MASTER'S THESIS

Memory-Efficient Learning Algorithms for Feature and Relation Extraction with Applications to Genomics Data and Clinical Notes

GENG, Yu

Date of Award:
2024

ABSTRACT

Clinical and genomics data have been made widely available where more personalized medicine and effective scientific discovery can be supported. One of the fundamental challenges is how to extract salient features and their relations from the raw data to facilitate the downstream analytics tasks. For genomics data, one important type of feature to extract is the regulatory relationship among the genes from the gene expression data. The problem is often called the Gene Regulatory Network (GRN) Reconstruction. This is important as the functional properties of a cell are known to be characterized by its gene regulatory interaction. For clinical data, extracting relations between entities detected in clinical notes is another important analytics task. For instance, the detection of adverse drug events from clinical notes, often known as Adverse Drug Event (ADE) Extraction, is crucial for pharmacovigilance. The consequence of the ADE could be fatal, and thus identifying them is important for patient safety.

Firstly, we address GRN Reconstruction challenges by proposing novel methods to infer GRNs from expression data, capturing the complex nonlinear relationships among genes. The methods leverage the Hilbert-Schmidt Independence Criterion (HSIC) to estimate nonlinear interactions, aiming to overcome limitations in existing linear models and reduce false positive identifications. Furthermore, we apply Factorization Machines (FM) as a supervised learning method to reduce the false negative predictions caused by the unsupervised method. Secondly, we introduce a memory-efficient FM model for capturing nonlinear feature interactions. Our approaches apply element-wise feature mapping for both individual features and feature interactions and further binarize model parameters, which significantly achieve high performance while reducing memory requirements. This is particularly beneficial for resourceconstrained applications such as mobile or IoT devices. Lastly, our research delves into pharmacovigilance by enhancing ADE Extraction from clinical notes. We integrate biomedical domain-specific knowledge into Transformer models and utilize Generative Pre-trained Transformer (GPT) for data augmentation, improving the model's capability to identify complex clinical entities and their relations. The methodologies not only enhance the accuracy of ADE identification but also contribute to efficient pharmaceutical analysis.

Keywords: Gene Regulatory Network Reconstruction, Adverse Drug Event Extraction, Memory-Efficient Algorithms