INFO411 Data Mining and Knowledge Discovery

Project 6

Instructions:

This tasks is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) your proposed data mining approach and methodology; (4) the strengths and weaknesses of your proposed approach; (5) the performance measures that can evaluate your data mining results; (6) the results a brief discussion and a conclusion.

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: BioData Mining

Background:

Biodata mining is one of the fundamental data mining problems. It focuses on discovering new biological knowledge through analyzing oceans of biological datasets. One important biological dataset is the DNA microarray according to chemical experiments (http://en.wikipedia.org/wiki/DNA_microarray).

Microarray time-series data provide temporal information of gene expressions that exhibit a strong dependence between adjacent time points. They have been widely used to study the cell cycle system, genetic regulatory interactions, development at the molecular level, and genes that act in response to a certain infectious disease. Because of the large number of genes that are profiled in time series data, the *clustering technique* plays an important role in providing a global overview of the underlying relationships of genes. In addition, the clustering can also determine the function of unknown genes or infer the regulatory networks. Generally, the computational issues for analyzing microarray time-series data have a hierarchy of four levels: (1) experimental design, (2) data analysis, (3) pattern recognition and (4) networks. In this task, we focus on the first three levels.

One major characteristic of microarray dataset is that there are missing values due to corruption of some expression measurements or other technical reasons such as hybridization failures. Estimating these missing values is an important step before further clustering analysis.

The attached microarray data is organized as a “522 x 18” matrix stored in the file “microarray”. Each row is a gene profile, and each column is the expression level (attribute) with an arbitrary unit. So, there are a total of 522 genes with 18 attributes. A “label” file is also provided, which is organized as a “522 x 10” matrix. This file contains the binary class labels (0 or 1) for each gene according to the biological pathways. The value 1 means that the gene belongs to this cluster, and the value 0 means that the gene does not belong to the cluster. To summarize, these genes involve in 10 biological clusters stored in the file “label”.

Requirements:

1. Present a general description of the dataset and present the general properties of the dataset as far as it is relevant for this task.
2. Analyse the dataset and estimate missing values.
3. Investigate into possible approaches (experimental design) to this task. Briefly explain those approaches.
4. Among the approaches that you identified, select the one that you think will work best. Explain why you think this approach will work best.
5. Provide performance measures of your approach.
6. Present your results (pattern recognition). Discuss the common properties of genes within the same cluster.
7. Summarize: What are the most interesting things that you discovered while working on this project?