FROM MEDICAL LITERATURE TO PREDICTIVE FEATURES: AN EVIDENCE-BASED KNOWLEDGE GRAPH APPROACH

Donghee Choi, Antoine D Lain & Joram M. Posma
Imperial College London
{donghee.choi,a.lain,j.posma11}@imperial.ac.uk

Mark Kozdoba[1], Binyamin Perets[1] & Shie Mannor[1,2]
[1] The Technion, IIT [2] NVIDIA Research
{markk@,sbp67250@campus.,shie@ee.}technion.ac.il

1 INTRODUCTION

Medical and biological studies often face a fundamental challenge: while large amounts of measurements can be collected per subject (biomarkers, genetic variants, etc.), the number of subjects typically remains limited compared to the number of features. Moreover, it is often the case that individual biomarkers, while relatively easy to measure, are only weak indicators of the more high level prediction questions (for instance, a biomarker that may be a symptom for many different otherwise unrelated disease). In such cases however, the *joint* information from multiple biomarkers may still be informative, indicating that the number of features can not be trivially reduced and all features may be required for good predictions to be possible.

This issue, known as the low sample-to-feature ratio, James et al. (2013), and closely related to the "curse of dimensionality" phenomenon, considerably complicates all prediction tasks in medical and biological domains. It typically implies that practitioners must choose between simple models that may miss complex relationships and complex models that risk overfitting.

Another common way to improve the performance in low sample-to-feature ratio situations is via the incorporation of *prior knowledge*. Perhaps one of the most obvious and interesting sources of prior knowledge is the formal academic literature. Indeed, centuries of medical research produced a wealth of knowledge about relationships between various medical entities. However, this knowledge typically takes the form of *evidence* - i.e. documented qualitative directional associations between entities (e.g., "calcium intake decreases blood pressure") rather than direct measurements of relationship strengths. As such, it is challenging to integrate this knowledge into traditional machine learning predictive models.

In this paper, we introduce a framework for the use of the literature information in machine learning models. The framework consists of three main components: **(i)** A construction of a Knowledge Graph (KG), which contains the evidence based relationships extracted from literature. **(ii)** A construction of a probability model which is as *consistent* with the above relationships as possible. The model provides a qualitative, coherent, and aggregated *representation* of a diverse and sometimes conflicting literature evidence. **(iii)** The use of the model to augment features in a given dataset with new KG features, and the use of *sparse* classifiers to exploit the added KG features for prediction.

We now describe these stages in slightly more detail.

2 METHOD OVERVIEW

2.1 KNOWLEDGE GRAPH CONSTRUCTION FROM LITERATURE

The knowledge graph (KG) used in this work is derived from on-going research focused on extracting structured biomedical knowledge from unstructured literature. The preliminary work consortium (2024) collected Open-Access articles that permit redistribution from publicly available scholarly

databases, including PubMed Central and Scopus. A total of 3,598 full-text documents were processed using current natural language processing (NLP) techniques based on pre-trained language models. Building on this collection of articles, our KG construction pipeline consists of three main steps: named entity recognition (NER), entity linking (normalisation) and relation extraction (RE).

First, a biomedical NER model identifies key entities categorised into 13 biomedical types, including diseases and phenotypes, potential biomarkers (e.g. SNPs, genes, proteins, metabolites), and diets/foods. After these are given standard identifiers based on existing ontologies, the extracted entities are then linked through an RE model (using an open-weight large language model trained with a novel RE dataset) that classifies relationships into eight predefined categories (from 'association' to 'unrelated', and 'correlation' to 'causal'). This large-scale processing of full-text documents enables the KG to capture a broad set of biomedical entities and their interactions, forming a structured representation of literature-derived knowledge. To give an example, a relationship could be: "*sitosterolemia* (a genetic disorder,entity) and *Niemann–Pick C1-like Protein* (a well studied biomarker, entity) are *positively correlated*". Note that the literature, and thus the extracted relationships, capture the direction of the association, but are otherwise qualitative.

