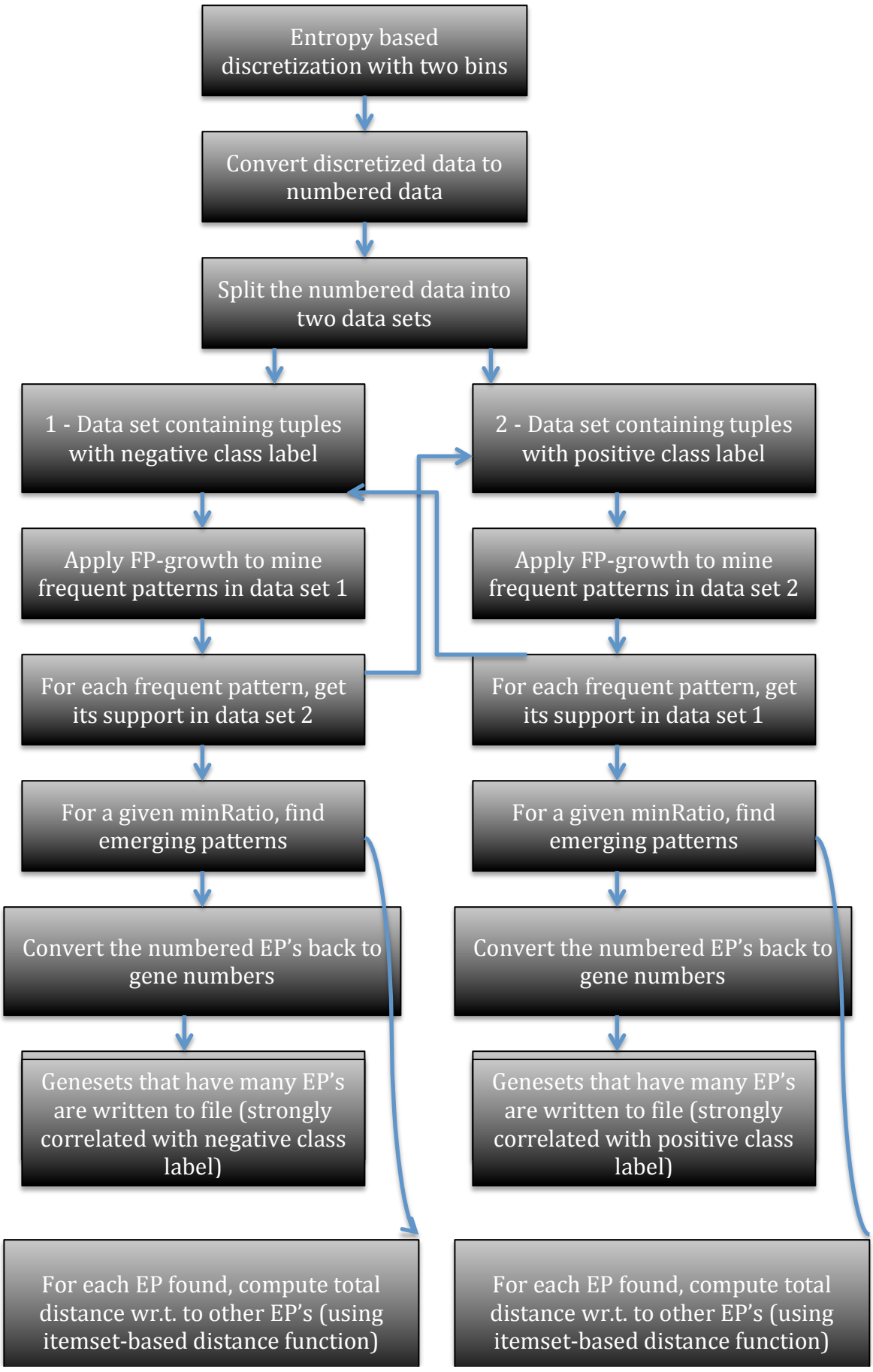
Cs 7720 –Project 2 : Mining Microarray Gene Expression Data for Cancers
Dattatreya Adusumilli

PROCEDURE



APPROACH

- First file scan and entropy based discretization:

The given dataset is scanned once to get the number of rows (tuples) and columns (attributes, class label). Entropy based discretization is performed for the given data with two bins. The discretized data is converted to numbered data, wherein the two states of a gene (low/high) are represented by numbers (2i-1) and 2i; 'i' being the gene number.

