Data Mining Project 2
Mining Microarray Gene Expression Data for Cancers

Released on Oct 29
Due 11:55pm, December 4

Background

In this project, the students are to write programs to perform several tasks concerning the mining and analysis of microarray gene expression datasets for cancers/diseases, and the students are to conduct analysis of one or more microarray gene expression datasets for cancers based on the mining tools and the mined results. In this project, we want to focus our attention on interesting (relatively small) gene sets.

For simplicity, we assume that each dataset has just two classes.

An example of gene expression datasets is the colon cancer dataset, which contains 62 samples collected from colon-cancer patients; among those samples, 40 are tumor biopsies labeled as "positive" and 22 are normal tissue biopsies labeled as "negative". Each tuple (row) in the data consists of the readings for the genes, and the class (which is the last column). Each gene is an attribute. The columns are separated by ",", which is a commonly used format in data mining. The dataset can be found on pilot under "Projects" called p1colon.txt. Two datasets were provided under "Projects" as part of project 1. You can use those datasets for project 2.

In your discussions and reports, refer to the genes as g1, ..., gN, corresponding to the left-to-right order in the data file.

Your program will need to be able to handle various datasets with formats similar to the ones given, so your program will need to go through the data once to determine the number of samples/rows and the number of attributes before doing other thing.

I will prefer to compile and test your program on gandalf/thor (UNIX machines), using an appropriate makefile provided by you.

In this project you can use various codes by others, as long as they were not originally designed to address the tasks discussed below. For example, you can use FP-growth code found from the web, e.g. at `http://www.borgelt.net/fpgrowth.html`. Also, weka codes may be useful. If you found programs that were designed to deal with the tasks of the projects, you should talk with the professor about whether you can use the codes. In any case, you should give credit to the sources of the codes in a clearly designated section your report. You could also use the codes you developed for project 1.

Tasks

The expectation is that your program will handle two or all three tasks. The marking will take this into consideration.

Task 1. Write code to perform correlation analysis. Produce the minimal gene sets that are strongly correlated with the disease classes. (Minimal – in the set containment sense.)

Task 2. Write code to mine minimal controlling gene sets. A set X of genes is said to be controlling if the matching data for the low-high bin-value combinations of the genes (discretized using entropy based method into two bins) of X have small total variations in the genes not in X. (Small should be with respect to the gene sets of the same size of X.)

The matching dataset of a pattern Z in a dataset D is defined to be $match(Z, D) = \{t \in D \mid t \text{ matches } Z\}$. If D is understood, we can simply write $match(Z)$ for $match(Z, D)$.

You may want to use some one-pass algorithm to compute variance (or other things if possible), using the formula:
$variance = \frac{1}{N}\sum_{i=1}^{N}(x_i - avg(x))^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - avg(x)^2$.

Task 3. Write code to perform neighborhood based interestingness analysis of emerging/frequent patterns. You can mine emerging patterns by comparing the frequent patterns mined using FP-growth on the discretized data for each of the classes. Neighborhood can be based on itemset-based neighbors, or dataset-based neighbors. You may want to define two distance functions: $dis_{items}(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$, and $dis_{data}(X, Y) = 1 - \frac{|match(X) \cap match(Y)|}{|match(X) \cup match(Y)|}$.

For example, it may be interesting to find emerging patterns X (with relatively high support) such that X is unusually far away from X's nearest neighbor emerging patterns Y (with relatively high support). Far away can be measured by either or both distance functions.

As another example, it may be interesting to compute neighbor based lift: Given X, find y such that $X \cup \{y\}$ gives the largest lift among all possible single item additions to X. Similarly, find $y \in X$ such that the deletion of y from X gives some extreme behavior.

You may combine the ideas in different tasks to make the results more interesting.

Your program should write interesting results into files. You should also produce files (e.g. split values, mapping between gene-bin pair and integers representing items) that are necessary to understand your mining results. The files should have informative names.

You have some freedom in variations in definition of "controlling" and "neighbor". Clearly, your definitions should make sense. You should clearly describe your definitions in your report, and explain why you think they make sense.

If necessary (because you do not know how to do better), your code can impose limitations. For example, you may consider limiting the size of gene sets to 3.

You want to make your program codes modularized. Each logical part should be coded as a function.

Your executable should be called genesetmine. It should use command line arguments with two arguments: dataset-name and task-number. Then the program should prompt the user to provide various thresholds and other necessary details.

Project 2 Submission Guidelines

Turn in all files needed to compile and execute your code (including a makefile) via the course web page. Put all files into a folder named as YourNameP2. That folder can have sub folders but not sub-sub folders. I should be able to compile your program just after typing "make" inside your directory/folder.

In the submitted folder, include a file called Report.pdf, to summarize your findings.

If there are any special instructions that I need to know about, be sure to include a file named README.TXT detailing them. Limit such instructions to a minimum.

You should not deviate from the guidelines and requirements discussed above. You should follow good programming/documentation practices.

Correctness, interestingness of mined results, efficiency, readability etc will be factors for marking, in addition to the number of tasks you complete. [A minimum of two is expected.] Moreover, to ensure that you make an early start on your project, you are required to send me an email on each Tuesday (except the last one) to describe how much you have done by that day – this will also be a factor in the marking.

In your code, you must include comments that show the major steps.

A demo may be scheduled for students when their projects cannot be successfully run by the professor.

Notes: The colon dataset was first made available by U. Alon, et al. ("Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays", PNAS, 96:6745-6750, 1999).

The Work Must Be Your Own: Your submission must be your own work, plus codes listed in the "credit" section of your report.