Diagnosis Code Prediction from Electronic Health Records as Multilabel Text Classification: A Survey

Jeong Min Lee
Department of Computer Science
University of Pittsburgh
jlee@cs.pitt.edu

Aldrian Obaja Muis
Language Technology Institute
Carnegie Mellon University
amuis@cs.cmu.edu

Abstract

This article presents a survey on diagnosis code prediction from various information in Electronic Health Records (EHR): both unstructured free text and structured data. Particularly, our interests are in casting the problem as text classification with multiple sources and using neural network based models. We will first present previous work in this area and describe some simple baseline models for the relevant tasks.

1 Introduction

Electronic Health Records (EHR), the comprehensive collection of patient care details such as order of medications, procedures, lab tests, and diagnosis, serve as an important source of information for healthcare technology, which can be used for training intelligent patient monitors, simulating basic research, providing well-documented case studies of clinically significant pathologies (Moody and Mark, 1996), and providing training data for a personalized healthcare system (Ma et al., 2017).

Being an important aspect of the healthcare sector, the patient records in these databases are usually recorded with lots of details, including both numerical, structured, and textual data. The numerical data comes from measurements such as heart rate, blood pressure, and clinical test results. The structured data come in the form of medical codes associated with each patient record meticulously assigned manually by hospitals originally for billing and administration purposes, while the unstructured data come from the clinical notes, which contain detailed natural language description of the healthcare provided to the patients during the admission.

The large amount of data poses a problem for manual analysis due to information overload. However, the large amount of data, both structured and unstructured, also provide a fertile ground to develop intelligent systems for automatically analyzing these records. For example, picking the appropriate diagnosis codes from more than ten thousands of possible codes is prone to error, not to mention that this requires expert knowledge in medicine (Farkas and Szarvas, 2008).

In this survey we will focus more on the textual content of the records, and as such we will describe existing works on utilizing those texts to develop an automated system that helps analysis of the health records. At the end of this survey, we will also describe one task and the state-of-the-art systems on that task, which we plan to investigate deeper for further analysis.

2 Electronic Health Records Databases

The work on collecting EHR databases is not new, dating back to at least as early as 1996 with the first version of Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) database (Moody and Mark, 1996), which was then developed further to MIMIC-2 (Saeed et al., 2011; Lee et al., 2011) and MIMIC-3 (Johnson et al., 2016). The latter two versions of MIMIC database also include texts in various categories.

Other than the popular and open access MIMIC databases, there are also other health records databases which are used in some previous works, such as Sutter (used by Choi et al. (2015, 2016a,b) and Zhang et al. (2017)), Medicaid (used by Ma et al. (2017)), and 2007 Computation Medical Challenge (introduced by Pestian et al. (2007) and used by Zhang (2008)).

There are also some extensions to the existing datasets, motivated by the desire to have more structured information attached to the clinical texts to alleviate the noise inherent in the text so that downstream tasks might benefit from it. In 2013, Suominen et al. (2013) conducted a shared task focusing on detecting disease and disorder names mentioned in the texts (they considered not just discharge summaries, but also radiology, EEG, and ECG reports).

3 Tasks and Related Work

Due to the many types of information available in EHR, a number of research studies have been conducted to utilize the diverse aspects of the data. In this survey, as the aim of our project is to predict diagnosis codes of patients, we will focus on various approaches to predict diagnosis codes across different types of features (from very unstructured free-text to the more structured data like medical and diagnosis codes) and models (including both neural and non-neural models).

3.1 Diagnosis Prediction Task

On patient management and care in hospital, status of patient is observed and diagnosis is identified by physicians. To standardize the coding scheme, the International Classification of Diseases (ICD) is used in most of hospitals around the world. The revision of the ICD is conducted periodically and current version is 10 (ICD-10. But the previous version ICD-9-CM(CDC, 2011) had been used from 1970s till recent and most of publicly available EHRs such as MIMIC-2 (Saeed et al., 2011; Lee et al., 2011) and MIMIC-3 (Johnson et al., 2016) are using ICD-9-CM. From general disease category to particular organ and site specific subtypes, the ICD-9-CM has tree-like hierarchical structure in which nodes closer to the root represent more abstract categories. In the tree, leaf nodes and intermediate nodes has its own ICD-9-CM code with string label. In the view of machine learning and data mining, the problem of diagnosis code prediction can be seen as a multi-class multi-label classification problem.[1] The large label space (14,025 unique diagnosis codes in total) in this task is one of the great challenges of this task that makes this task interesting.The following sections will describe the various approaches that people have used to do this task.

