Homework 2: Mining Association Rules from Gene Expression Data

CSE 601: Data Mining and Bioinformatics

AVIJEET MISHRA
(AVIJEETM@BUFFALO.EDU)
UB#:50169242

PRITHVI GOLLU INDRAKUMAR
(PGOLLUIN@BUFFALO.EDU)
UB#:50169089

DEPT OF COMPUTER SCIENCE,
UNIVERSITY AT BUFFALO

Contents

Contents

Implementation details

We have implemented the Apriori algorithm in C# using Microsoft Visual Studio 2015. Support and confidence are made dynamic in the UI so that we could test using different values. Dictionary is used as the data structure for storing all the manipulated data. Dictionary is similar to a Hash map with unique key and value. This data structure has been used because we would be accessing the data structure very frequently and it has constant time complexity for the search of a record so it would provide the best performance for this scenario. We have uploaded the data of gene_expression.txt to the SQL server and got the data using SQL connection as it was less time consuming comparing to uploading data from excel. G1, G2…G100, G101 is added to all the items corresponding to the 100 samples.

1st Part

1st scan is performed on the data to calculate C1 which is basically the count of items with their support (frequency of item set in the data). So we loop on the data to generate the itemset with length 1 and their corresponding frequency in the data. If the same key already exist in the hash table then we continue with the loop else added it to the hash table.

```
for (int k = 0; k < 100; k++)
                {
                    //

                    string[] words = data[k].Split(',');

                    int c = 1;
                    for (int i = 1; i < words.Length; i++)
                    {
                        int val1 = 0;
                        if (ht.TryGetValue(words[i], out val1))
                        {
                            continue;
                        }
                        else
                        {
                            ht.Add(words[i], 1);
                        }
```

```
                    for (int p = (k + 1); p < 100; p++)
                    {
                        if (data[p].Contains(words[i]))
                        {
                            int val = 0;
                            if (ht.TryGetValue(words[i], out val))
                            {
                                val++;
                                ht[words[i]] = val;
                            }
                        }
                    }
                }//inner for i

            }
```

Now according to the threshold support provided in the UI, item set is reduced if the support is less than the threshold support.

```
foreach (KeyValuePair<string, int> kp in ht)
            {
                if (kp.Value >= sup)
                {
                    l30.Add(kp.Key, kp.Value);
                    l1.Add(kp.Key, kp.Value);
                    count30++;
                }


             }
```

Now we start a recursive loop which would create the new itemset on the basis of itemset generated in previous step. The logic for combining the items is done by looping on the all itemset from first index and then again looping it from the next index and then add “,” between them. New item is added only to the database if it is already not present in the database. Then again there combination is scanned in the original gene_expression data to get the support. Now this process is recursively called to get the all results corresponding to the variable lengths of itemset.

2nd Part

Initially Apriori algorithm is called using the same code explained in part 1. The difference is that as soon as itemset at each scan are generated, all the possible combination of the items are calculated to generate the rules by looping on each itemset. This would be calculated recursively to generate all the possible rules for all the possible length of itemset. For Ex: itemset {G59_UP,G72_UP,G82_Down} would generate following rules :

G72_UP,G82_Down->G59_UP

G59_UP,G82_Down->G72_UP

G59_UP,G72_UP->G82_Down

G82_Down->G59_UP,G72_UP

G72_UP->G59_UP,G82_Down

G59_UP->G72_UP,G82_Down

Results for part1:
Support is set to be 30%

Number of length-1 frequent itemset: 5338

Number of length-2 frequent itemset: 5287

Number of length-3 frequent itemset: 1518

Number of length-4 frequent itemset: 438

Number of length-5 frequent itemset: 88

Number of length-6 frequent itemset: 11

Number of length-7 frequent itemset: 1

Number of length-8 frequent itemset: 0

Total: 12681

Support is set to be 40%

Number of length-1 frequent itemset: 167

Number of length-2 frequent itemset: 753

Number of length-3 frequent itemset: 156

Number of length-4 frequent itemset: 8

Number of length-5 frequent itemset: 1

Total: 1085

Support is set to be 50%

Number of length-1 frequent itemset: 109

Number of length-2 frequent itemset: 63

Number of length-3 frequent itemset: 2

Number of length-4 frequent itemset: 0

Total: 1074

Support is set to be 60%

Number of length-1 frequent itemset: 34

Number of length-2 frequent itemset: 2

Number of length-3 frequent itemset: 0

Total: 36

Support is set to be 70%

Number of length-1 frequent itemset: 7

Number of length-2 frequent itemset: 0

Total: 7

Results for part 2:

Support is set to 50% and confidence to 60%

For template 1:

1. RULE HAS ANY OF G6_UP: **10**

G13_Down-G6_UP
G6_UP-G13_Down
G28_Down-G6_UP
G6_UP-G28_Down
G59_UP-G6_UP
G6_UP-G59_UP
G38_Down-G6_UP
G6_UP-G38_Down
G32_Down-G6_UP
G6_UP-G32_Down

2. RULE HAS 1 OF G1_UP: **14**

G59_UP-G1_UP
G1_UP-G59_UP
G72_UP-G1_UP
G1_UP-G72_UP
G38_Down-G1_UP
G1_UP-G38_Down

G54_UP-G1_UP
G1_UP-G54_UP
G70_Down-G1_UP
G1_UP-G70_Down
G10_Down-G1_UP
G1_UP-G10_Down
G67_UP-G1_UP
G1_UP-G67_UP

3. RULE HAS 1 OF (G1_UP, G10_DOWN): **26**

G59_UP-G1_UP
G1_UP-G59_UP
G72_UP-G1_UP
G1_UP-G72_UP
G38_Down-G1_UP
G1_UP-G38_Down
G54_UP-G1_UP
G1_UP-G54_UP
G70_Down-G1_UP
G1_UP-G70_Down
G67_UP-G1_UP
G1_UP-G67_UP
G10_Down-G28_Down
G28_Down-G10_Down
G10_Down-G59_UP
G59_UP-G10_Down
G10_Down-G38_Down
G38_Down-G10_Down
G10_Down-G47_UP
G47_UP-G10_Down
G10_Down-G88_Down
G88_Down-G10_Down
G10_Down-G70_Down
G70_Down-G10_Down
G10_Down-G94_UP
G94_UP-G10_Down

4. BODY HAS ANY OF G6_UP: **5**

G6_UP-G13_Down
G6_UP-G28_Down
G6_UP-G59_UP
G6_UP-G38_Down

G6_UP-G32_Down

5. BODY HAS NONE OF G72_UP: **124**

G59_UP-G1_UP
G1_UP-G59_UP
G1_UP-G72_UP
G38_Down-G1_UP
G1_UP-G38_Down
G54_UP-G1_UP
G1_UP-G54_UP
G70_Down-G1_UP
G1_UP-G70_Down
G10_Down-G1_UP
G1_UP-G10_Down
G67_UP-G1_UP
G1_UP-G67_UP
G13_Down-G6_UP
G6_UP-G13_Down
G28_Down-G6_UP
G6_UP-G28_Down
G59_UP-G6_UP
G6_UP-G59_UP
G38_Down-G6_UP
G6_UP-G38_Down
G32_Down-G6_UP
G6_UP-G32_Down
G28_Down-G13_Down
G13_Down-G28_Down
G59_UP-G13_Down
G13_Down-G59_UP
G13_Down-G72_UP
G82_Down-G13_Down
G13_Down-G82_Down
G54_UP-G24_Down
G24_Down-G54_UP
G88_Down-G24_Down
G24_Down-G88_Down
G52_Down-G28_Down
G28_Down-G52_Down
G59_UP-G28_Down
G28_Down-G59_UP
G2_Down-G28_Down
G28_Down-G2_Down

