Analysis of Dataset 2

Preliminary Report

Henry Xiao

xiao@cs.queensu.ca

School of Computing

Queen's University

Queen's UNIVERSITY

Microarray Data Cleaning

Microarray dataset needs to be preprocessed before mining.

- Remove the “Gene Description” and change the “Gene Accession Number” to “ID”.
- Normalize all data values to the range [20, 16,000]. (i.e. the value less than 20 or over 16,000 was considered as unreliable by biologists.)
- Add a “Class” row to indicate the type of leukemia. (note for the test dataset all values for “Class” are “?”.)
- Transpose the dataset making each column to be an attribute, and each row to be a testing sample.

Queen's UNIVERSITY

Data Clean Dataset

After cleaning the microarray data, our dataset is ready for mining.

- 7070 genes(attributes) for each sample.
- Class Variable: $Class \in ALL,\ AML$.
- “Class” is treated as nominal here.
- “ID” should be further removed.

Queen's UNIVERSITY

Attribute Selection

It is generally impossible to do mining with 7070 attributes. *Weka* gives out of memory error when running attribute selection algorithms. Attribute selection has to be done manually!

Queen's UNIVERSITY

Attribute Selection

It is generally impossible to do mining with 7070 attributes. *Weka* gives out of memory error when running attribute selection algorithms. Attribute selection has to be done manually!

- Attribute's Discrimination Ability.

Queen's UNIVERSITY

Attribute Selection

It is generally impossible to do mining with 7070 attributes. *Weka* gives out of memory error when running attribute selection algorithms. Attribute selection has to be done manually!

- Attribute's Discrimination Ability.
- The Signal-to-Noise (*S2N*) ratio and T-value:

$$S2N = \frac{\mu_1 - \mu_2}{\delta_1 + \delta_2}$$

$$T - value = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\delta_1^2}{N_1} + \frac{\delta_2^2}{N_2}}}$$

where μ is the average, δ is the standard derivation, and N is the number of samples.

Queen's UNIVERSITY

Attribute Selection

It is generally impossible to do mining with 7070 attributes. *Weka* gives out of memory error when running attribute selection algorithms. Attribute selection has to be done manually!

- Attribute's Discrimination Ability.
- The Signal-to-Noise (*S2N*) ratio and T-value:

$$S2N = \frac{\mu_1 - \mu_2}{\delta_1 + \delta_2}$$

$$T - value = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\delta_1^2}{N_1} + \frac{\delta_2^2}{N_2}}}$$

where μ is the average, δ is the standard derivation, and N is the number of samples.

- Join the two measures to hope a better set.