[1]There are also others treating this as an information retrieval task, such as Rizzo et al. (2015), using the concept of "soft-classification", but in this survey we focus on works that treat this as a classification task.

3.2 Diagnosis Prediction with Non-textual Data

EHR contains various structured data such as past histories of medication, procedure, lab test, and demographic information of patient. Several studies have been conducted using those data. Parthiban and Srivatsa (2012) used SVM and Naive Bayes to predict heart-related diseases with patient demography, cholesterol level, and, blood pressure as features. Choi et al. (2016a) used bidirectional RNN with attention mechanism to predict heart failure diagnosis codes. They used medication code and procedure code as features.

On the neural models, Lipton et al. (2016) used LSTM to predict the 128 most common diagnosis codes. One particular thing to note is that they modeled sequential data in multivariate time series from physiological variables. With time series of 13 variables such as diastolic and systolic blood pressure, CO2, heart rate, and body temperature, it shows moderate performance (30.35% micro-average F1 score) on 128 diagnosis codes.

3.3 Diagnosis Prediction with Unstructured Text Data

Along with numerical variables, unstructured free text also contains valuable information for diagnosis code perdiction. Especially, as it is written by medical experts, it summaries complex physiological states of patient and details of care management and treatment. However, as its unstructuredness nature, particular choice of processing and modeling is needed.

Farkas and Szarvas (2008), Goldstein et al. (2007) and Crammer et al. (2007) use rule-based approaches. Farkas and Szarvas (2008) uses hybrid of C4.5 decision tree and Maximum entropy classifier and Crammer et al. (2007) uses rule-based system that matches the input text with the medical code description, in addition to a keyword-based system. Goldstein et al. (2007) compares rule based model with N-grams and TF-IDF based models. Perotte et al. (2014) leverages hierarchical structure of the ICD-9-CM codes to utilize hierarchical SVM classifier. In addition to leaf node of ICD-9-CM as base labels, they create augmented label sets for each intermediate nodes of ICD-9-CM and trained individual SVM for all label sets. It uses MIMIC-2 database and with 5,030 ICD-9-CM codes it achieves micro F1-score of 39.5%. Saria et al. (2010) incorporated unstructured free notes with structured variables such as clinical events of medication order, ventilator settings and tube placements. Most of these approaches process free text into bag of word or N-gram features and there is inevitable drawback that losing information of order of words and sentences.

More recently, Vani et al. (2017) proposes to utilize a variant of RNN which they call **Grounded RNN (GRNN)** to learn word-level representation and optimize modeling diagnosis codes together. In order to improve interpretability of the RNN model, it ties each dimension of hidden states to the label that is to be predicted. It recorded a micro F1-score of 58.0% in MIMIC-2 (7,042 labels) and 46.4% in MIMIC-3 (5,000 labels[2]). They proposed to use *grounded dimensions* in the RNN hidden states as a mean to give more interpretability to their models, which is an important aspect in this domain, on top of the prediction accuracy. The idea is to force-associate some dimensions in the RNN hidden states to the label space, effectively modeling the model's belief of each label at each word in the document. To prevent the associated dimensions from being used to store non-label-specific information, they did a *semi-diagonal update* (see Fig 1) to help the model to store label-specific information in the associated dimensions $\mathbf{g}$ and other information in the normal hidden states $\mathbf{c}$.

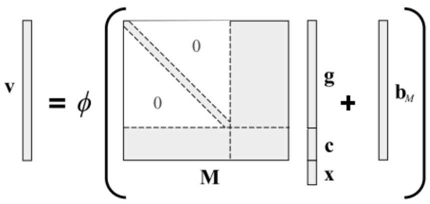

Figure 1: Semi-diagonal update in GRNN (Vani et al., 2017).