G38_Down-G28_Down
G28_Down-G38_Down
G47_UP-G28_Down
G28_Down-G47_UP
G87_UP-G28_Down
G28_Down-G87_UP
G88_Down-G28_Down
G28_Down-G88_Down
G32_Down-G28_Down
G28_Down-G32_Down
G10_Down-G28_Down
G28_Down-G10_Down
G41_Down-G28_Down
G28_Down-G41_Down
G38_Down-G52_Down
G52_Down-G38_Down
G59_UP-G72_UP
G82_Down-G59_UP
G59_UP-G82_Down
G96_Down-G59_UP
G59_UP-G96_Down
G38_Down-G59_UP
G59_UP-G38_Down
G87_UP-G59_UP
G59_UP-G87_UP
G88_Down-G59_UP
G59_UP-G88_Down
G32_Down-G59_UP
G59_UP-G32_Down
G10_Down-G59_UP
G59_UP-G10_Down
G82_Down-G72_UP
G96_Down-G72_UP
G97_Down-G72_UP
G38_Down-G72_UP
G32_Down-G72_UP
G96_Down-G82_Down
G82_Down-G96_Down
G97_Down-G82_Down
G82_Down-G97_Down
G38_Down-G2_Down
G2_Down-G38_Down
G47_UP-G38_Down
G38_Down-G47_UP

G65_Down-G38_Down
G38_Down-G65_Down
G88_Down-G38_Down
G38_Down-G88_Down
G32_Down-G38_Down
G38_Down-G32_Down
G70_Down-G38_Down
G38_Down-G70_Down
G91_UP-G38_Down
G38_Down-G91_UP
G94_UP-G38_Down
G38_Down-G94_UP
G10_Down-G38_Down
G38_Down-G10_Down
G41_Down-G38_Down
G38_Down-G41_Down
G67_UP-G38_Down
G38_Down-G67_UP
G10_Down-G47_UP
G47_UP-G10_Down
G88_Down-G54_UP
G54_UP-G88_Down
G88_Down-G87_UP
G87_UP-G88_Down
G8_UP-G88_Down
G88_Down-G8_UP
G10_Down-G88_Down
G88_Down-G10_Down
G41_Down-G88_Down
G88_Down-G41_Down
G10_Down-G70_Down
G70_Down-G10_Down
G10_Down-G94_UP
G94_UP-G10_Down
G59_UP,G82_Down-G72_UP
G82_Down-G59_UP,G72_UP
G59_UP-G72_UP,G82_Down
G59_UP,G96_Down-G72_UP
G96_Down-G59_UP,G72_UP
G59_UP-G72_UP,G96_Down

6. BODY HAS 1 OF (G1_UP, G10_DOWN): **15**

G1_UP-G59_UP
G1_UP-G72_UP
G1_UP-G38_Down
G1_UP-G54_UP
G1_UP-G70_Down
G10_Down-G1_UP
G1_UP-G10_Down
G1_UP-G67_UP
G10_Down-G28_Down
G10_Down-G59_UP
G10_Down-G38_Down
G10_Down-G47_UP
G10_Down-G88_Down
G10_Down-G70_Down
G10_Down-G94_UP

7. HEAD HAS ANY OF G6_UP: **5**

G13_Down-G6_UP
G28_Down-G6_UP
G59_UP-G6_UP
G38_Down-G6_UP
G32_Down-G6_UP

8. HEAD HAS NONE OF (G1_UP, G6_UP): **126**

G1_UP-G59_UP
G1_UP-G72_UP
G1_UP-G38_Down
G1_UP-G54_UP
G1_UP-G70_Down
G1_UP-G10_Down
G1_UP-G67_UP
G6_UP-G13_Down
G6_UP-G28_Down
G6_UP-G59_UP
G6_UP-G38_Down
G6_UP-G32_Down
G28_Down-G13_Down
G13_Down-G28_Down
G59_UP-G13_Down