Queen's UNIVERSITY

Selected Attributes

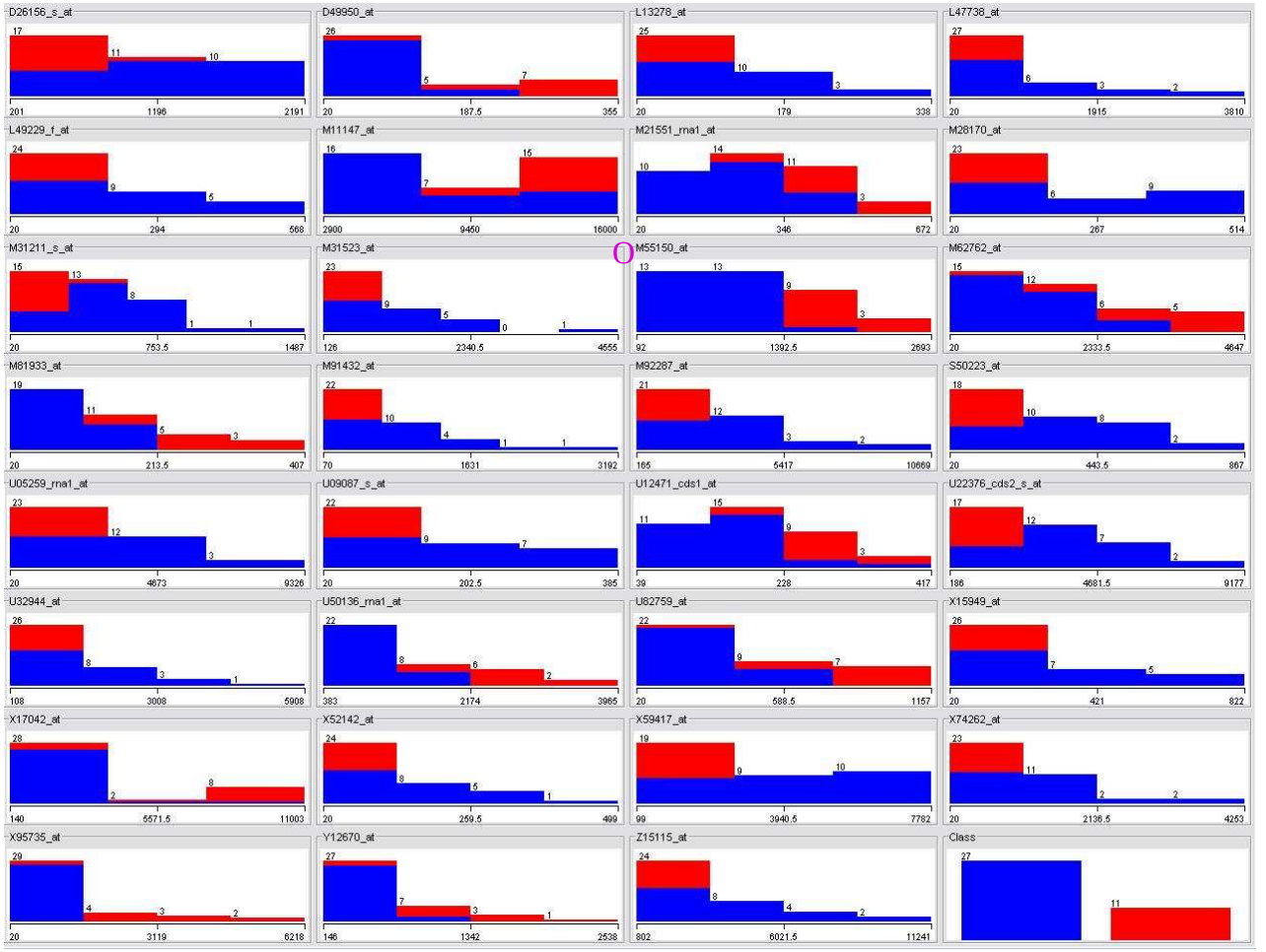
Three attribute sets have been selected by the following approaches.

- Rank the genes by the absolute value of *S2N* ratio and get top 50 genes.
- Rank the genes by the absolute value of *T-value* and get top 50 genes.
- Combine the previous two sets with the *common* genes.

Detail information of the selected genes can be found on my page: *http://www.cs.queensu.ca/home/xiao/dm.html*.

Queen's UNIVERSITY

Attribute Visualization



Preliminary Experiment

Naïve Bayes and Bayes Net have been applied.

- Load training dataset and provide testing dataset.
- Testing dataset attributes are mirrored from the training dataset's.
- Save the "error distribution" to see the classification result of the testing dataset.
- Compare the classification results (i.e. *ALL* or *AML* for each testing sample with different gene sets and different mining techniques.

Queen's UNIVERSITY

Some Preliminary Result

The results of applying two techniques with different gene sets are showed in below table.

Class	*NB(S2N)*	*BN(S2N)*	*NB(T)*	*BN(T)*	*NB(C)*	*BN(C)*
ALL	20	23	27	25	24	24
AML	14	11	7	9	10	10

- NB: Naïve Bayes; BN: Bayes Net.
- T: *T-value*; C: Common gene set.
- Disagree sample ID's: $\{51, 54, 60, 61, 62, 63, 64, 67\}$.

Queen's UNIVERSITY

Some Preliminary Results

Some knowledge from the experiments.

- The most different two classifications comes from applying Naïve Bayes between *S2N* set and *T-value* set.
- The most likely two classifications come from applying Naïve Bayes and Bayes Net on Common set.
- Common set may have a better discrimination power.
- Sample 61 is the most confused case with 3 *ALL* votes and 3 *AML* votes.

Queen's UNIVERSITY

Discussion

Some related questions and further mining plan.

- *S2N* and *T-value* measures, or others?
- Validate the results.
- Improvement from Biology point of view.
- Possibility of other types of leukemia.

Queen's UNIVERSITY

Ending

Questions regarding Analysis results?

Information Site: http://www.cs.queensu.ca/home/xiao/dm.html

E-mail: xiao@cs.queens.ca

Thank you

Queen's UNIVERSITY