Classifier System Basing On Gene Expression Data For The MAIGES Environment

Introduction

Microarray as a newly emerged technology is playing a very crucial role in the modern molecular biology and medicine science researches. It not only opens a new way to express thousands and millions gene information, but also brings challenges to researchers to process the information efficiently. And one of these challenges is to classify the genes according to their corresponding gene expression data in a robust and trustworthy way.

This project, as my "Trabalho de Conclusao de Curso" (a.k.a.: T.C.C.), pretends to bring on a classifier system into MAIGES (Mathematical Analysis of Interacting Gene Expression System) by using supervised learning technique.

Objective

The objective of the project can be divided into the following steps:

•Study and research the related topics.

•Create a classifier system to MAIGES by using the study and research result from the step 1.

Concepts

•**Classification** In the context of this project, classification is a way to put the given data to the known groups according to their corresponding characters. In other words, if we have a set of K known classes {*c1, c2, c3, …., ck*}, named L, and a set of data {x1,x2, …., xm}, named X. Classification should be a function F, by which $F(X)=ci \in L$, $1 \leq i \leq k$.

• **Supervised Learning** According to the Wikipedia's definition, supervised learning is "a machine learning technique for learning a function from training data. …". It means that if a predicable system, as the classifier system of this TCC, adopts the supervised learning technique, the system must be able to understand and find out the basis for the predication (in our case, classification) from a set of known training data. Eventually, this basis information is then used in the future predication experiments.

• **Classifier and Classification Algorithm** A classifier is a function which could classify an object into one of the known classes on the basis of an observed measurement or features. A classification algorithm is a statistical technique used to conduct predictive analysis, and in sequence, to generate classifiers.

Activities

• **CART – Classification And Regression Tree**
CART is a tree-based algorithm adopted by this classifier system. It has the following 3 principle rules:

1. Splitting rule. At each node, choose the split that maximizes the decrease in impurity.
2. Split–stopping rule. Grow large tree, selectively prune the tree upward, getting a decreasing sequence of subtrees, then use cross–validation to identify the subtree having the lowest estimated misclassification rate.
3. Class assignment rule. For each terminal node, choose the class that minimizes the resubstitution estimate of the misclassification probability, given that a case falls into this node.

- **Measurement Functions** The big challenge is to find out the reliable measurement functions from a given set of training data which could be used in the future classification experiments for a gene expression data.

More information:
Email: proj.information@gmail.com