A data model for heterogeneous healthcare knowledge

Nigar Alishzade
Azerbaijan National Academy of Sciences Institute of Control Systems
Baku, Azerbaijan
nigar.alish@isi.az

ABSTRACT

Healthcare is a high priority sector where the highest level of care and service is expected. Due to the tremendous advancement of data acquisition in novel diagnostic devices the healthcare data is quite large and is moving to big data, which makes efficient the application of machine learning techniques in the analysis of this data. Before it comes to data, the knowledge in healthcare is of heterogeneous type. This paper provides a semantic data modeling method for the knowledge that comes from the healthcare sector to make applicable machine learning methods on it better.

KEYWORDS

Heterogeneous knowledge, knowledge graph, healthcare data, machine learning

CCS CONCEPTS

• Healthcare knowledge • semantic graphs • machine learning in healthcare

1 Introduction

The rapidly expanding field of big data analytics has started to play a pivotal role in the evolution of healthcare practices and research. To make the benefits of big data unlocked and harness the insights that come from it we need to manage and analyze the big data in a systematic manner [1]. Since in practice the biomedical knowledge usually comes from different domains and has different types, it meets the definition of heterogeneous knowledge. The huge size and highly heterogeneous nature of big data in healthcare render it relatively less informative using conventional technologies. Relying on the previous research [2], we assert that the knowledge graph is a suitable model for this purpose, due to its capability of expressing heterogeneous knowledge from various domains and fitting as many use cases as possible.

2 Health data in machine learning context

Traditionally, when faced with heterogeneous knowledge in a machine learning context, data scientists preprocess feature vectors (i.e. create dummy variables) so they can be used as input for learning algorithms. These transformations can result in loss of information and introduce bias that is unallowable in this domain. To solve this problem, we require a versatile data model to represent this heterogeneous knowledge.

Concretely, the term knowledge graph uses to refer to entities (things), their relations, and their attributes. For instance, in a healthcare database, we may find entities such as patients, medical workers, hospitals, health insurance companies, etc. Relations express which patient takes control of a particular doctor, which hospital a doctor works for, and so on. Attributes can be simple strings such as names and insurance numbers, but also richer media like short biographies, photographs like a medical card. Furthermore, these attributes can play the role of the labels in the supervised machine learning process on the next stages of the building of a healthcare system.

3 Current research challenges

In our study, we have gathered data from several sources and considering our aim to transform it into knowledge graph, the main breakthroughs we encountered are:

1. Data privacy and security: whilst the privacy problem is tackled by encoding, the security problem is remaining as an uphill task to build an efficient healthcare ecosystem.

2. Missing data: this problem can be handled by copying existing values for categorical records or filling average values of recorded observations for numerical records. But in clinical decision-making, this approach on electronic health records can be absurd, especially when the condition is still emerging.

3. Formatting features for model training: as we mentioned in section 2, our research implies to feed a machine learning model with the knowledge graph triples (entities, relations, and attributes). To do so, we have to preprocess the raw data have been gathered into a format that fits this structure. Unstructured data has to be streamlined via a method that ensures the least bias. Thus, we have chosen a distributed approach.

4 Conclusion and future work

Appropriate works [3], [4] show that the semantic graph approach promises better results and versatile solutions. Our work represents a data model for heterogeneous knowledge, to develop end-to-end machine learning that can directly consume health data. The main research gap that our work aims to cover, is semantic feature engineering on ontology triples of knowledge graphs. We suggest that further work along these tracks should aim the development of deep learning models on knowledge bases that will enable us to provide insights out of heterogeneous healthcare knowledge from different perspectives.

REFERENCES

[1] Ambigavathi M, Sridharan D (2020). A Survey on Big Data in Healthcare Applications. Intelligent Communication, Control and Devices, Advances in Intelligent Systems and Computing 989: 755-763. DOI: 10.1007/978-981-13-8618-3_77.

[2] Wilcke X, Bloem P, de Boer V (2017). The Knowledge Graph as the Default Data Model for Machine Learning. IOS Press: Amsterdam, The Netherlands, DOI: 10.3233/DS-170007.

[3] Irene Y. Chen, Monica Agrawal, Steven Horng, David Sontag (2020). Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. Pacific Symposium on Biocomputing 25:19-30.

[4] Zhenfeng Lei, Yuan Sun, Y.A. Nanehkaran, Shuangyuan Yang, Md. Saiful Islam, Huiqing Lei, Defu Zhang (2020). A novel data-driven robust framework based on machine learning and knowledge graph for disease classification. Future Generation Computer Systems 102: 534–548.