Predicting patient's diagnoses and diagnostic categories from clinical-events in EHR data

Seyedsalim Malakouti[1] and Milos Hauskrecht[1]

University of Pittsburgh, Pittsburgh PA 15260

Abstract. In this paper we develop and study machine learning based models based on latent semantic indexing capable of automatically assigning diagnoses and diagnostic categories to patients based on structured clinical data in their Electronic Health record (EHR). These models can be either used for automatic coding of patient's diagnoses from structured EHR data at the time of discharge, or for supporting dynamic diagnosis and summarization of the patient condition. We study the performance of our diagnostic models on MIMIC-III EHR data.

Keywords: Lower Dimensional Representation, Singular Value Decomposition, Electronic Health Records, Machine Learning, ICD-9 Diagnosis

1 Introduction

Healthcare is one of the most promising areas for applications of data mining and machine learning methodologies. Since the adoption of electronic health records (EHRs), there has been an explosion in digital clinical data available for learning and analysis. However, the development of models that are derived from such data and that can solve important clinical problems still lags the advances in data collection. One important that can use such data is the problem of automated assignment of diagnoses to EHRs. Motivation behind solving this problem can be summarized as follows. First, automated diagnostic assignments can be used as a utility that informs clinician about the diagnoses associated with the current patient. Second, it can be used as a patient condition summarization tool to define proper context for analysis of patient management steps or to support improved prediction of future outcomes.

Going from structured EHR data to automated diagnoses is not easy. First, structured EHRs consist of a large number of time series that represent variety of labs, physiological measurements, symptoms, treatments, procedures, etc. Hence it is not easy to automatically associate the signals in these time series with specific diagnoses, especially when the diagnoses are defined by a combination of these signals or the same diagnosis can be confirmed by multiple alternative signals. This problem is even more challenging when data are sparse (data are collected at irregular times) and many time series for the patient cases are unknown or missing. Second, the assignment of diagnoses to patient case is typically done at the time of the discharge, which means it is not only unclear

what the signals related to the specific diagnosis are but also when they occurred in time. Finally, some diagnoses are very rare and even with moderate to large EHR repositories the number of patients suffering from the specific disease is very small, so learning of diagnostic models for such diseases is not feasible.

In this work we study this important problem by investigating methods from text mining, natural language processing (NLP) and information retrieval, but apply them to structured EHR data. Briefly we consider each patient's EHR to be equivalent to a document, and clinical events of different kinds recorded in EHR as words or terms in the document. To represent different events describing the patient case we consider the bag-of-word (BoW) representation that uses individual event counts and transform it using a lower-dimensional projection, based on Latent Semantic Indexing[5] that aims to better reflect semantic relations between events. The advantage of such a representation is that it permits us to consider a large number of events of different types typically found in the EHR data, and is also robust in handling missing and unknown data sources very common in EHRs. Additionally, it helps us to define the meaningful similarity among the patients as well as similarities among the words (clinical events). We use this new patient case representation to build models for individual diagnoses, as well as, diagnostic categories we define with the help of icd-9 hierarchy. Through experiments we demonstrate our new representation is able define accurate diagnostic models at different levels of abstractions.

2 Related Work

Majority of existing work modelling patient diagnostic process fall into one of two categories. The first group tackles prediction of future patient visit diagnosis. Lipton et al. proposed a Recurrent Neural Network (RNN) architecture based on Long Short Term Memory units to predict future patient visit diagnosis from a collection of 13 clinical variables [8]. GRAM [3] is an attention based RNN network that uses a BoW representation of patient's previous diagnosis as their input and take advantage of diagnosis hierarchies to extend low level diagnosis to categories.The second category of existing work studies the problem of automatic diagnosis assignment at the end of hospitalization and it is mainly motivated by improving hospital billing process. Other solutions were also proposed based on Autoencoder and LSTM neural network architectures [10, 12].