3.4 Other Prediction Tasks using EHR

In addition to the task of predicting diagnosis codes, there are many other interesting works utilizing EHR data. Zhang et al. (2017) uses GRU with content-based attention to predict medication prescription given the disease codes. Suresh et al. (2017) compares LSTM and CNN models on predicting clinical interventions from both numerical and textual data.

With regard to usage of clinical notes, Ghassemi et al. (2014) tackles in-hospital mortality prediction problem by generating latent topics using LDA (Blei et al., 2003) and building an SVM classifier. Caballero and Akella (2015) predicts patient readmission to ICU using clinical notes. They created features from unstructured free text using name entity recognition frameworks. Jo et al. (2015) models patient mortality by modeling state transition of latent topics of notes and training SVM classifier.

4 Experiment

We conducted a few experiments with simple models to estimate the difficulty of the diagnosis code prediction task.

4.1 Dataset

For the experiments we use MIMIC-3 Database, a publicly available, multi-granular, deidentified EHR contains more than 60,000 hospital admissions of critical care patients. From the dataset we extracted medication orders, lab test orders, and procedures that occurred more than 20 different admissions. As a result, we have 1,140 medications, 518 lab events, and 423 procedures. Although we do not make use of these for this first checkpoint, they might potentially be used in the final project. Similarly, in total there are 6,914 ICD-9-CM labels in the full dataset. Unlike the setup of Vani et al. (2017), we use this full set of labels for this preliminary experiments.

The main data that we use for this project would be the discharge summaries, which contain textual description of each patient's admission. We follow the text preprocessing procedure of Muis and Lu (2016) in using Stanford CoreNLP sentence splitter with additional heuristics to handle some semi-structured text in the dataset, including bullet lists and section names. The sentences are then tokenized using regex-based tokenizer and anonymization tokens (e.g., "[**doctor first name 77**]") are normalized (e.g., "DOCTOR_NAME"). We prefer oversplitting (most punctuations are considered separate tokens) during tokenization, which has the advantage of smaller vocabulary size and less sparsity, since there are lots of chemical and medicine names in the dataset.

For the purpose of this assignment, we also took about 1/10th of the full dataset to be used as a development set.

The dataset statistics can be seen in Table 1.

[2]They use the most common 4,000 diagnosis codes plus 1,000 procedure codes as the label space.

	Full dataset				Development			
	Train	Valid	Test	Total	Train	Valid	Test	Total
#Docs	31,676	10,442	10,556	52,674	3,151	1,075	1,045	5,271
#Sentences	4,951,612	1,636,034	1,650,258	8,237,904	496,424	168,517	162,069	827,010
Avg. #Sents/doc	156.32	156.68	156.33	156.39	157.54	156.76	155.09	156.90
Avg. #Tok/doc	1990.57	1999.21	1992.24	1992.62	2006.18	2012.10	1960.45	1998.32
Avg. #Labels	11.76	11.79	11.66	11.74	11.77	11.71	11.46	11.70

Table 1: Full MIMIC-3 dataset statistics.

Model	Full dataset						Development					
	Valid			Test			Valid			Test		
	P	R	F	P	R	F	P	R	F	P	R	F
SVM (C=32)	59.11	25.99	36.11	59.20	26.20	36.32	68.11	13.15	22.05	67.83	13.50	22.52
SVM (C=128)	58.18	26.16	36.09	58.21	26.36	36.28	67.74	13.22	22.12	67.49	13.68	22.75
MLP (h=1024)	64.16	21.68	32.41	64.03	21.58	32.28	57.04	13.84	22.27	57.97	14.30	22.95
GRNN-best	-	-	-	-	-	46.40	-	-	-	-	-	-

Table 2: Result (micro-average F1 scores) of some basic models on the task without label filtering.

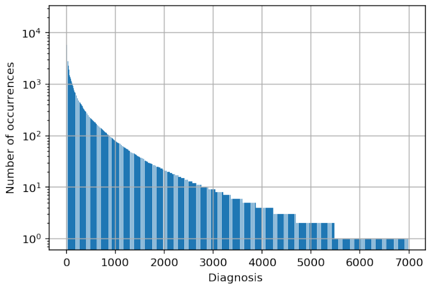

Figure 2: Occurrences of diagnosis codes in full dataset (note the log scale in y-axis).

