HST 951 Homework - Fuzzy Rough Rules

Due : 2 Days after Mid-Term Exam

1 Preliminaries

This homework is on rough sets and generating simple fuzzy membership functions. You will need to download (and possibly install)

the data set (in the homework bundle this file is in),

the R environment (http://www.r-project.org/),and

the ROSETTA system for the rough sets part (http://rosetta.lcb.uu.se/general/).

1.1 Objectives

The objectives are

1. Learning fuzzy membership functions and applying them in a discretization scheme.
2. Learning rough rules from the discretized data.
3. Gaining experience in using the R environment and the ROSETTA system.

2 Tasks

2.1 Task 1

Write an R program that

reads the data set on tissues U' into a data frame,

- discretizes the data according to the fuzzy discretization scheme outlined in the lecture creating another data frame, and
- writes the discretized data to a file in ROSETTA import format.

2.1.1 Hints

Following the ideas outlined in the lecture, we consider genes to be regulated up (u), neutral (n), or down (d). Let $G = \{g_j\}_j$ be a set of gene symbols, and let U be a set of tissue samples. Abusing notation slightly, let $g(x)$ denote the value of expression of the gene g in tissue sample x. For each gene g in question, we for now assume the existence of three corrsponding fuzzy membership functions $g_u, g_n, g_d : U \to [0,1]$. Let ramp : $\mathbf{R}^3 \to [0,1]$ be defined as

$$\text{ramp}(x, v, w) = \begin{cases} 0 & \text{if } x \leq v \\ 1 & \text{if } x > w \\ \frac{x-v}{w-v} & \text{otherwise} \end{cases}$$

Associate with each gene g a triplet $(g_{\min}, g_{\text{median}}, g_{\max}) \in \mathbf{R}^3$ such that $g_{\min} \leq g_{\text{median}} \leq g_{\max}$. Using these we define

$$\begin{aligned} g_u(x) &= \text{ramp}(g(x), g_{\text{median}}, g_{\max}) \\ g_d(x) &= 1 - \text{ramp}(g(x), g_{\min}, g_{\text{median}}) \\ g_n(x) &= 1 - \max(g_u(x)), g_d(x)). \end{aligned}$$

The definition of the membership function for a gene expression attribute value set is then parametrisized via the values $(g_{\min}, g_{\text{median}}, g_{\max})$. These we determine by letting $g_{\min} = \min\{g(x)|x \in U'\}$, $g_{\text{median}} = \text{median}\{g(x)|x \in U'\}$, and $g_{\max} = \max\{g(x)|x \in U'\}$. Given a tissue $x \in U'$, we determine its discretized value for gene g as $\arg\max_{l \in \{u,n,d\}}(g_l(x))$, encoding elements of u, n, d as the integers $2, 1, 0$, respectively. Assume that you can break ties arbitrarily.

The ROSETTA import format is

```
<attribute name> <attribute name> ... <attribute name>
<attribute type> <attribute type>     <attribute type>
<value> <value> ... <value>
```

where

`<attribute name>` is either a string containing no whitespace, or a (double) quoted string. Neither can contain (double) quotes (").

`<attribute type>` is either `String`, `Integer` (which I would guess that you will use), or `Float(`n`)`, where n is a decimal number describing the number of decimals after the integer part.

`<value>` is a value of a type given above.

Each element on each line is separated by whitespace (space or tab).

2.1.2 Deliverables

The file(s) containing the R program.

2.2 Task 2

The task is to

- Import the discretized data set produced in Task 1 into the ROSETTA system,
- compute "reducts" (reduce) using a standard greedy algorithm for the set cover problem attributed to D.S. Johnson (use the non-RSES one),
- generate rules, and
- export them to a file in rosetta export format.
- Comment briefly on the rules.

2.2.1 Hints

The ROSETTA system GUI (which only runs under Windows, while the kernel and command line system runs on several platforms) has two main components, structures and algorithms. The latter can be applied to the former to produce the latter. Both can be found in the project tree view that the GUI presents. Algorithms can also be found by right-clicking on a structure. The selected algorithm will then applied to the structure, and the result will usually appear as a sub structure of the structure it was computed from. For this task the algorithm types to apply (after importing the data by file->open) are of the type reducer (object related and modulo decision), rule generator, and exporter (plain format).

2.2.2 Deliverables

- The file containing the exported rules.
- A file containing your comments about the rules. Comment on how they reflect the relationship between gene expression and tissue type.