The data models and SVD-based lower dimensional projections we propose in our work are typically used for analysis of text data. For example, SVD has been applied in addition to information retrieval and document analysis [2]. In terms of clinical applications, SVD and other lower dimensional representation methods including non-negative matrix factorization have been used on EHR data for missing value imputation [1], future visit diagnosis prediction [9] and medical phenotyping [11]. Despite numerous studies in diagnoses prediction and assignment, the existing work has not attempted to take advantage of the entire span of structured clinical data in EHR nor they have studied the advantage of looking at diagnostic categories as target variables.

3 Methodology

Let V_i denote a patient visit i and let $D = \{V_1, V_2, ..., V_{|D|}\}$ be a set of all patient visits in our data. A visit can be defined as $V_i = \{x_i, y_i\}$ where x_i and y_i are respectively a set of clinical events and diagnoses assigned to the patient during the visit. Clinical events are formed by a discrete representation of clinical information derived from Electronic Health Records (see below for details). Additionally, we adopt a bag-of-word (BoW) representation of a patient's EHR, therefore, $x_i \in \mathbb{N}^E$ reflects the number of occurrences of each clinical event during a patient's stay where E is the total number of event types.

Low dimensional representation of patient's clinical information is a key step in summarizing the information important for learning of diagnostic models. We define a low dimensional embedding as a mapping $E \mapsto \mathbb{R}^k : x_i : u_i$ that maps a patient's visit's data to a new lower dimension dense vector $u_i \in \mathbb{R}^k$ while $k << |E|$. Automatic learning of a low dimensional representation of complex data vectors is one of the most actively studied topics in machine learning research [4]. Our goal in this work is to show that these methods are capable supporting our problem - automatic assignment of diagnoses to patient's clinical data. Briefly, Electronic Health Records contain tens of thousands of different information including medications, procedures and surgeries, lab results, vital signs, pain scores and etc. However, often this data contain missing values. Additionally, much of this information is interrelated, conveying interchangeable or opposite information regarding patient condition. For example, various medications are used to treat blood pressure related conditions including Diuretics, Beta blockers and Alpha-1-Agonist medications. However, the first two are prescribed to patients with high blood pressure and the third group is ordered for patients with low blood pressure. Therefore, with the help of lower dimensional representation methods one can learn compact representations of patient data that a simple bag-of-word model fails to do.

Latent semantic indexing is a statistical method for analyzing the relationship between a set of documents and terms used in information retrieval by finding underlying concepts [5] . This is done by finding a Singular Value Decomposition(SVD) of original term-document matrix A. We consider each patient's EHR to be equivalent to a document, and clinical events of different kinds recorded in EHR as words or terms in the document. The underlying concepts are in fact eigenvectors of symmetric matrix $X^T X$ and are represented in the left singular vector matrix in $A = U\Sigma V^T$. Therefore, rank k Singular value decomposition of patient matrix $X_{|D|,|E|}$ can be obtained as:

$$X_{|D|,|E|} = U_{|D|,k} \Sigma_{kk} {V^T}_{k,|E|} \tag{1}$$

The lower dimensional representation of u_i can be obtained as $u_i = x_i V \Sigma^T$.

Learning diagnostic models includes learning one model per y_i (diagnosis or diagnostic category) using logistic regression with L2 regularization to capture the input-output relations. All models use low dimensional vectors as their inputs. If the lower dimensional representation is successful in capturing all important information about the patient visit in a compact form, we expect it to

be sufficient. We note that this approach is not optimized to capture the relations among different diagnoses and their categories. We leave the study of these models to our future work.

4 Experiments

We experiment with our models on MetaVision part of MIMIC-III [7], an open access EHR dataset obtained over a 12-year time span that covers 22K patient visits or hospitalization to ICU. MIMIC-III encodes patients' diagnoses using standard ICD-9 codes. We enrich the ICD-9 codes with diagnostic categories defined by ICD-9 hierarchy. We limited our experiments to ICD-9 codes with at least 0.02 for prior probability of positive examples chosen to guarantee enough positive examples for learning and cross validation. This results in 421 diagnoses and diagnosis categories. We evaluate the performance of our models on the post-discharge diagnostic assignments expressed in terms of icd-9 diagnoses and their categories using the area under receiver operating characteristics curve (AUROC) and area under precision recall curve (AUPRC). The latter statistics is known to be more appropriate in the presence of imbalanced data [6].

