［課程－２］

審査の結果の要旨

氏名　馬　驍珺

Deep phenotyping, observing patient phenotypic abnormalities, is an essential component of precision medicine which aims to provide best available care for each patient based on stratification into disease subclasses. This dissertation proposed to present patients' data through disease ontology (presenting a disease as a causal chain of abnormal states) grasping the transition of disorders in order to realize deep phenotyping and disease progression management. For that purpose, the author put forward to a framework of constructing a system of automatically mapping electronic health records (EHR) data to disease ontology. The framework consists of methodology of building mapping modules from heterogeneous EHR data and a system composed of these modules. The system automatically extracts information from EHR, identifies presence of abnormal states, and maps them to disease ontology. Taking chronic kidney disease (CKD) as an exemplar, this dissertation constructed a mapping system which was evaluated based on EHR data from The University of Tokyo Hospital. The experimental results demonstrated that the proposed framework has achieved high performance in identifying those abnormal states, which can easily reach agreement among medical experts when annotation. The significance and positive outcomes drawn from the dissertation are summarized as follows:

- Social implication

The contribution of this study first comes from the combination of disease ontology and real EHR data to identify patients' fine-grained abnormal states serving clinical applications with significant social implications.

(1) This study proposed to apply disease ontology to curate EHR data for the purpose of identifying medical concept (abnormal states), which is very promising in contributing to deep phenotyping.

(2) Through mapping EHR data to the disease ontology which describes the transitions between abnormal states, this study made it possible to visualize the disease progression for individual patients.

- Information technology

Wide adoption of EHR have let to fast accumulation of patient data, where high throughput technologies are required for its analysis. However, EHR data is characteristic of heterogeneity, sparsity, noise, and incomplete data, which poses challenges in its information extraction. Furthermore, lack of labeled medical data hampers the utilization of the power of machine learning approaches which have seen rapid advance over recent years. This study solves these problems by introducing the following information technologies, which seem to be interesting to scientists in biomedical informatics.

(1) This study introduced EHR-driven unsupervised approaches to solve the heterogeneity of medical concept expressions in clinical texts.

(2) This study exploited the causal chain structure of disease ontology to implement imputation, solving data sparsity problem.

(3) This study proposed a method introducing the idea of transfer learning and active learning, which improves the generalizability of the proposed framework.

This dissertation addresses interesting topics in biomedical informatics and represents high level scientific work. This dissertation meets the requirements for awarding a PhD degree in Medicine.