Classifier System Basing On Gene Expression Data For The MAIGES Environment

Tutor: Prof. Dr. Eduardo Jordão Neves

Author: Xieli Zhaofu

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 DNA sequence characters, named features, in a robust and large scale way.
- MAIGES – Mathematical Analysis of Interacting Gene Expression System developed by Gustavo Henrique Esteves.

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**
- **Supervised Learning**
- **Gene Expression Data**
- **Classifier and Classification Algorithm**

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 $F(X)=ci \in L$, $1 \leq i \leq k$.
- In the microarray research areas, classification has a very important role for classifying gene expression data. For example: cancer classes, tumor classes etc. And at the same time, the large and complex multivariate gene expression data generated by microarray experiments brings more challenges to classification.

Supervised Learning

- Training Data:

$$\begin{bmatrix} X11 & X12 & \dots & X1m \\ X21 & X22 & \dots & X2m \\ \vdots & & & \\ Xn1 & Xn2 & \dots & Xnm \end{bmatrix}$$

- Known Classes: {A,B}
- Task: Find out a Predictive Function F

  $F(X) \in \{A,B\}$

Classifier & Classification Algorithm

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

Activities

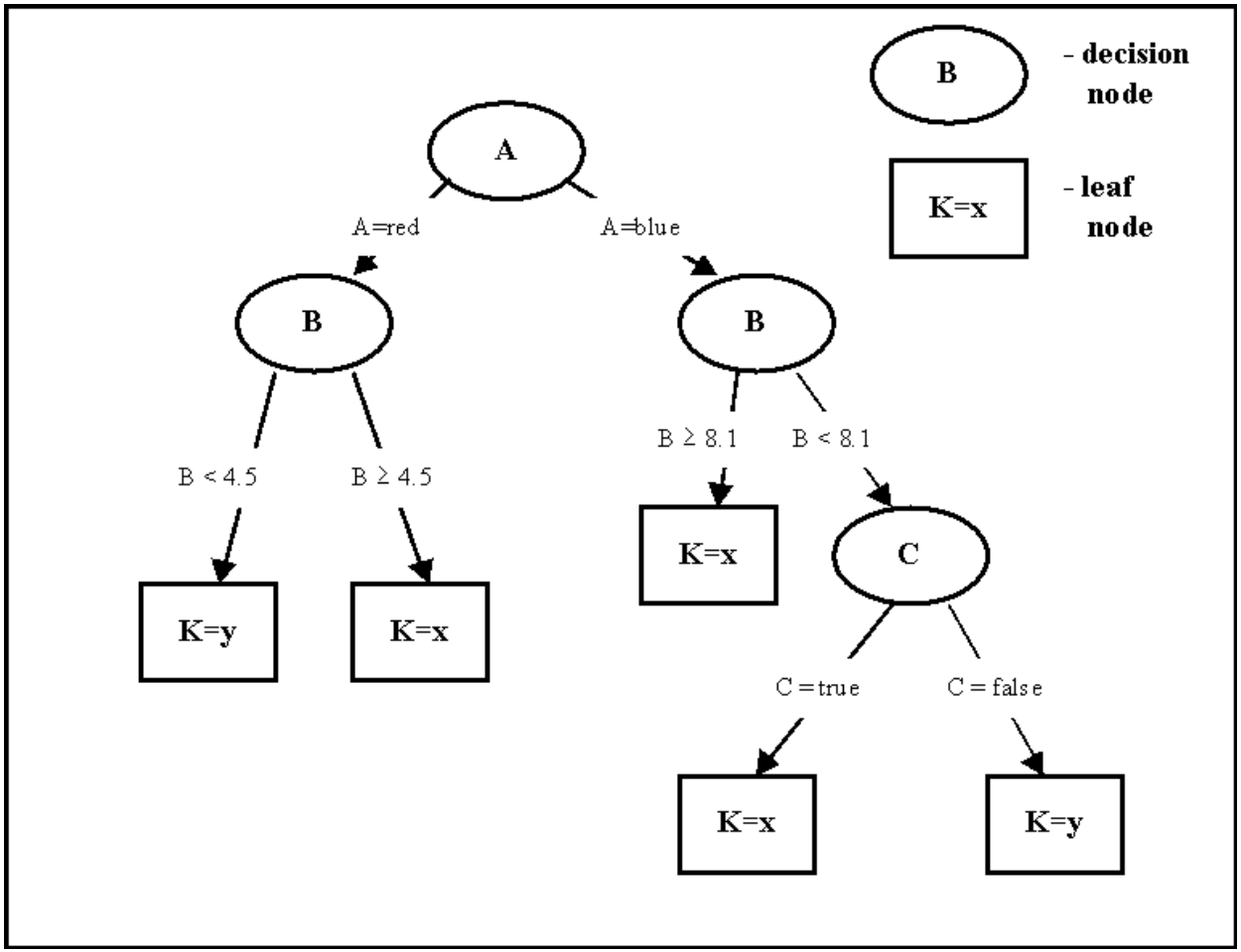
- CART – Classification And Regression Tree
- Measurement Functions

CART

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

CART - Illustration



Measurement Functions

- Entropy Impurity Equation: $Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$
- p_i is the proportion of S belonging to class *I*

Classifier System

- Still being under development

Conclusion

- Challenges :
  - A lot of study and research
  - New way to think
  - Frustrations

Information

- Source code: http://zhaofu.lee.googlepages.com/source.zip

- Email: proj.information@gmail.com