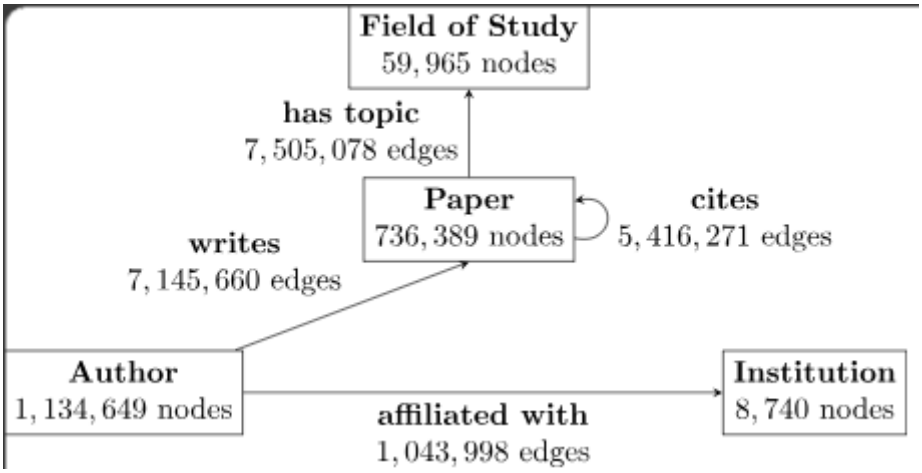
Heterogeneous Graph Learning

Knowledge graphs are visualization of information with multi-type relations (edges) among some multi-type entities (nodes) within an environment of interest. The heterogeneous aspect comes from the graph having two or more types of nodes and two or more types of edges.



Usage: Alzheimer's Drug Repurposing

In recent studies on Alzheimer's drug repurposing (Reference 1) Graph Neural Networks (GNN) were used to identify biological interactions from prior knowledge databases containing information on effective treatments and genes associated with high risk. The nodes of the graph included drugs, genes, pathways and gene ontology (GO) connected by interactions including drug-target interaction, drug-drug structural similarity, gene-gene interaction, gene-pathway association, gene-GO association and drug-GO association. A comprehensive graph can be created combining existing information from multiple sources on the biological interactions of complex drug-gene relationships, and from there we can use machine learning to train a model and address the incomplete knowledge in the graph. Note that in this paper the genes are encoded/embedded, simply meaning they are represented as numerical vectors/matrices to capture their function. To give a broad overview of the study's workflow:

1. A knowledge graph is built to describe the interaction between drugs, genes, gene ontology and pathways
2. Node embeddings are derived using a multi-relational Variational Graph Auto-Encoder (VGAE)
3. Machine learning model ranks drug-candidates based on multi-level evidence
4. Drug combinations are searched for complementary exposure patterns, using previously ranked drug candidates

5. Validate drug combinations with "oxidative stress responses" and randomly removing edges to see if the model can correctly predict interactions

By keeping the focus previously approved drugs, the graph can identify synergy between medications that treat complex diseases. The "Complementary Exposure Pattern" of analysis used suggests drug combinations are effective when the target of the drugs hit the disease module without overlap.

While this study breaks ground in applying knowledge graphs with multi-task learning to fragmented multi-modal data, it is subject to the limitations of each dataset it combines. The knowledge graph alone extracts data from:

- 5,874,261 Universal protein-protein interactions from STRING, Drug Bank, HetioNet, GNBR, and IntAct
- Interaction between genes, drugs, GO, and pathways from Comparative Toxicogenomic Database (CTD)
- Drug-Drug associations based on structural similarities using scores from the RDkit package
- Gene IDs from National Center for Biotechnology information (NCBI)
- High confidence Alzheimer's associated genes from Agora's nominated gene list

If any of the data-driven drug efficacy is biased, then so is the model. Some of the datasets contained *in vitro* studies (lab based experiments usually in a test tube), which does not guarantee identical treatment outcomes. Still, given the lack of clinical evidence from a huge list of compounds, this research could provide insight on drug combinations in future research opportunities.

VGAE

One of the key elements in the knowledge graph is the customized deep Variational Graph neural Auto-Encoder (VGAE), used to incorporate multiple types of relationships (edges). It is a self-supervised technique which encodes the nodes into a latent vector (embedding) and reconstruct the graph with the latent vector. This approach allows the incorporation of the uncertainty of our knowledge of a given node, thus creating a probabilistic distribution. Different weight matrices were also determined for each type of edge.

Within the Alzeimer's drug GNN Pytorch Geometric was used for the model implementation. It was trained to reconstruct missing interactions using the node embeddings as an autoencoding manner. To do this, each node's embedding is iteratively updated by aggregating their neighbors' embedding, in which the average of the neighbors' features are taken, concatenated with the current embedding, and applied to a single layer of a neural network.

Transfer Learning

Transfer learning is to transfer knowledge from previously learned universal models to domain-specific models. Injecting universal knowledge on pharmacological interactions while prioritizing

Alzheimer's is a critically important step in identifying uncertainty. The **Drug Repurposing Knowledge Graph** (DRKG) network was used as the universal data source, containing over 15 million pharmacological entities including 39 different interactions between compounds and genes from various biomedical databases. The pretrained embeddings for drugs, genes, pathways and GOs were used as initial node features and then fine-tuned, but the edges were not always directly incorporated as the interactions were less accurate.

Node Classification

A node classification task was used to differentiate 743 Alzheimer's associated genes from the remaining genes. The ongoing hypothesis is that one can predict these genes using the gene interaction with other genes, GO, and pathways. This classification tool consisted of 2 GraphSAGE graph convolution layers that sample a node's adjacent neighbors and aggregate their representation. This was jointly optimized with the VGAE. PcGrad, a research technique that resolves conflict among multiple tasks' gradients by normal vector projections, was also used to achieve a better local optimum.

Usage: SPOKE

In *Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis* (reference 2) a modified version of the PageRank algorithm (the algorithm used by Google to rank web pages) was implemented to embed millions of EHRs into a biomedical knowledge graph (SPOKE), which contains information from 29 public databases. The high-dimensional patient health signatures were subsequently used as features in a random forest to classify patients risk of MS.

The PageRank algorithm outputs a probability distribution to represent the likelihood a person randomly clicking links will arrive at any particular page. This probability can be calculated for collections of documents of any size. At the start it is assumed the distribution is evenly divided but over several iterations the values are adjusted to their theoretical value. If any document links to another document, the destination document probability rises. In SPOKE, we are looking for connections between medication, diagnosis, genes and etc rather than links on webpages, and iterations are continued until the target threshold is reached. Then the final node ranks are used to create the weights for the patient population, called the Propegated SPOKE Entry Vector (PSEV). Then vector/matrix arithmatic is applied to produce the Patient-Specific SPOKE Profile Vectors (SPOKEsigs), and random forest classifiers are used to determine their significance in predicting the disease.

So to summarize the above in the fewest words, the process is:

1. Find overlapping concepts between SPOKE and EHR
2. Chose any term or concept from the graph to make the cohort
3. Perform PageRank
4. Final node ranks are used to create weights

References

[1] Synthesize heterogeneous biological knowledge via representation learning for Alzheimer’s disease drug repurposing

[2] Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis

---

Revision #18
Created 21 April 2024 20:59:38 by Elkip
Updated 2 April 2025 02:18:51 by Elkip