4.2 Models

In general, we consider the task to be a multi-class multi-label classification problem, where the input is a document (a sequence of words or can also be considered a sequence of sentences) and the output is all the ICD-9-CM labels associated with the document.

We implemented two models following a simple bag-of-words approach to do the prediction for our experiments. We implemented a multi-class multi-label SVM with linear kernel using one-vs-rest strategy using `sklearn` package in Python. We did some tuning on the C parameter on development set. We also implemented multilayer perceptron (MLP) with one hidden layer of size 1024 (with Tanh) in PyTorch with MSE loss function (with Sigmoid activation at the end, since we are doing multi-label classification). We use batch size 64 and learning rate of 0.002 and trained the network for 3700 and 300 epochs in development and full dataset, respectively, using AdaGrad optimizer.

4.3 Results and Discussion

We report micro-average F1 scores in Table 2.

Looking at the bag-of-words models, we can see that it got a descent performance compared to state-of-the-art score (46.4%, although this is not directly comparable due to our larger label space). Although these models do not consider word orders, it can capture some prominent keywords in the text which are predictive of the labels. We speculate the higher result of SVM compared to MLP in the full dataset due to the better TF-IDF values obtained from larger dataset.

5 Conclusion and Future Work

The task of diagnosis code prediction from discharge summary is a challenging task, due to the large number of labels and the long textual description as input, yet interesting, due to the hierarchy in the label space, and suitable for neural models, due to the ample amount of training data.

We would like to investigate more how GRNN performs and how we can utilize other information on top of it, such as the label hierarchy and other non-textual data including orders of medication, lab test and procedures. We extracted these additional data for the next step.

In addition, one of the characteristics of the dataset is the the high class imbalance (diagnosis codes) as seen in the Figure 2. Treatment to class imbalance problem in CNN has been studied in Buda et al. (2017). It uses oversampling, undersampling, and thresholding that compensates for prior class probabilities. It would be interesting to explore the idea to our problem.

Another interesting path toward exploiting the EHR data is to introduce time dimensionality into the diagnosis prediction task. That is, modeling previous admission's notes and diagnosis codes into predicting next diagnosis codes. For that path, it would be feasible make hierarchical model that uses MLP or CNN at the lower level and uses recurrent neural network models at top level.

Work Distribution

Jeongmin I worked on the implementation of LSTM (although it was not included to the report due to some implementational bugs, the code is in the repository) and data processing on admissions and non-textual features such as medications, lab test orders and procedures which will be used for next task. For report, I contributed to the section of tasks and related work, and get the statistics of the dataset and for the section of experiment, I contributed to writing of the parts that I worked on implementation side.

Aldrian I worked on the implementation of MLP and SVM and also the text preprocessing (sentence splitting, tokenization), including the corresponding sections in the report, such as the result table and the data statistics. In addition to that I also contributed in the report through the introduction and the initial overall draft, and final touch up of the report.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

M. Buda, A. Maki, and M. A. Mazurowski. 2017. A systematic study of the class imbalance problem in convolutional neural networks. *ArXiv e-prints* .

Karla Caballero and Ram Akella. 2015. Dynamic Estimation of the Probability of Patient Readmission to the ICU using Electronic Medical Records. *AMIA Annual Symposium* 2015:1831–40. http://ncbi.nlm.nih.gov/pmc/articles/PMC4765609.

CDC. 2011. ICD-9-CM Official Guidelines for Coding and Reporting. http://cdc.gov/nchs/data/icd/icd9cm_guidelines_2011.pdf.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016a. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Neural Information Processing Systems (NIPS)* (Nips). http://arxiv.org/abs/1608.05745.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2015. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *Proceedings of Machine Learning Research (PMLR)* 56. http://arxiv.org/abs/1511.05942.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2016b. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *Knowledge Discovery and Data Mining (SIGKDD)* pages 787–795. https://doi.org/10.1145/3097983.3098126.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. 2007. Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007 Biological, Translational, and Clinical Language Processing - BioNLP '07* (June):129. https://doi.org/10.3115/1572392.1572416.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC bioinformatics* 9 Suppl 3(Suppl 3):S10. https://doi.org/10.1186/1471-2105-9-S3-S10.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 75–84.

