Chapter 2
Performance Analysis of Classifiers on Filter-Based Feature Selection Approaches on Microarray Data

Arunkumar Chinnaswamy
Amrita Vishwa Vidyapeetham University, Coimbatore, Campus, India

Ramakrishnan Srinivasan
Dr. Mahalingam College of Engineering and Technology, India

ABSTRACT

The process of Feature selection in machine learning involves the reduction in the number of features (genes) and similar activities that results in an acceptable level of classification accuracy. This paper discusses the filter based feature selection methods such as Information Gain and Correlation coefficient. After the process of feature selection is performed, the selected genes are subjected to five classification problems such as Naïve Bayes, Bagging, Random Forest, J48 and Decision Stump. The same experiment is performed on the raw data as well. Experimental results show that the filter based approaches reduce the number of gene expression levels effectively and thereby has a reduced feature subset that produces higher classification accuracy compared to the same experiment performed on the raw data. Also Correlation Based Feature Selection uses very fewer genes and produces higher accuracy compared to Information Gain based Feature Selection approach.

1. INTRODUCTION

Statistical analysis of differentially expressed genes helps to assign them to different classes. This process enhances the basic understanding of the biological processes in the system. The activity of thousands of genes could be investigated simultaneously using the concept of microarray gene expression technology. Gene expression profiles are used to predict the relative abundance and presence of mRNA in the genes. The results obtained using suitable discriminant analysis represent the state of the cell that serves as a tool for the diagnosis, prediction and treatment of diseases. The hybridization process is used for generating DNA microarray samples. This process can be done in two ways. In the first method, during the process of hybridization, the messenger RNA (mRNA) taken from sample tissues or from the blood stream is converted to cDNA if it uses spotted arrays. RNA profiles may be noisy and might be unequally sampled over time. The second method involves the use of Affymetrix chips that hybridizes the oligonucleotides on the surface of the chip array. The simultaneous measurement and monitoring of thousands of genes using a single experiment is made possible by using DNA microarray technology (Li Yeh Chuang, Kuo-Chuan Wu, & Cheng-Hong Yang, 2008). The production of proteins in a gene signifies the gene expression level that aids in identifying the membership of the different classes. The presence of a wide variety of gene expression problems helps in advancement in the field of clinical medicine using results produced by several microarray experiments. Microarray data finds its application in the areas of cancer classification, disease diagnosis, prediction and treatment and most importantly in the area of gene identification that would be used in drug development at later stages. This has been a recent advancement in the area of clinical research. Microarray cancer data is combined with statistical techniques to analyze the gene expression patterns to identify potential bio markers for the diagnosis and treatment of different types of cancer (Arunkumar C & Ramakrishnan S, 2014).

The most common challenge in bioinformatics is the process of selecting relevant and non redundant genes from the dataset. Complex biological problems can only be solved by predicting and classifying the genes in the most efficient way. Feature Selection and Classification are considered to be the two key tasks in microarray gene expression analysis. The process of classification purely depends on Feature selection as the fewer gene subsets will contribute to adequate increase in classifier accuracy. Identification of a subset of differentially expressed genes is the main goal of feature selection. This identified subset would exhibit strong correlation between different classes and this helps to distinguish features between these classes. Another important measure is to avoid overfitting and build faster and cost effective models. During the process of feature selection there might be situations wherein a weakly ranked gene might perform well and a critical gene might be left out during the
Related Content

Reinforcement Learning
Virtual Organization Support through Electronic Institutions and Normative Multi-Agent Systems
www.igi-global.com/chapter/virtual-organization-support-through-electronic/21166?camid=4v1a

Online Evolution of Adaptive Robot Behaviour
www.igi-global.com/article/online-evolution-of-adaptive-robot-behaviour/113296?camid=4v1a

Predictive Regulation in Affective and Adaptive Behaviour: An Allostatic-Cybernetics Perspective
www.igi-global.com/chapter/predictive-regulation-in-affective-and-adaptive-behaviour/176190?camid=4v1a