Embedding Representations of Electronic Medical Records

Internship at INCLUDE (CHU de Lille)

1. Medical context

The High Council on Public Health estimates that 15 million people in France suffer from one or more chronic diseases. This number has increased substantially in recent years, for two main reasons: population ageing, and improved treatments (meaning that formerly fatal, acute diseases have become chronic conditions). In the move towards personalized medicine, there is an increasing need for tools that can detect chronic diseases, provide prognostic information, and predict the occurrence of complications and exacerbations leading to hospitalization.

Furthermore, medical nosology is becoming increasingly complex. The numbers of disease entities, diagnostic tests, biomarkers and treatment modalities have increased exponentially over recent years. As a result, clinical decision-making has also become more complex, and requires the integration and synthesis of a large amount of clinical information. Based on the patient's complaint and initial signs and symptoms, the physician seeks to rule out a number of potentially serious differential diagnoses. The most useful disease characteristics are identified; when the probability of one of the diagnoses reaches a predetermined level of acceptability, the process is stopped, and the diagnosis is accepted. In some senses, the physician acts as a classifier.

All French hospitals collect medical and administrative data as part of hospital invoicing. Their electronic medical records (EMRs) notably contain original **data on lab results, drug prescriptions, and clinical notes**. Furthermore, information on outpatients and causes of death are provided as part of the SNDS (Système National des Données de Santé). It is now possible to consider linking outpatient and inpatient data, for a full analysis of the care pathways followed by patients with chronic diseases. Most of the today's research projects have mainly used structured SNDS data. **The search for predictive elements in EMRs requires a focus on more complex, unstructured data** such as free text, event sequences, and changes over time in laboratory parameters. Structured information is coded using terminologies with a high number of component items (for example, there are 32,000 codes in the International Classification of Diseases, 10th Edition (ICD-10)); some of these correspond to very similar medical concepts, and could be grouped together and summarized.

2. Methodological context

Unsupervised and weakly-supervised statistical learning methods make it possible to consider building unified representations by synthesizing information from heterogeneous and (in some cases) unstructured data associated with patients. This can be done via "embedding" transformations that provide an alternative representation of the initial objects within a structured mathematical space. The typical purpose of an embedding is to represent initial objects in a small space that preserves or even reveals the structure of the relationships between objects. This is achieved *via* the introduction of a mathematical distance and/or a reduction in dimension with respect to the initial space.

In the case of natural language in general and words in particular, embedding makes it possible to replace words (represented by a simple numerical index within the vocabulary, or by vectors whose dimension V is that of the vocabulary) by vectors in a Euclidean space of dimension m (where m <<< V) whose relative positions reflect linguistic properties (e.g. semantics or syntactics). This embedding can be based exclusively on informational criteria, e.g. using the GloVE algorithm ([1]); this consists in (i) factoring the co-occurrence matrix of words within a given corpus, and (ii) using one of the resulting matrices as a representation of words in a space of a chosen dimension. In other cases (word2vec [2], BERT [3], etc.), we will rely on an artificial neural network trained on one or more linguistic tasks - the prediction of a hidden word from those surrounding it in a given sentence, for example. Embedding will then corresponds to one of the intermediate representations learned by this network at the end of its training.

The methods initially developed to produce "word embedding" in natural language could be transposed to other types of data and objects. In a medical context, it makes sense to produce embeddings of a nomenclature's items, so as to represent mathematically their relationships and notably capture similarities and possible redundancies between items. This can for example be applied (separately) to drugs, symptoms, exams or medical acts. In turn, these representations are of value in developing (for example) decision support and alert systems.

For quantitative measures, such as those in a standardized biological exam, embedding methods can be desirable to construct vectors synthetizing multivariate information, at a given date and/or through time. For these data, an embedding can be constructed using simple methods such as PCA or t-SNE, or more complex models such as auto-encoder artificial neural networks, or even adaptations of some network architectures and tasks initially conceived for word embedding.

Lastly, following recent research on **combining embeddings of different types of related data** (e.g. Mixture-of-Embedding-Experts (MEE) [4], Mixture-of-Experts (MOE) [5], or Multi-channel Variational Auto-encoder [6]), learned representations from heterogeneous medical data could be combined into a "patient embedding" representation. The latter (built from massive data in warehouses) could then be transferred to other contexts, providing greater statistical power and thus reducing the number of participants needed for prospective research.

