Reasonably integrating data for predicting the drug toxicity by machine learning

Naiyang Guan[1], Zhilong Jia[2], Xiang Zhang[1], Bin Luo[1], Bin Mao[2], Qing Liao[3] and Zhigang Luo[1]

[1] National Laboratory for Parallel and Distributed Processing, School of Computer Science, National University of Defense Technology, Changsha, China
[2] Department of Chemistry and Biology, College of Science, National University of Defense Technology, Changsha, China
[3] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China

E-mail: zhilongjia@gmail.com

The toxicogenomics is usually expected to aid in the risk assessment of drugs. Drug-induced liver injury (DILI) is a leading reason of drugs failing during clinical trials as well as being withdrawn from the market. In this paper, we focused on the two problems in prediction of DILI by analyzing the toxicogenomics data provided by CAMDA2013.

Firstly, we predict the DILI using a more reasonable data-collapsing method although with a relatively lower classification performance. Some of previous work consider the data-collapsing but neglect the following objective situation. During predicting the risk of a new drug, the drug toxicity in any doses and any time-point conditions should be blind to us. More specifically, for predictive machine learning models, it is more reasonable to use some drug profiling data for training while test on another drugs profiling data. Thus, we present a new and more reasonable method for collapsing the multi-doses and multi-time-points expression profiling data of microarray. We use the averaged value and maximum value of the differential expression values in all doses and time-points for each drug within each gene with or without trim to construct the two datasets for learning. By using some classification methods including both LDA and linear SVM with ten folds cross-validation, we found the accuracies of predicting DILI are all about 60% using the rat in vivo single dose type profiling data. This unsatisfactory result may be caused by the fact that different drugs may result in DILI in various approaches. In other words, this problem becomes a small sample problem in spite of many features. Therefore, the differential expression genes have litter in common among the profiling of drugs. This result suggests that subtyping the coarse-grained DILI based on the related pathways of differential expression genes may contribute to the ultimate risk assessment of new drugs.

Secondly, we utilize Canonical Correlation Analysis (CCA) to integrate the profiling data from rat in vivo single and in vitro for predicting DILI. A typical use for CCA is to take two sets of variables, e.g., profiling data from rat in vivo single and in vitro, and see what is common subspace, e.g. DILI, amongst the two sets. So, it is suitable to integrate these data via CCA. Using the integrated data, the prediction result is not better than the result from that of rat in vivo. Our result reveals that the agreement between in vivo and in vitro is poor for predicting the DILI. It seems that the noises in each dataset mask the common profiling pattern of DILI or the subtypes of DILI may result in a less common profiling pattern of DILI.

In summary, we present a more reasonable data-integrating method for the classification of IDLI and the poor agreement between in vivo and in vitro for predicting the DILI. It seems that subtyping the DILI may be essential to better assess the risk of new drugs.