

Informative top-k class associative rule for cancer biomarker discovery on microarray data

ABSTRACT

The discovery of reliable cancer biomarkers is crucial for accurate early detection and clinical diagnosis. One of the strategies is by identifying expression-based cancer biomarkers through integrative microarray data analysis. Microarray is a powerful high-throughput technology, which allows a genome-wide analysis of human genes with various biological information. Nevertheless, more studies are needed on improving the predictability of the discovered gene biomarkers, as well as their reproducibility and interpretability, to qualify them for clinical use. This paper proposes an informative top-k class associative rule (iT CAR) method in an integrative framework for identifying candidate genes of specific cancers. iT CAR introduces an enhanced associative classification algorithm that integrates microarray data with biological information from gene ontology, KEGG pathways, and protein-protein interactions to generate informative class associative rules. A new interestingness measurement is used to rank and select class associative rules for building accurate classifiers. The experimental results show that iT CAR has excellent predictability by achieving the average classification accuracy above 90% and the average area under the curve above 0.8. Besides, iT CAR has significant reproducibility and interpretability through functional enrichment analysis and retrieval of meaningful cancer terms. These promising results suggest the proposed method has great potential in identifying candidate genes, which can be further investigated as biomarkers for cancer diseases.

Keyword: Microarray gene expression; Associative classification; Information gain; Biomarker discovery; Colorectal cancer; Breast cancer