A robust re-rank approach for feature selection and its application to pooling-based GWA studies

Jia-Rou Liu[1]* (劉佳柔) and Hung Hung[2] (洪弘)
[1]Instituite of Statistical Science, Academia Sinica
[2]Institute of Epidemiology and Preventive Medicine, National Taiwan University,

Abstract

Large-*p*-small-*n* datasets are commonly encountered in modern biomedical studies. To detect the difference between two groups, conventional methods would fail to apply due to the instability in estimating variances in *t*-test, and a high proportion of tied values in AUC (area under the receiver operating characteristic curve) estimates. The Significance Analysis of Microarrays (SAM) may also not be satisfactory, since its performance is sensitive to the tuning parameter, and its selection is not straightforward. In this work, we propose a robust Re-Rank Approach to overcome the above-mentioned difficulties. In particular, we obtain a rank-based statistic for each feature based on the concept of "rank-over-variable". Techniques of "random subset" and "re-rank" are then iteratively applied to rank features, and the leading features will be selected for further studies. The proposed Re-Rank Approach is especially applicable for large-*p*-small-*n* datasets. Moreover, it is insensitive to the selection of tuning parameters, which is an appealing property for practical implementation. Simulation studies and real data analysis of pooling-based genome-wide association (GWA) studies demonstrate the usefulness of our method.

Keywords: large-*p*-small-*n*, feature selection, rank-over-variable, re-rank, random subset