Data processing is needed before creating a bag-of-word representation of patient data. We convert patient information in EHR to a set of meaningful binary events. We used medication and procedure orders by converting them to occurrence indicators. Laboratory results and physiological measurements with numerical values were converted to Abnormal Low, Normal or Abnormal High events based on their standard normal ranges, discrete valued measurements were converted to events matching these values. Finally, pain assessments were converted to special events reflecting the different pain levels. After the conversion our new events data covered 4826 clinical events including 2420 for medication orders, 116 for procedure orders, 2012 for laboratory results and 278 for physiological and pain assessment measurements.

Results in Table 1 show that more accurate models can be learned by using higher level (more general) diagnoses from disease hierarchy by taking advantage of their higher priors. However, it is important to mention that moving up the hierarchy may not always improve the models as generic categories might be harder to learn. An example of this case is the category "Diseases of Genitourinary" in Table 1 that has lower AUPRC and AUROC from its immediate sub-category. Additionally, generic categories may not be as informative.

5 Acknowledgement

This work was supported by NIH grant R01GM088224. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

1. Beaulieu-Jones, B.K., Moore, J.H.: Missing data imputation in the electronic health record using deeply learned autoencoders. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017. pp. 207–218. World Scientific (2017)

Task Name	Prior	AUROC	AUCPRC	Task Name	Prior	AUROC	AUPRC
Root ICD9 codes average	0.434	0.74	0.647	Forms of Heart Failure	0.509	0.822	0.822
All ICD-9 codes average	0.096	0.771	0.262	Heart failure	0.249	0.852	0.681
Diseases of Genitourinary	0.495	0.862	0.869	Systolic heart failure	0.096	0.808	0.324
Nephritis related diseases	0.371	0.931	0.891	Chr systolic hrt failure	0.035	0.732	0.091
Acute renal failure	0.269	0.878	0.718	Diastolic heart failure	0.103	0.81	0.331
Ac kidny fail, tubr necr	0.056	0.897	0.376	Cardiac dysrhythmias	0.352	0.793	0.674
Ac kidney failure NOS	0.213	0.832	0.527	Cardiac arrest	0.026	0.849	0.211
Chronic kidney disease	0.198	0.917	0.742	Atr fibrillation & flutter	0.269	0.831	0.642
Chr kidney dis stage III	0.029	0.863	0.158	Atrial fibrillation	0.261	0.829	0.63
End stage renal disease	0.053	0.971	0.79	Atrial flutter	0.032	0.751	0.095
Chronic kidney dis NOS	0.097	0.838	0.323	liver disease & cirrhosis	0.073	0.876	0.586
Diseases of Urinary Sys.	0.183	0.732	0.351	Alcohol cirrhosis liver	0.028	0.924	0.419
Disease of male Genital..	0.061	0.657	0.102	Cirrhosis of liver NOS	0.03	0.885	0.273

Table 1: Performance of models for diagnoses on different ICD-9 hierarchy levels

2. Berry, M.W., Drmac, Z., Jessup, E.R.: Matrices, vector spaces, and information retrieval. SIAM review **41**(2), 335–362 (1999)
3. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: Gram: graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 787–795. ACM (2017)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(Aug), 2493–2537 (2011)
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
6. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering (9), 1263–1284 (2008)
7. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
8. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.: Learning to diagnose with lstm recurrent neural networks. arXiv preprint arXiv:1511.03677 (2015)
9. Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports **6**, 26094 (2016)
10. Pakhomov, S.V., Buntrock, J.D., Chute, C.G.: Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. Journal of the American Medical Informatics Association **13**(5), 516–525 (2006)
11. Wang, Y., Chen, R., Ghosh, J., Denny, J.C., Kho, A., Chen, Y., Malin, B.A., Sun, J.: Rubik: Knowledge guided tensor factorization and completion for health data analytics. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1265–1274. ACM (2015)
12. Xie, P., Xing, E.: A neural architecture for automated icd coding. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1066–1076 (2018)