Interaction-Based Feature Selection and Classification for High-Dimensional Biological Data

Maggie Haitian Wang[1], Shaw-Hwa Lo[3], Tian Zheng[3], and Inchi Hu[2]

[1]: Division of Biostatistics, School of Public Health and Primary Care, CUHK, Shatin, Hong Kong

[2]: Department of ISOM, HKUST, Clearwater Bay, Kowloon, Hong Kong

[3]: Department of Statistics, Columbia University, New York, USA

Emails: MHW (maggiew@cuhk.edu.hk); SHL (slo@columbia.edu); ZT(tz33@columbia.edu); IH (imichu@ust.hk)

Epistasis or gene-gene interaction has gained increasing attention in studies of complex diseases. Its presence as an ubiquitous component of genetic architecture of common human diseases has been contemplated. However, the detection of gene-gene interaction is difficult due to combinatorial explosion. We present a novel feature selection method incorporating variable interaction. Three gene expression datasets are analyzed to illustrate our method, although it can also be applied to other types of high-dimensional data. The quality of variables selected is evaluated in two ways: first by classification error rates, then by functional relevance assessed using biological knowledge. We show that the classification error rates can be significantly reduced by considering interactions. Secondly, a sizable portion of genes identified by our method for breast cancer metastasis overlaps with those reported in breast cancer database as disease associated and some of them have interesting biological implication. In summary, interaction-based methods may lead to substantial gain in biological insights as well as more accurate prediction.

Key Words: Interactions effect, boosting, microarray