ABSTRACT

Cancer is a leading cause of death worldwide although a significant proportion of it can be cured if it is detected early. In recent decades, technology called microarray takes an important role in the diagnosis of cancer. In microarray, we can see the level of gene expression in specific cell samples to analyze thousands of genes simultaneously. By using data mining technique, microarray data classification can be performed to improve the accuracy of cancer diagnosis compared to traditional techniques. The characteristic of microarray data is small sample but it has huge dimension. Since that, there is a challenge for researcher to provide solutions for microarray data classification with high performance in both accuracy and running time. Several researches showed that dimension reduction could decrease the running time significantly while also potentially could overcome the noises in the data thus could improve the accuracy. This thesis proposed the usage of Principal Component Analysis (PCA) as a dimension reduction method along with Support Vector Method (SVM) optimized by kernel functions as a classifier for microarray data classification. The proposed scheme was applied on seven data sets using 5-fold cross validation and then evaluation and analysis conducted on term of both accuracy and running time. The result showed that the scheme can obtained 100% accuracy for Ovarian and Lung Cancer data when Linear and Cubic kernel functions are used. In term of running time, PCA greatly reduced the running time for every data sets.

Keyword: *kernel function, k-fold cross validation, microarray data classification, Principal Component Analysis, Support Vector Machine*.