Knowledge Graphs to Help with Data-driven Clinical Decision-making

Diarmuid Cahalane*, Alexandre Kouznetsov† and Christophe Guéret*

Accenture Labs* and Accenture Applied Intelligence†
The Dock, 7 Hanover Quay, Dublin
diarmuid.cahalane@accenture.com
https://www.accenture.com/labs

Abstract: Clinical decision-making in a given case can benefit from the experience of treating similar, prior cases. We are investigating the potential for Knowledge Graph Embeddings to represent rich contextual information about a given patient case, and for that to enable the efficient retrieval and classification of prior cases which are decision-relevant.

Keywords: Knowledge Graphs, Knowledge Graph Embeddings, Electronic Health Records, Clinical Data

1 Background and Summary

For physicians, it can be complex to choose whether and how to intervene in a given patient case, using drugs, surgery or other therapies. This is particularly so in oncology, due to the genetic and highly dynamic nature of the disease. The choices and outcomes of other cases which share similar attributes can help inform decisions. However, identifying meaningfully similar cases and classifying them is an onerous task. We hypothesize knowledge graphs (KGs) representing patient data, and analytics applied to KG embeddings (KGEs) can help physicians with that task. We are building a proof of concept to test this. We will share our progress to date in this paper.

We constructed a KG representing medical records of 17,948 prostate cancer patients, including their medical conditions, drugs and medical procedures. We used AmpliGraph [1] to learn embedding vectors for the KG nodes. We isolated the KGE vectors representing patients and applied clustering techniques to partition the set of patient cases. Future work will seek to validate that the partitioning is clinically useful.

2 Our Approach

The nodes of our KG represented the 17,948 patients along with (distinct) instances and classes of their prescribed drugs (numbering 17,400), medical conditions (38,930) and a subset of their medical procedures (279). The graph contained over 3.3 million edges (relationships between pairs of nodes). The majority of the edges related patient nodes to the instances of drugs, conditions and procedures appearing in the respective patient medical records: our "hasPrescription" (c. 1.25 million), "hasDiagnosis" (c. 1.53 million) and "hasProcedure" (36,128) predicates. The remaining c. 0.5 million edges related instances to each other and to the classes of relevant hierarchies and ontologies. For drugs, conditions and procedures, in each case we used the classes and relationships found in an appropriate publicly available ontology or taxonomy.

AmpliGraph [1], an open-sourced suite of models for graph embedding, consumed the KG to generate 200-dimensional KGE vectors. Thus we learned vector representations for all the node types, and in particular the patient nodes. We used dimensional reduction to project the patient nodes' 200-dimensional KGE vectors to a low (three) dimensional space before applying the DBSCAN clustering algorithm [2].

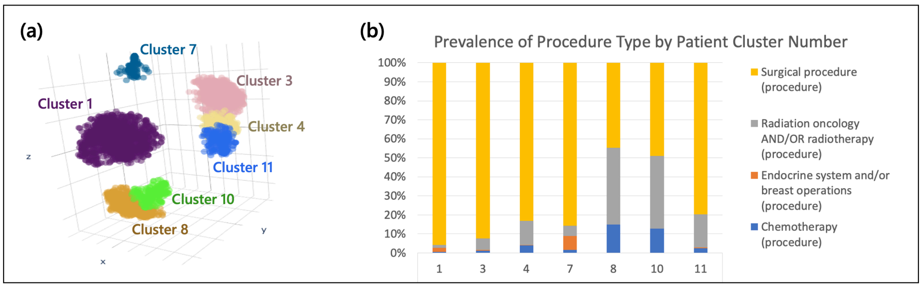

Fig. 1. (a) A subset of the patient clusters generated after projecting the embeddings into a 3D space. (b) The prevalence of different types of procedures (therapies) in patient medical records varied strongly by cluster.

3 Conclusions

We observed qualitatively that the KGE plus clustering approach yielded clusters of patients cases (Figure 1(a)) that varied strongly in the prevalence of diferenent classes of procedures, drugs and medical conditions (Figure 1(b)). If validated as clinically meaningful, this may help physicians to more efficiently find, making sense of and select among decision-relevant example cases.

4 References

1. Costabello, L., Pai, S., Van, C.L., McGrath, R., McCarthy, N., Tabacof, P.: Ampli-Graph: a Library for Representation Learning on Knowledge Graphs (Mar 2019). https://doi.org/10.5281/zenodo.2595043, https://doi.org/10.5281/zenodo.2595043
2. Ester, M., Kriegel, H., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.