




How to Empower Disease Diagnosis in a Medical Education System Using Knowledge Graph

Samuel Ansong^(✉) , Kalkidan F. Eteffa, Chao Li, Ming Sheng, Yong Zhang, and Chunxiao Xing

Research Institute of Information Technology, Beijing National Research Centre for Information Science and Technology, Department of Computer Science and Technology, Institute of Internet Industry, Tsinghua University, Beijing 100084, China
{ssml8, ajql8}@mails.tsinghua.edu.cn, {li-chao, shengming, zhangyong05, xingcx}@tsinghua.edu.cn

Abstract. Disease diagnosis is an important function in a medical training system, an integrated system which is aimed at providing the necessary skills and know-how to health practitioners. As one of the most vital features of a medical training system, many researchers and industry alike have channelled time and resources to engage several techniques and practices in a bid to find a way to accurately predict diseases with minimal margin of error. This has motivated several variations in feature selection, data representations and techniques in machine learning. In this paper, we explore some of these variations with prime focus on how knowledge graphs have helped address issues like insufficient data and interpretation to help empower the construction of a disease diagnosis feature in a medical training systems.

Keywords: Knowledge graph · Neural networks · Disease diagnosis

1 Introduction

Today, the pathway to build robust educational technologies is to integrate several vital features tailored for its target group. Medical training systems just like any other training system will provide a way for medical practitioners and medical researchers to accomplish tasks ranging from predicting adverse drug reactions, viewing statistical health data and insights, calculating patient similarity and diagnosing patients. A medical training system will complement traditional medical training methodologies to provide a holistic approach to medical training. There are several features that make up a medical training system, however the disease diagnosis feature is the more relevant for junior medical doctors. According to research by Singh, 5% of patients, that is approximately 12 million patients are misdiagnosed a year in the united states of America alone. This contributes significantly to the overall percentage of medical mistakes resulting in death or serious complications for patients, making it the third leading cause of death in the USA according to a recent publication by CNBC.

The disease diagnosis component of a medical training system demands a very high accuracy in its prediction task, an accuracy that cannot be guaranteed solely by the use of EHR (Electronic Health Record). It is therefore important that at the construction stage of such a feature the diagnosis of a medical practitioner is compared to that of a disease prediction algorithm and their differences reconciled by medical experts. However, building an accurate disease prediction algorithm does come with its own challenges. Competent and accurate implementation of such a system needs a comparative study of various techniques available.

This paper will explore how machine learning algorithms combined with knowledge graphs can improve the construction of a disease diagnosis feature in a medical training system. The paper will discuss how Knowledge graphs have contributed to resolving the challenges associated with EHR.

2 Methodology

The orientation of the paper follows the SLR guidelines for computer engineering. The key phases of this methodology include primary planning, selection and searching of primary studies which consists of formulating queries and keywords, performance of quality assessment and a selection procedure. The studies were limited to recent publications in the field of computer science and medicine which addresses the disease diagnosis task. Recent publications between the years 2013 and 2019 were considered.

3 Overview of Disease Diagnosis

Clinical decision making has evolved to become more complex and requires the evaluation of large volumes of data expressive of clinical information Liang et al. [1]. Artificial intelligence methods have emerged as possibly powerful tools to mine EHR data to aid in disease diagnosis and management. Data mining techniques have been adopted in variety of applications in the healthcare industry. For example, data mining proved essential in diagnosing diseases with good outcomes according to Zriqat et al. [2]. Due to the nature of the EHR, Hug et al. [3] suggests that comparable to spam filtering, sentiment analysis and language identification, disease prediction is an important medical text classification problem. Feature extraction from medical text has been a prevailing issue for most researchers over the years and several techniques have been deployed to address this issue, a study of eleven publications shows that most researchers favoured the bag of words and N-gram methodologies as seen in Fig. 1. However, BOW (Bag of Words) does not consider word order and grammar a worrying sign in medical text classification. To address this, several authors chose N-gram, although the approach takes into consideration word order, it does not solve the issue entirely since it does not account for word inversion, a persistent problem in healthcare data reports.

The general traditional approach process of diagnosing a patient can be seen as a classification task since it may be possible to achieve an acceptable level of conviction of a diagnosis with only a few features without having to process the whole feature set.

