Modeling Healthcare Data Using Multiple-Channel Latent Dirichlet Allocation

盧信銘
臺灣大學資訊管理學系

Abstract

Information and communications technologies have enabled healthcare institutions to accumulate large amounts of healthcare data that include diagnoses, medications, and additional contextual information such as patient demographics. To gain a better understanding of big healthcare data and to develop better data-driven clinical decision support systems, we propose a novel multiple-channel latent Dirichlet allocation (MCLDA) approach for modeling diagnoses, medications, and contextual information in healthcare data. The proposed MCLDA model assumes that a latent health status group structure is responsible for the observed co-occurrences among diagnoses, medications, and contextual information. Using a real-world research testbed that includes one million healthcare insurance claim records, we investigate the utility of MCLDA. The results indicate that MCLDA is capable of capturing the comorbidity structures and linking them with the distribution of medications. Moreover, MCLDA is able to identify the pairing between diagnoses and medications in a record based on the assigned latent groups. MCLDA can also be employed to predict missing medications or diagnoses given partial records. Our experimental results show that, in most cases, MCLDA outperforms alternative methods such as logistic regressions and the k-nearest-neighbor (KNN) model for the medication and diagnosis prediction tasks. Thus, MCLDA is a promising approach to modeling healthcare data for clinical decision support.