Abstract

This paper presents an application of supervised machine learning approaches to the classification of the colon cancer gene expression data. Established feature selection techniques based on principal component analysis (PCA), independent component analysis (ICA), genetic algorithm (GA) and support vector machine (SVM) are, for the first time, applied to this data set to support learning and classification. Different classifiers are implemented to investigate the impact of combining feature selection and classification methods. Learning classifiers implemented include K-Nearest Neighbors (KNN) and support vector machine. Results of comparative studies are provided, demonstrating that effective feature selection is essential to the development of classifiers intended for use in high dimension domains. This research also shows that feature selection helps increase computational efficiency while improving classification accuracy.

References

- Canul Reich, J. L. Hall, O., Goldgof, D., and Esrichy, S. A. 2008 "Feature selection for microarray data by auc analysis," in IEEE SMC.
Science and Information Security.

**Index Terms**

Computer Science          Pattern Recognition

**Keywords**

Hold Out Pca SVM Knn Ica Features Classification Feature Selection Accuracy Colon Cancer