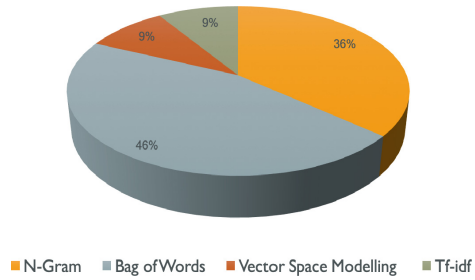


Fig. 1. Feature extraction techniques for medical text classification

This analogy has paved way for several implementations of classification algorithms for disease prediction. Researchers over the years have preferred neural networks over statistical approaches. These authors [4–6] used RNNs (Recurrent Neural Networks) in an attempt to model the sequential relations amongst medical codes. RNNs and all other deep learning algorithms however suffer from insufficient data and lack of interpretation. In the next section, we will focus on knowledge graphs and how they have helped in addressing these issues.

4 Knowledge Graph Methodology

Traditional machine learning algorithms take as input numeric or categorical qualities of an object known as a feature vector [7]. Knowledge graphs model information in the form of entities and relationships. Thus, a knowledge graph can contain an object together with its relationships to other objects in the form of a graph made up of nodes and labelled edges (which represents the relationships between the nodes). Knowledge Graphs (KGs) therefore offer semantically organized information that is interpretable by computers something that is needed in order to build more intelligent systems [7].

The speedy growth in volume and multiplicity of health care data from EHR and other sources [8] further necessitates the use of KGs. Training deep learning models typically involves large amounts of data that often cannot be met by a single health system or provider organization. Using knowledge graphs, distributed medical data sources can be aggregated into one meaningful data source. Several authors [8–14] have used knowledge graphs to solve the data inefficiency and interpretation drawbacks of neural networks. Choi et al. [8] applied knowledge graph in supplementing the EHR with hierarchical information from medical ontologies to improve interpretation whereas Ma et al. [9] used it to improve consistency by learning medical code representations. Authors [10–12] used knowledge graph to aggregate domain related data sources to solve the data insufficiency problem and achieved admirable results. Deploying knowledge graphs in the healthcare services space has proven to be an effective method to map relationships between the enormous variety and structure of healthcare data. Graphs provide an uncanny ability to model concealed relationships between information sources and capture linked information that other data models fail to capture. This enables researchers to more easily embed the information they need among a wide array of variables and data sources.

4.1 Knowledge Graph Representation

Machine learning on graphs is an important and pervasive task. The principal challenge in this area is finding an efficient method for the representation and encoding of graph structures in a way suitable for machine learning models [15]. Learning the representations of graphs is a hot research topic which has driven the proposal of various methods. Unlike traditional hand-engineered approaches, representation learning approaches treat this problem as machine learning task itself, using a data-driven approach to learn embeddings that encode graph structure [15]. Node embedding which is an approach use by several authors, is a process where nodes are encoded into low-dimensional vectors keeping intact their structure and relationships. DeepWalk, Node2Vec, LSHM, LINE, Metapath2Vec and Struc2Vec are some approaches proposed by authors.

All the listed methods focus on learning good representations for graph data, the majority of these methods belong to the direct encoding class under node embedding. Direct encoding however, fails to leverage node attributes which are sometimes highly helpful with regard to the node's position and role in the graph and are sometimes also computationally inefficient. To address these issues, methods like Deep Neural Graph Representations and SNDE (Structural Deep Network Embeddings) [16] have been proposed where graphs are directly incorporated into the encoder algorithm.

4.2 Discussion on Knowledge Graph Methodology

The task of disease prediction is an interesting field yet challenging one which has motivated several research and optimizations in pre-processing techniques, feature extraction techniques, feature selection techniques, algorithm design techniques and evaluation techniques. Knowledge graphs present us with the opportunity to incorporate large domain related datasets from arbitrary sources into one meaningful data source as input for neural networks to improve accuracy as demonstrated by authors [8, 10]. EHR are considered rich in data and can be greatly utilized [17]. In a sector like health where interpretation and accuracy is on a higher demand, knowledge graphs offer a way for understandable readings into the task of diagnosing a disease since it provides concise and meaningfully accurate relationships between features of a dataset. The integration of knowledge graphs into E-health systems will go a long way to drastically improve healthcare by offering an appreciable level of support to health professionals. In Liu et al. [12], the extraction of prediction rules generated from observing both classically professional paediatrics textbooks and clinical experiences of paediatric doctors respectively to form a knowledge graph indicates that junior doctors can benefit widely from existing knowledge graphs generated from more experience doctors.

