ABSTRACT

Based on data from the Center for Data and Information Ministry of Health, in 2012 about 8.2 million cases of death caused by cancer. Recent developments show that DNA microarray technology is able to handle cancer detection problems early on, but the main disadvantage of microarray is the problem of curse of dimensionality.

Analysis of Variance (ANOVA) is one of the feature selection methods that can overcome the weakness of microarray. ANOVA can find an informative gene pair that can assist in the classification process performed by the Support Vector Machine (SVM). In SVM, the kernel trick when learning model is helpful in overcoming the feature space problem. The selection of the kernel affects the resulting accuracy. Through a series of processes such as correlation calculations, feature selection and classification using SVM, accuracy is obtained from the four datasets used. For leukemia and ovarian cancer datasets, the greatest accuracy is generated by the polynomial kernel at 100% and 97.54%. As for the largest lung cancer dataset accuracy obtained from linear kernel is 100% and for the dataset colon tumor the greatest accuracy is obtained from the RBF kernel of 85.15%. The kernel difference that produces the highest accuracy on each dataset depends heavily on the characteristics of the cancer dataset.

Keywords: *cancer detection, DNA microarray, dimension reduction, correlation, analysis of variance, support vector machine, kernel trick*