- Finding frequent patterns using FP-growth algorithm:

The numbered data is split into two data sets, one with tuples containing positive class label and, the other dataset with tuples containing negative class label. FP-growth algorithm is applied to each of these two datasets - to find the frequent patterns, given a minimum support threshold.

Once the two sets of frequent patterns are attained via FP-growth algorithm, for each frequent pattern found, its support in the other numbered dataset is obtained. For a given minimum support ratio, the emerging patterns corresponding to each of the two class labels are obtained using 'big support ratio'.

Big support ratio: supp2 (x)/supp1 (x) >= minRatio.

For example,

- With *minRatio* being 6, a frequent pattern {2, 3, 19} with supp2 (x) = 20 and supp1 (x) = 2 qualifies as an emerging pattern as it's big support ratio is greater than 6.
- Similarly, a frequent pattern {2,1 31, 19} with supp2 (x) = 20 and supp1 (x) = 19 does not qualify as an emerging pattern
- It is worth noting that if the support of a pattern is ~ 100% in one class and zero in the other class then, it is considered as a *jumping emerging pattern*.

*** The above steps are common for task 1 and 3*

Task 1: Perform correlation analysis. Produce the minimal gene sets that are strongly correlated with the disease classes.

With steps mentioned in the above section, as emerging patterns have been obtained, the next step is to derive the minimal gene sets that are strongly correlated with the class. To accomplish this, the emerging patterns are converted to corresponding gene numbers. Subsequently, FP-growth algorithm is applied to find the gene sets which are frequent, given a minimum support threshold. In other

words, FP-growth algorithm is used to find the minimal gene sets which have many emerging patterns. The minimal gene sets which are strongly correlated with each of the two class labels are written to file(s).

Task 3: Perform neighborhood based interestingness analysis of emerging patterns. Neighborhood can be based on itemset-based neighbors. 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).

Emerging patterns pertaining to each of the two class labels have already been obtained. Distance between each pair of emerging patterns is computed using itemset-based neighbors. Subsequently, total distance is computed for each emerging pattern. The emerging patterns with their corresponding total distances are sorted based on the distance values.

The emerging patterns that have the highest distance value are considered as interesting, as these patterns are far away from other patterns, based on the item-set - distance measure.

For example, consider four emerging patterns:

{1,2}, {2,3,4}, {2,3,4,6} and {9,10}

The total distance for each pattern would be,

- {1,2} = (1 – (1/4)) + (1 – (1/5)) + (1 – 0) = 2.55
- {2,3,4} = 2
- {2,3,4,6} = 2.05
- {9,10} = 3

As we can observe from the above data, based on item-set distance measure, the patterns {1,2}, {2,3,4} and {2,3,4,6} are close to 2.0 whereas the pattern {9,10} is far away from its neighbors and, hence it is considered as interesting.

LIMITATIONS

Based on test runs with various combinations of minimum support threshold for FP-growth and minRatio, I have observed that FP-growth takes relatively long time (10+ minutes) to produce frequent patterns - when more than 50 genes were given as input, with a minimum support threshold of 0.8. Hence, for the purpose of this project, the input to FP-growth algorithm is limited to 50 genes and, a maximum item-set size of 4 – for frequent patterns.

FINDINGS

When the top 35 genes based on information gain order are considered for mining frequent patterns using FP-growth with,

- Minimum support of 0.6 – to find frequent patterns
- Minimum ratio of 20 – to find emerging patterns
- Minimum threshold of 0.004 for obtaining gene sets

The following gene sets - that are strongly correlated with the corresponding class label were obtained.

