ABSTRACT

Method to Impute Missing Features in Metabolomics Data using Rank-Transformation and Matrix Factorization

Metabolomics data helps computational oncologists identify biomarkers of disease and therapeutic response. During the conversion of a raw tissue sample into an observation in a metabolomics data matrix, the abundances of key metabolites of interest may not be recorded. Rather than discard the data and repeat the study which is missing measurements for metabolites of interest, a researcher could instead augment their old data with imputed metabolite abundances, saving time and money. This thesis describes a method to impute the abundances of unmeasured metabolites in batches of metabolomics data using non-negative matrix factorization. The method learns the abundance of an unmeasured metabolite by modeling metabolite covariation in datasets where the metabolite of interest is measured. It then transfers this knowledge to a hold-out dataset where the metabolite of interest is unmeasured. The effectiveness of this imputation method is benchmarked with three test cases on Memorial Sloan Kettering's nine batch pancancer metabolomics dataset. The significance of the method is demonstrated in use cases such as biological interpretation of embeddings and active learning of core features. The sensitivity of the model is analyzed through experiments on simulated metabolomics data.