The support from models powered by knowledge graphs could help junior doctors perform both differential diagnosis which requires an iterative step likened to the work of a classifier more efficiently and also allow junior doctors perform pattern recognition

diagnosis which requires the use experience to recognize a pattern of clinical characteristics. Pattern recognition diagnosis can only be achieved if a practitioner has enough experience. Representing the knowledge of experience doctors and clinical books as a knowledge graph as demonstrated in Liu et al. [12] will result in a more efficient disease prediction algorithm which can predict diseases with minimal error.

5 Conclusion

In this paper, we establish the fact that, in order to build an effective disease diagnosis, feature for a medical training system, a hybrid approach must be utilised where results from a disease prediction algorithm are compared to that obtained from querying an electronic health record and reconciled by experts. The paper then systematically review the various techniques and methodologies deployed by researchers in building accurate disease prediction algorithms whose results can be used to validate query results from an EHR in constructing an efficient disease diagnosis feature for a medical training system. The paper has shown that researchers over the years deployed knowledge graphs as a tool to remove ambiguity in medical concepts to help improve feature extraction. Also, researchers have applied knowledge graph as a tool to aggregate the various data sources of EHR and also provide sufficient input for deep learning models to help improve their accuracy. An accurate disease diagnosis algorithm will not only provide a means for validating diagnosis records in an EHR but can also become central in the construction of a diagnosis feature in a medical training system.

Acknowledgements. This work was supported by NSFC (91646202), National Key R&D Program of China (2018YFB1404400, 2018YFB1402700).

References

1. Liang, H., et al.: Evaluation and accurate diagnoses of paediatric diseases using artificial intelligence. *Nat. Med.* **25** (2019). <https://doi.org/10.1038/s41591-018-0335-9>
2. Zriqat, E., Altamimi, A., Azzeh, M.: A comparative study for predicting heart diseases using data mining classification methods (2017)
3. Huq, K.T., Mollah, A.S., Sajal, Md.S.H.: Comparative study of feature engineering techniques for disease prediction. In: Tabii, Y., Lazaar, M., Al Achhab, M., Enneya, N. (eds.) *BDCA 2018*. *CCIS*, vol. 872, pp. 105–117. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-96292-4_9
4. Choi, E., et al.: Multi-layer representation learning for medical concepts. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)* (2016)
5. Choi, E., et al.: RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In: *Advances in Neural Information Processing Systems (NIPS 2016)* (2016)

6. Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., Gao, J.: Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017), pp. 1903–1911 (2017)
7. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. *Proc. IEEE* **104**(1), 11–33 (2016)
8. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: GRAM: graph-based attention model for healthcare representation learning. In: KDD (2017)
9. Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., Gao, J.: KAME: knowledge-based attention model for diagnosis prediction in healthcare, pp. 743–752 (2018). <https://doi.org/10.1145/3269206.3271701>
10. Jha, A., et al.: Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer (2018)
11. Jiang, J., et al.: Medical knowledge embedding based on recursive neural network for multi-disease diagnosis (2018)
12. Liu, P., et al.: HKDP: a hybrid knowledge graph based pediatric disease prediction system. In: Xing, C., Zhang, Y., Liang, Y. (eds.) ICSH 2016. LNCS, vol. 10219, pp. 78–90. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59858-1_8
13. Yang, et al.: GrEDeL: a knowledge graph embedding based method for drug discovery from biomedical literatures. *IEEE Access*, p. 1 (2018)
14. Bean, D.M., et al.: Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.* **7**, 16416 (2017)
15. Hamilton, W.L., Ying, R., Leskovec, J.: Representation Learning on Graphs: Methods and Applications (2017)
16. Cao, S., et al.: Deep neural networks for learning graph representations. In: AAAI Conference on Artificial Intelligence (2016). pag.web. 14 April 2019
17. Tian, B., Xing, C.: Deep learning based temporal information extraction framework on Chinese electronic health records. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 203–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_19