Genesets strongly correlated with negative class label	**Genesets strongly correlated with positive class label**
{1227, 822}	{652, 249}
{964, 765}	{1325, 493}
{897, 513}	{642, 66}
{249, 513, 822}	{399, 249, 493}
{415, 513, 822}	{1153, 1325, 493}
{1060, 1227, 822}	{1325, 780}

Note: only a subset of the gene-sets obtained in the result has been presented in the table.

The emerging patterns that are interesting based on itemset based distance measure are,

Emerging patterns – negative class label	**Total distance of pattern**	**Emerging patterns – positive class label**	**Total distance of pattern**
{652+, 1060+}	1424.65	{1325-, 897+, 652-, 377+}	1222.03
{1325+, 415-}	1388.53	{780-, 1325-, 897+, 377+}	1221.76
{964+, 66-}	1378.08	{964-, 493+, 1325-, 467-}	1195.37
{467+, 513+}	1375.68	{964-, 493+, 1325-, 43-}	1191.54
{964+, 493-}	1370.35	{780-, 1325-, 897+, 1047-}	1214.17

Note: 66- represents low state of gene 66; and, 513+ represents high state of gene 513

FILES

A brief description of the results written to files:

Entropy.data
Contains the discretized data using entropy-based discretization. Two bins are used for this project; a gene value's low state is indicated by "a" and it's high state is indicated by "b".

Gene-maxgain-splitvalue.data
The maximum information obtained for each gene while performing entropy-based discretization and, the best split value are written to this file.

Example: g2, 0.064, 5396.91. Here g2 indicates gene 2, followed by the maximum information gain and, the best split value.

Top-genes-infogain.data
Each line in this file contains a gene number and its corresponding information gain. The genes are sorted based on information gain order.

Top-genes-numbered.data
Each line in this file represents a tuple containing numbered data that corresponds to the top 50 genes and, is followed by the class label. The fifty genes limit is imposed as FP-growth takes relatively long time to generate frequent patterns when more than 50 genes are given as input (as per my test runs).

Numberd.data
Similar to the *entropy.data* file, this file represented the data in numbered format.

For instance,
In line one – 2,4,5,7 … means gene1 +, gene2 +, gene3-, g4- and so on.

Negativetuplesnumberd.data and ***Positivetuplesnumbered.data***
As mentioned in the approach section, the numbered data is divided into two data sets one containing tuples with the negative class label and the other with positive class label. These two files represent the numbered data files that are given as input to FP-growth algorithm to mine frequent patterns.

Frequent_patterns_neg.data
The frequent patterns obtained from the data set that contain the negative class label are saved to this file. Each line represents a pattern and it's corresponding support in the data set.

Example: 1304,2120,1250:24
{1304,2120,1250} Indicates the pattern and 24 indicates its support in the data set with tuples containing negative class label.

Frequent_patterns_pos.data
The frequent patterns obtained from the data set that contain the positive class label are saved to this file. Each line represents a pattern and it's corresponding support in the data set.
Example: 1162,2585,2305:14
{1162,2585,2305} Indicates the pattern and 14 represents its support in the data set with 'positive' class label.

Emerging_patterns_neg_pos.data
This file contains one emerging pattern per line followed by its support in the data set with 'negative' class label and then, its support in the data set with 'positive' class label.

Emerging_patterns_pos_neg.data
This file contains one emerging pattern per line followed by its support in the data set with 'positive class label and then, its support in the data set with 'negative' class label.

Sorted_corr_genesets_neg.data and ***Sorted_corr_genesets_pos.data***
The correlated gene sets that have *many* (taken as input by user) emerging patterns are written to these files.

Sorted_ep_avg_distance_neg.data and ***Sorted_ep_avg_distance_pos.data***
Emerging patterns along with their total distance are written to these files one emerging pattern per line. Emerging patterns with a high total distance are perceived as interesting as they are far away from other neighbor patterns as per item-set based distance measure.

CREDITS

- FP-growth algorithm implementation for java has been obtained from 'SPMF' http://www.philippe-fournier-viger.com/spmf/index.php?link=documentation.php#growth