2.2 A KG Consistent Probability Model

Considering the set of entities as variables, we construct a probabilistic model of a joint distribution of these variables, that is as *consistent* with the relationships above as possible. While the relationships naturally land themselves to be interpreted as covariances, there is no straightforward way to directly obtain a covariance matrix from the relations. Indeed, first, the relationships extracted from different papers may be conflicting, and thus some conflict aggregation/resolution mechanism must be present. And second, the true covariance matrices are structured (e.x: if X_1 is highly correlated with X_2, and X_2 is highly correlated with X_3, then X_1 is quite highly correlated with X_3). This structure would not necessarily show up directly in the KG, since, for instance, the relevant evidence may simply be missing.

We solve the above issues by performing a regularized optimization in the space of covariance matrices, where the KG relationships are viewed as soft constraints, and the regularization imposes maximum independence (equivalently maximum entropy) under the constraints.

2.3 Feature Augmentation And Prediction

Given a fixed dataset, we use the model constructed above to *augment* the given features with additional (termed *latent*) features, derived from the graph. The classifiers are then applied to the augmented dataset, *with sparsity constraints*. The advantage of this augmentation is that the extended features may represent higher level, more informative and aggregated signals (such as a disease, rather than just one of its particular biomarker level symptoms), which in turn can then be used by sparse classifiers to improve the performance while maintaining the generalization ability. It is also worth emphasizing that the relation between the dataset features and the higher level signals is encoded by the KG and its model, and typically can not be obtained from the dataset alone. Finally, the use of sparsity is critical, as it allows us to obtain better generalization despite having a bigger feature set. In the augmentation process, the dataset features are first matched with some subset F of all KG entities E. Then, the KG probability model is used to augment the features in F by the features in $E \setminus F$. This is done by taking conditional expectations w.r.t the KG model, of values in $E \setminus F$, conditioned on values in F. This procedure is somewhat similar to that of some *data imputation* methods,Emmanuel et al. (2021), but our probability model is external to the dataset.

3 Experimental Evaluation

A preliminary evaluation of the approach was performed on the NutriTech dataset, Rundle et al. (2023). In this dataset, there are 32 subjects, for whom we predict several post-diet body composition metrics, based on about 130 pre-diet biomarkers (normalised to identifiers found in the KG), and on 200 additional features from KG augmentation. On one of the five uncorrelated body composition markers (liver fat) we have found a significant performance improvement for the classifier with augmentation, compared to the non-augmented features baseline. Here the significance was evaluated on 1000 random train/test splits, with an 18% error decrease on average.

MEANINGFULNESS STATEMENT

A meaningful representation of life is a model of a living system which helps making verifiable predictions about the future, or about unobserved parts of the system in present. Such representations would typically describe the entities in the system, the interactions between them, and possibly the evolving nature of both.

The literature is arguably humanity's best resource identifying what the interesting entities are, and some of their interactions. However, papers mostly are *local*, i.e. describing few entities in isolation. In this work we make the first steps towards synthesizing a global and quantitative representation from this knowledge.

ACKNOWLEDGMENTS

This work was funded by the European Union under Horizon Europe grant number 101084642 and supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 101084642].

REFERENCES

CoDiet consortium. Release of an automatically annotated and manually verified corpus of 1,000 open access articles. Technical report, Imperial College London, 2024. URL `https://cordis.europa.eu/project/id/101084642/results`.

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big data*, 8:1–37, 2021.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Milena Rundle, Jarlei Fiamoncini, E Louise Thomas, Suzan Wopereis, Lydia A Afman, Lorraine Brennan, Christian A Drevon, Thomas E Gundersen, Hannelore Daniel, Isabel Garcia Perez, et al. Diet-induced weight loss and phenotypic flexibility among healthy overweight adults: a randomized trial. *The American Journal of Clinical Nutrition*, 118(3):591–604, 2023.