Abstract

In the era of information technology advancement, data are being collected and accumulated at a dramatic pace. The ability to extract useful hidden knowledge from a large amount of data and make decisions based on this knowledge has offered significant benefits to business organizations. Predicting health risks of patients using Electronic Health Records (EHR) has attracted considerable attention in recent years, especially with the development of deep learning techniques. Health risk refers to the probability of the occurrence of a specific health outcome for a specific patient. The predicted risks can be used to assist the decision-making of healthcare specialists. EHRs contain a chronological set of medical records, and within each medical event, there is a set of clinical/medical activities. Various risk prediction models have been developed and introduced to health risk prediction on EHR data.

Predictive analytics is the process of analyzing large datasets to discover patterns and meaningful insights where machine learning and statistical methods are applied to build predictive models for decision support. However, EHR data has its unique characteristics, such as high dimensionality, sparsity, irregularity, heterogeneity, random errors, temporality, and systematic biases. It is technically challenging to apply existing risk prediction models to EHR data with a high degree of irregularity, which contains many missing values and varying time intervals between medical records. Existing studies on EHR data irregularity have been focused on the provision of deep learning-based solutions. These studies impute missing values by incorporating deep neural networks to learn variable correlations and introduce time decay mechanisms to capture the impact of varying time intervals. The complete data matrices obtained from the imputation task are used for downstream risk prediction tasks. However, the existing imputation methods lead to less than desirable prediction accuracy. Further improvements in risk prediction models are necessary before they can be adopted for real-world applications.

This thesis investigates and develops new risk prediction models for handling the irregularity of EHR data and predicting patients' health risks. We proposed compound density networks, an end-to-end, novel, and robust model to train the imputation method and prediction model simultaneously within a single framework. The purpose is to handle missing values in EHR data, enhance imputed values' reliability, and quantify their uncertainties. We then developed deep imputation-prediction networks by extending the compound density network to perform imputation and prediction in EHR data. The focus is to capture the impact of varying time intervals. We further introduced multi-task learning to perform risk prediction tasks by incorporating the imputation task as an auxiliary task while carrying out both simultaneously.

To improve imputation performance, we considered patient similarity via incorporating graph analysis techniques. We proposed contrastive learning-based imputation prediction networks, which mainly impute missing values in EHR data by exploiting similar patient information as well as patients' personal contextual information. Similar patients are generated from patient similarity computing during stratification modeling and analysis of patient graphs. We further proposed contrastive graph similarity networks, which incorporate graph contrastive learning in representation learning for EHR data. The graph similarity networks were extended to multi-graph neural networks. This enables the learning of multiple graph structures from input EHR data, which aggregates the information from similar patients to offer a richer representation of the patient and allows the extraction of patient health context for both imputation and prediction tasks. The result is an optimised graph structure that incorporates the characteristics of these graphs with attention mechanisms.

We empirically investigated the proposed deep imputation-prediction models on two tasks using the Medical Information Mart for Intensive Care Database and eICU Collaborative Research Database. These tasks are (i) multivariate clinical time series imputation and (ii) in-hospital mortality risk prediction. The empirical results indicated that our models outperform state-of-the-art imputation-prediction models by significant margins.

The strength of these models lies in their ability to present transparency and interpretability of the decision process and provide the estimation of epistemic and aleatoric uncertainties of the model decisions.

Our research work made novel contributions to the improvement of methodologies for dealing with the irregularity of EHR data in the context of health risk prediction. These methodologies are potentially applicable to other medical applications such as hospital length of stay prediction and phenotype classification.