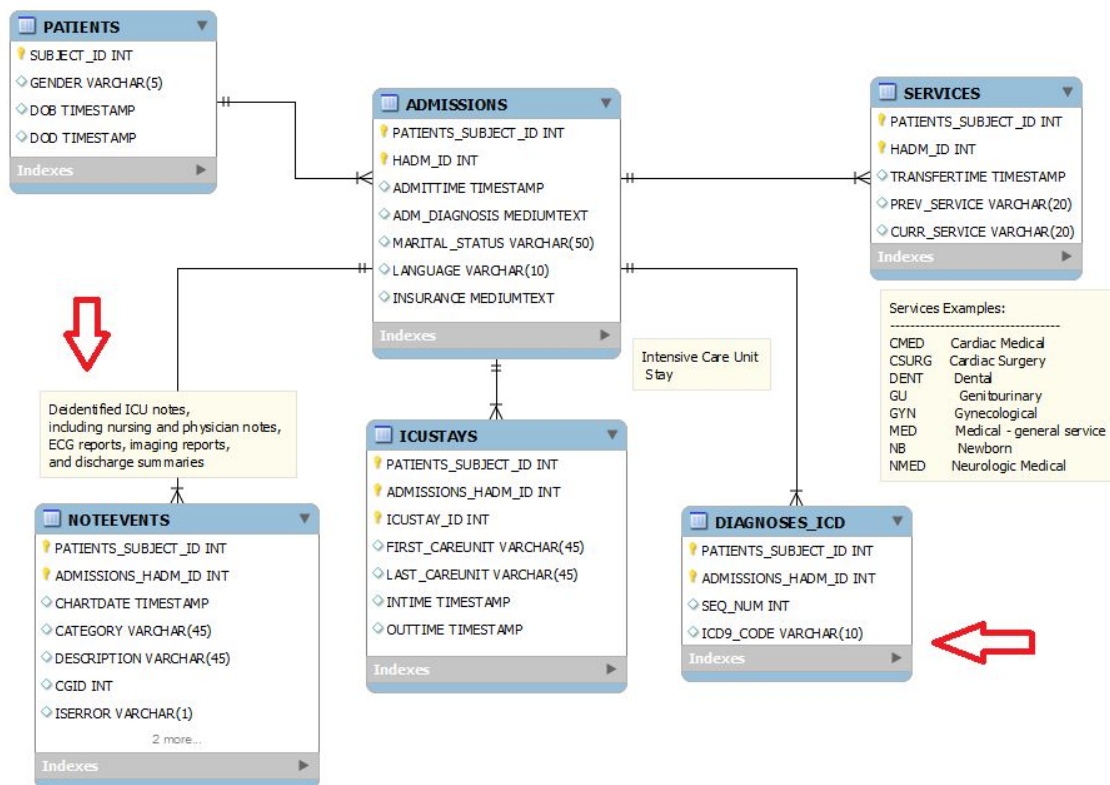
- An SVM hierarchical model got better F1-scores than a SVM flat with F-measures of 39.5% and 27.6%, respectively [7].
- It is possible that a “Hierarchical” CNN model will yield better results
- Hierarchical CNN models had been applied to image classification [10] lowering the top-1 error of the standard CNNs by 2.65%, 3.1% and 1.1%, respectively.

Appendix

Database

[MIMIC III](#) [8] is a dataset developed by the MIT Lab for Computational Physiology, it contains de-identified health records from about 53,000 patients, who stayed in critical care units between 2001 and 2012. MIMIC-III includes several types of clinical notes, including discharge summaries (n = 52,746) labelled by ICD-9 codes (International Classification of Diseases).

Following is an ER diagram of a few relevant tables to our project. The database has about 30 tables, but we will be mostly using two of them, the one with the clinical notes (NOTEEVENTS), and the one with the ICD-9 codes assigned to the discharge summaries (DIAGNOSES_ICD).



The basic and NN baseline focuses on the top-20 ICD-9 codes, based on number of patients with that label (by counting records in the **DIAGNOSES_ICD** table), then selecting only the diagnose records that contained at least one of those top 20 ICD-9 codes and removing the ICD-9 codes that are not in the top 20. We then identified the discharge notes related to these diagnose reports (via a JOIN with the **NOTEEVENTS** table) and exported results to a csv file.

References

- [1] Comparing Rule-Based and Deep Learning Models for Patient Phenotyping
(MIT, Harvard, Tufts, Washington University School of Medicine, etc)
Patrick D Tyler, Leo Anthony Celi. March 25, 2017
- [2] Applying Deep Learning to ICD-9 Multi-label Classification from Medical Records
Priyanka Nigam, Department of Computer Science, Stanford University, June 2016
- [3] ICD-9 Coding of Discharge Summaries
Luke Lefebure, Department of Statistics, Stanford University, June 2016
- [4] Clinical Entity Recognition for ICD-9 Code Prediction in Clinical Discharge Summaries
Jonathan Brauer. Institut fur Informatik. January 25, 2017
- [5] Multi-label neural networks with applications to functional genomics and text categorization. IEEE Transactions on Knowledge and Data Engineering, M.L. Zhang and Z.H. Zhou, 2006
- [6] Large-scale Multi-label Text Classification - Revisiting Neural Networks. ArXiv e-prints, J. Nam, J. Kim, E. Loza Menc'ia, I. Gurevych, and J. Furnkranz., December 2013.
- [7] Diagnosis code assignment: models and evaluation metrics
A. Perotte,R. Pivovarov,K.Natarajan,N. Weiskopf,F. Wood, N. Elhadad, 2014
- [8] MIMIC-III, a freely accessible critical care database.
Johnson AEW, Pollard TJ, Shen L, et al. ,*Sci data*. 2016;3:160035. doi:10.1038/sdata.2016.35
- [9] Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES)
G. Savova, J. Masanz, P.Ogren, J. Zheng, S. Sohn, K.Kipper-Schuler, C. Chute, 2010
- [10] HD-CNN: Hierarchical Deep Convolutional Neural Network for Large Scale Visual Recognition
Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, 2014
- [11] National Center for Health Statistics, International Classification of Diseases, ninth revision, Clinical Modification (ICD-9-CM) <https://www.cdc.gov/nchs/icd/icd9cm.htm>