DOCTORAL THESIS

Graph Network Representation of Traditional Chinese Medicine Prescriptions: Bridging Ancient Wisdom with AI Model Development

CHENG, Xing Ye

Date of Award:
2023

ABSTRACT

Traditional Chinese medicine (TCM) is a traditional medical system with a rich historical background that has garnered global attention and study in recent years. TCM's fundamental concept revolves around syndrome differentiation, wherein diagnosis and treatment plans are formulated based on a combination of symptoms. This approach emphasizes the holistic consideration of the entire human body and establishes a macroscopic model. Presently, TCM research primarily focuses on pharmacology, delving into the identification of ingredients, targets, and biological processes associated with herbs. Such bottom-up investigations concentrate on the specific details of TCM but fail to adequately recognize its inherent holistic perspective. Hence, the need for a top-down TCM research framework arises, aiming to offer effective guidance for bottom-up TCM research and facilitate the scientific comprehension of TCM itself. This thesis endeavors to analyze TCM prescription data from a top-down standpoint, encompassing the construction of a graphical representation of prescriptions and employing network medicine and Artificial Intelligence (AI) techniques to examine it, along with an initial exploration of an end-to-end AI-based prescription generation model.

In this thesis, we made contributions in the following aspects.

1. The process of standardizing TCM prescription data involved the collection of more than 80,000 prescriptions, followed by data cleaning and standardization using various techniques such as statistical methods, word segmentation algorithms, manual inspection, and a preliminary attempt with a large language model (LLM). The standardized dataset encompasses symptoms, syndromes, and herbs associated with each prescription, which were subsequently utilized in further analysis to enhance the normalization of TCM terminology.

2. A graph representation of prescriptions was established, and various analysis results were obtained. Based on the co-occurrence of symptoms, syndromes, and herbs in the standardized TCM prescription data, this thesis selected relevant correlations and constructed a three-layered network which consisted of 705 common herbs, 981

symptoms, and 143 syndromes. The structural properties of the herb layer and symptom layer, such as degree distribution, betweenness distribution, and clustering coefficient, were analyzed. The communities and their sub-communities were found through community detection on the networks. The findings revealed a hierarchical structure in both networks, highlighting the presence of hubs and transportation junctions. Building upon the structure of this three-layer network, two applications were explored. The PageRank algorithm was employed to determine the importance of herbs in the prescription, aligning with the Chief-Deputy-Assistant-Envoy theory in TCM, albeit in a more quantitative manner. Additionally, the network facilitated the visualization of disease progression. During the course of progression, changes in symptoms on the network were centered around nodes with high betweenness, starting from nodes with low betweenness, traversing nodes with high betweenness, and ultimately reaching other nodes with low betweenness.

3. This study delved into an in-depth analysis of the three-layered network using AI models and explored their potential applications. Specifically, the Graph Neural Network (GNN) was employed to predict the functions of herb nodes within the herb network and simultaneously obtain herb embeddings. This model facilitated the prediction of potential herb functions, with some of the results being supported by ancient TCM books. Additionally, the herb embeddings derived from the AI model reflected traditional herb properties and enabled the effective clustering of prescriptions. In conjunction with the hierarchical structure of prescriptions, the high-frequency herbs in clusters matched similarities or even identical actual prescriptions in the data, aligning with historical practices. The GNN was also capable of predicting syndromes based on the subgraph of herbs or symptoms. This model supplemented prescriptions with missing syndrome data through its predictions, which were corroborated by TCM databases. Remarkably, all the aforementioned models achieved their performance without relying on herb properties, TCM knowledge, or semantic information about symptoms, surpassing scenarios where such information was incorporated. These findings indicate that herb and symptom networks convey medical meaning more effectively than traditional herb properties and symptom semantics.

4. An exploratory effort was made to design an end-to-end AI-based prescription generation model. In this regard, the Transformer architecture was employed to generate prescriptions with complex nonlinear properties, outperforming the Fully

Connected Neural Network by a significant margin. The attention mechanism with interpretability shed light on the prescription generation process and the relationships between symptoms and herbs. Certain herbs obtained a preferential focus and association with other complementary herbs, and accurately predicting the initial herbs played a pivotal role in the success of the model. These findings align with the actual TCM diagnosis and treatment process and lay the groundwork for the future development of practical AI-based TCM diagnosis and treatment systems.

The ancient books and prescriptions of TCM substantiate several findings presented in the graph network representation of prescriptions. These findings illustrate the parallels between statistical and AI-based approaches and the profound wisdom of the ancients. The findings and outcomes presented in this thesis offer valuable insights and perspectives for future research endeavors in the field of TCM.