G13_Down-G59_UP
G72_UP-G13_Down
G13_Down-G72_UP
G82_Down-G13_Down
G13_Down-G82_Down
G54_UP-G24_Down
G24_Down-G54_UP
G88_Down-G24_Down
G24_Down-G88_Down
G52_Down-G28_Down
G28_Down-G52_Down
G59_UP-G28_Down
G28_Down-G59_UP
G2_Down-G28_Down
G28_Down-G2_Down
G38_Down-G28_Down
G28_Down-G38_Down
G47_UP-G28_Down
G28_Down-G47_UP
G87_UP-G28_Down
G28_Down-G87_UP
G88_Down-G28_Down
G28_Down-G88_Down
G32_Down-G28_Down
G28_Down-G32_Down
G10_Down-G28_Down
G28_Down-G10_Down
G41_Down-G28_Down
G28_Down-G41_Down
G38_Down-G52_Down
G52_Down-G38_Down
G72_UP-G59_UP
G59_UP-G72_UP
G82_Down-G59_UP
G59_UP-G82_Down
G96_Down-G59_UP
G59_UP-G96_Down
G38_Down-G59_UP
G59_UP-G38_Down
G87_UP-G59_UP
G59_UP-G87_UP
G88_Down-G59_UP
G59_UP-G88_Down
G32_Down-G59_UP

G59_UP-G32_Down
G10_Down-G59_UP
G59_UP-G10_Down
G82_Down-G72_UP
G72_UP-G82_Down
G96_Down-G72_UP
G72_UP-G96_Down
G97_Down-G72_UP
G72_UP-G97_Down
G38_Down-G72_UP
G72_UP-G38_Down
G32_Down-G72_UP
G72_UP-G32_Down
G96_Down-G82_Down
G82_Down-G96_Down
G97_Down-G82_Down
G82_Down-G97_Down
G38_Down-G2_Down
G2_Down-G38_Down
G47_UP-G38_Down
G38_Down-G47_UP
G65_Down-G38_Down
G38_Down-G65_Down
G88_Down-G38_Down
G38_Down-G88_Down
G32_Down-G38_Down
G38_Down-G32_Down
G70_Down-G38_Down
G38_Down-G70_Down
G91_UP-G38_Down
G38_Down-G91_UP
G94_UP-G38_Down
G38_Down-G94_UP
G10_Down-G38_Down
G38_Down-G10_Down
G41_Down-G38_Down
G38_Down-G41_Down
G67_UP-G38_Down
G38_Down-G67_UP
G10_Down-G47_UP
G47_UP-G10_Down
G88_Down-G54_UP
G54_UP-G88_Down
G88_Down-G87_UP

G87_UP-G88_Down
G8_UP-G88_Down
G88_Down-G8_UP
G10_Down-G88_Down
G88_Down-G10_Down
G41_Down-G88_Down
G88_Down-G41_Down
G10_Down-G70_Down
G70_Down-G10_Down
G10_Down-G94_UP
G94_UP-G10_Down
G72_UP,G82_Down-G59_UP
G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G82_Down-G59_UP,G72_UP
G72_UP-G59_UP,G82_Down
G59_UP-G72_UP,G82_Down
G72_UP,G96_Down-G59_UP
G59_UP,G96_Down-G72_UP
G59_UP,G72_UP-G96_Down
G96_Down-G59_UP,G72_UP
G72_UP-G59_UP,G96_Down
G59_UP-G72_UP,G96_Down

9. HEAD HAS 1 OF (G6_UP, G8_UP): **6**

G13_Down-G6_UP
G28_Down-G6_UP
G59_UP-G6_UP
G38_Down-G6_UP
G32_Down-G6_UP
G88_Down-G8_UP

10. RULE HAS 1 OF (G1_UP, G6_UP, G72_UP): **48**

G59_UP-G1_UP
G1_UP-G59_UP
G38_Down-G1_UP
G1_UP-G38_Down
G54_UP-G1_UP
G1_UP-G54_UP
G70_Down-G1_UP
G1_UP-G70_Down
G10_Down-G1_UP