3. Objectives of the internship

Main objective

The main objective will be the construction of embedding representations for various types of medical data, starting with drugs (based on prescriptions) and biology results (for a set of standard exams). These embeddings will then be **used as inputs for supervised machine learning methods** in order to predict the occurrence of a health outcome, in a classical biomedical research context. Comparisons to baseline models making use of the “raw” data will be conducted to assess the interest in building (and possibly transferring) embeddings to gain predictive power.

This objective will require the intern to reflect (possibly relying on a literature review) on the adequate embedding methods *for each type of data* depending on its specificities, to implement the embedding (using open-source software and/or new implementations), and to train both baseline and embedding-based machine learning predictive models. Research on optimal model classes and/or hyper-parameters may be conducted both at the embedding and classification steps.

Additional objectives

Depending on the time, the results and the personal inclinations of the intern, **a variety of follow-up and complementary objectives may be pursued**:

(a) Embed additional types of data.

Diagnostic and/or medical act codes may be considered for embedding. As for clinical notes (text), existing work conducted in the team may be leveraged, or new solutions may be explored.

(b) Define (and implement) additional tasks to assess the quality of the embeddings.

This open topic, which may involve a literature review, could either leverage available information, or produce specifications for a task involving data annotation to be obtained from clinicians.

(c) Construct a “patient embedding”, aggregating multi-channel information.

This exploratory topic, which should involve a literature review, will aim at producing a “unified” representation of patient data to be leveraged for prediction tasks (and/or additional ones). This global embedding may either be constructed by aggregating data-type-wise embeddings, or by starting back from the raw data.

(d) Reflect on explicability concerns emerging from using embeddings as inputs in a supervised task.

This open topic, which could involve a literature review, may be taken in a variety of directions: searching for embedding and prediction methods that enable deriving clinical knowledge from trained models; defining application cases that do not suffer from the potential lack of explicability arising from using embeddings; etc.

4. INCLUDE and the Lille University Hospital

With a community of 16,000 professionals, the Lille University Hospital is one of the largest campuses of northern Europe dedicated to healthcare, and has become a reference for teaching, innovation and research over the recent years. In 2017, more than 5,000 patients were included in clinical studies, representing a budget of nearly 70 million euros. During the year 2019, INCLUDE, the Integration Centre for the Lille University Hospital for Data Exploration, was authorized by the CNIL to reuse patients’ data (e.g. through routine EMRs) for clinical and methodological research. Data integration is carried out within the data warehouse while the statistical development (in every sense, including machine learning and deep learning) is provided by a team of data scientists with significant computing resources (GPU server). Thanks to the extremely rich scientific environment available on the campus, INCLUDE actively collaborates with various clinical research teams - but also teams from INRIA, INSERM and the University of Lille, to explore the potential of artificial intelligence techniques in healthcare.

Data and practical concerns

The dataset used for this internship will consist in biomedical measurements, drugs prescriptions and clinical notes for a wide group of patients of the Lille University Hospital between 2008 and 2019. This data is currently held in the hospital's data warehouse, administered by INCLUDE, and is therefore ready to be made available securely on a self-hosted computational infrastructure; an authorization from the CNIL allows its lawful statistical exploitation. The data will not comprise information allowing to identify individuals (identity variables will be dropped, an *ad hoc* patient id will be used to align data tables, and text will undergo a de-identification process removing sensitive information).

The internship will take place at INCLUDE, with an articulation between remote and office work depending on the sanitary situation and on national and institutional rules and recommendations. Secure remote access to computational resources holding the data will be provided at any rate.

Supervision and support will be handled by both clinicians and data scientists, respectively providing with expert knowledge of the data and biomedical studies in general, and with scientific and technical knowledge and assistance.

5. Bibliography

[1] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. *GloVe: Global Vectors for Word Representation*.

[2] Tomas Mikolov, Ilya sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Distributed Representations of Words and Phrases and their Compositionality*. arXiv:1310.4546 [cs.CL]

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs.CL]

[4] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. *Learning a Text-Video Embedding from Incomplete and Heterogeneous Data*. arXiv:1804.02516 [cs.CV]

[5] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E. Gonzalez. 2018. *Deep Mixture of Experts via Shallow Embedding*. arXiv:1806.01531 [cs.CV]

[6] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. 2019. *Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data*. ICML 2019. hal-02154181