Feature Subset Selection by using Optimal Margin of Support Vector Classifier

Khin May Win, Nan Sai Moon Kham
Winn.km05@gmail.com, moonkhanucsy@gmail.com
University of Computer Studies, Yangon, Myanmar

Address correspondence to:
Khin May Win, SintGu Hall, Parami Road, Hlaing Township, Yangon,Myanmar

Abstract

Identification of cancer genes that might anticipate the clinical behaviors of different types of cancers is challenging due to the huge number of genes and small number of patients samples. The new method is being proposed based on supervised learning of classification like support vector machines (SVMs).A new solution is described by the introduction of the Minimized Margin (MM) in the subset criterion, which permits to get near the least generalization error rate. The performance of the new method was evaluated with real-world data experiment. It can give the better accuracy for classification.

Keywords: DNA microarray, feature selection, cancer classification, embedded method, support vector machines

1. Introduction

Now many researchers are investigating the class prediction methodology, especially for cancer classification. Recent advances in microarray technology allow scientists to measure the expression levels of thousands of features. New analytical methods are needed to be developed to identify the features of genes which have distinct signatures. Recently, Brown et al. applied a collection of supervised learning techniques to a set of microarray expression levels from yeast data . They showed that an algorithm known as a support vector machine (SVM) provides excellent classification for performance. SVMs are members of a larger class of algorithms, known as kernel methods .Kernel methods are non-linearly mapped to a higher-order feature space by replacing the dot product operation in the input space with a kernel function K(x, y) . Mercer's theorem[7] shows that every positive semi-definite kernel function corresponds to the dot product operation in some higher-dimensional feature space. In this work, we construct an explicitly heterogeneous kernel function by computing separate kernels for each data type and summing the results. The resulting kernel incorporates prior knowledge about the heterogeneity of the data by accounting for higher-order correlations among features of one data type but ignoring higher-order correlations across data types. This heterogeneous kernel leads to improved performance with respect to an SVM trained directly on the concatenated data.