Ira Goldstein, Anna Arzrumtsyan, and Ozlem Uzuner. 2007. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2007:279–83. http://www.ncbi.nlm.nih.gov/pubmed/18693842.

Yohan Jo, Natasha Loghmanpour, and Carolyn Penstein Rosé. 2015. Time series analysis of nursing notes for mortality prediction via a state transition topic model. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM, pages 1171–1180.

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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3:160035. https://doi.org/10.1038/sdata.2016.35.

Joon Lee, Daniel J. Scott, Mauricio Villarroel, Gari D. Clifford, Mohammed Saeed, and Roger G. Mark. 2011. Open-access MIMIC-II database for intensive care research. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* pages 8315–8318. https://doi.org/10.1109/IEMBS.2011.6092050.

Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzell. 2016. Learning to Diagnose with LSTM Recurrent Neural Networks. In *International Conference on Learning Representations (ICLR)*. http://arxiv.org/abs/1312.6229.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. *Knowledge Discovery and Data Mining (SIGKDD)* pages 1903–1911. https://doi.org/10.1145/3097983.3098088.

G.B. Moody and R.G. Mark. 1996. A database to support development and evaluation of intelligent intensive care monitoring. *Computers in Cardiology 1996* pages 657–660. https://doi.org/10.1109/CIC.1996.542622.

Aldrian Obaja Muis and Wei Lu. 2016. Learning to Recognize Discontiguous Entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 75–84. https://doi.org/10.18653/v1/D16-1008.

G Parthiban and SK Srivatsa. 2012. Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)* 3:2249–0868.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment : models and evaluation metrics. *JAMIA* 21(2):231–237. https://doi.org/10.1136/amiajnl-2013-002159.

John P Pestian, Christopher Brew, Paweł Matykiewicz, D J Hovermale, Neil Johnson, K Bretonnel Cohen, and Włodzisław Duch. 2007. A shared task involving multi-label classification of clinical free text. *Proceedings of the Workshop on BioNLP 2007 Biological Translational and Clinical Language Processing BioNLP 07* 1(June):97–104. https://doi.org/10.3115/1572392.1572411.

Stefano Giovanni Rizzo, Danilo Montesi, Andrea Fabbri, and Giulio Marchesini. 2015. ICD Code Retrieval: Novel Approach for Assisted Disease Classification. In Naveen Ashish and Jose-Luis Ambite, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer International Publishing, Cham, volume 9162 of *Lecture Notes in Computer Science*, pages 147–161. https://doi.org/10.1007/978-3-319-21843-4$_1$2.

Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine* 39(5):952–960. https://doi.org/10.1097/CCM.0b013e31820a92c6.

Suchi Saria, Gayle McElvain, Anand K Rajani, Anna A Penn, and Daphne L Koller. 2010. Combining structured and free-text data for automatic coding of patient outcomes. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2010, page 712.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. 2013. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In P. Forner, editor, *Information Access Evaluation: Multilinguality, Multimodality, and Visualization*, Springer-Verlag Berlin Heidelberg, volume 8138, chapter 24, pages 212–231. https://doi.org/10.1007/978-3-642-40802-1$_2$4.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical Intervention Prediction and Understanding using Deep Networks. *Computing Resource Repository (CoRR)* pages 1–16. http://arxiv.org/abs/1705.08498.

Ankit Vani, Yacine Jernite, and David Sontag. 2017. Grounded Recurrent Neural Networks http://arxiv.org/abs/1705.08557.

Yitao Zhang. 2008. A hierarchical approach to encoding medical concepts for clinical notes. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Student Research Workshop - HLT '08* (June):67. https://doi.org/10.3115/1564154.1564168.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart Sutter, Health Stewarwf@sutterhealth Org, and Jimeng Sun. 2017. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity. *Knowledge Discovery and Data Mining (SIGKDD)* pages 1315–1324. https://doi.org/10.1145/3097983.3098109.