Guyon, Isabelle; **Weston, Jason**; **Barnhill, Stephen**; **Vapnik, Vladimir**

Gene selection for cancer classification using support vector machines. (English)

Zbl 0998.68111
Mach. Learn. 46, No. 1-3, 389-422 (2002).

Summary: DNA micro-arrays now permit scientists to screen thousands of genes simultaneously and determine whether those genes are active, hyperactive or silent in normal or cancerous tissue. Because these new micro-array devices generate bewildering amounts of raw data, new analytical methods must be developed to sort out whether cancer tissues have distinctive signatures of gene expression over normal tissues or other types of cancer tissues.

We address the problem of selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA micro-arrays. Using available training examples from cancer and normal patients, we build a classifier suitable for genetic diagnosis, as well as drug discovery. Previous attempts to address this problem select genes with correlation techniques. We propose a new method of gene selection utilizing support vector machine methods based on recursive feature elimination. We demonstrate experimentally that the genes selected by our techniques yield better classification performance and are biologically relevant to cancer.

In contrast with the baseline method, our method eliminates gene redundancy automatically and yields better and more compact gene subsets. In patients with leukemia our method discovered 2 genes that yield zero leave-one-out error, while 64 genes are necessary for the baseline method to get the best result (one leave-one-out error). In the colon cancer database, using only 4 genes our method is 98% accurate, while the baseline method is only 86% accurate.

MSC:

68T05 Learning and adaptive systems in artificial intelligence
68U99 Computing methodologies and applications

Keywords:

DNA micro-arrays

Full Text: DOI