G1_UP-G10_Down
G67_UP-G1_UP
G1_UP-G67_UP
G13_Down-G6_UP
G6_UP-G13_Down
G28_Down-G6_UP
G6_UP-G28_Down
G59_UP-G6_UP
G6_UP-G59_UP
G38_Down-G6_UP
G6_UP-G38_Down
G32_Down-G6_UP
G6_UP-G32_Down
G72_UP-G13_Down
G13_Down-G72_UP
G72_UP-G59_UP
G59_UP-G72_UP
G82_Down-G72_UP
G72_UP-G82_Down
G96_Down-G72_UP
G72_UP-G96_Down
G97_Down-G72_UP
G72_UP-G97_Down
G38_Down-G72_UP
G72_UP-G38_Down
G32_Down-G72_UP
G72_UP-G32_Down
G72_UP,G82_Down-G59_UP
G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G82_Down-G59_UP,G72_UP
G72_UP-G59_UP,G82_Down
G59_UP-G72_UP,G82_Down
G72_UP,G96_Down-G59_UP
G59_UP,G96_Down-G72_UP
G59_UP,G72_UP-G96_Down
G96_Down-G59_UP,G72_UP
G72_UP-G59_UP,G96_Down
G59_UP-G72_UP,G96_Down

11. RULE HAS ANY OF (G1_UP, G6_UP, G72_UP): **50**

G59_UP-G1_UP
G1_UP-G59_UP

G72_UP-G1_UP
G1_UP-G72_UP
G38_Down-G1_UP
G1_UP-G38_Down
G54_UP-G1_UP
G1_UP-G54_UP
G70_Down-G1_UP
G1_UP-G70_Down
G10_Down-G1_UP
G1_UP-G10_Down
G67_UP-G1_UP
G1_UP-G67_UP
G13_Down-G6_UP
G6_UP-G13_Down
G28_Down-G6_UP
G6_UP-G28_Down
G59_UP-G6_UP
G6_UP-G59_UP
G38_Down-G6_UP
G6_UP-G38_Down
G32_Down-G6_UP
G6_UP-G32_Down
G72_UP-G13_Down
G13_Down-G72_UP
G72_UP-G59_UP
G59_UP-G72_UP
G82_Down-G72_UP
G72_UP-G82_Down
G96_Down-G72_UP
G72_UP-G96_Down
G97_Down-G72_UP
G72_UP-G97_Down
G38_Down-G72_UP
G72_UP-G38_Down
G32_Down-G72_UP
G72_UP-G32_Down
G72_UP,G82_Down-G59_UP
G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G82_Down-G59_UP,G72_UP
G72_UP-G59_UP,G82_Down
G59_UP-G72_UP,G82_Down
G72_UP,G96_Down-G59_UP
G59_UP,G96_Down-G72_UP

G59_UP,G72_UP-G96_Down
G96_Down-G59_UP,G72_UP
G72_UP-G59_UP,G96_Down
G59_UP-G72_UP,G96_Down

For template 2:

1. SIZE OF RULE >= 3: **12**

G72_UP,G82_Down-G59_UP
G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G82_Down-G59_UP,G72_UP
G72_UP-G59_UP,G82_Down
G59_UP-G72_UP,G82_Down
G72_UP,G96_Down-G59_UP
G59_UP,G96_Down-G72_UP
G59_UP,G72_UP-G96_Down
G96_Down-G59_UP,G72_UP
G72_UP-G59_UP,G96_Down
G59_UP-G72_UP,G96_Down

2. SIZE OF BODY >= 2: **6**

G72_UP,G82_Down-G59_UP
G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G72_UP,G96_Down-G59_UP
G59_UP,G96_Down-G72_UP
G59_UP,G72_UP-G96_Down

3. SIZE OF HEAD >= 2: **6**

G82_Down-G59_UP,G72_UP
G72_UP-G59_UP,G82_Down
G59_UP-G72_UP,G82_Down
G96_Down-G59_UP,G72_UP
G72_UP-G59_UP,G96_Down
G59_UP-G72_UP,G96_Down

For template 3:

1. BODY HAS ANY OF G1_UP AND HEAD HAS 1 OF G59_UP: **1**

G1_UP-G59_UP

2. BODY HAS ANY OF G1_UP OR HEAD HAS 1 OF G6_UP: **12**

G1_UP-G59_UP
G1_UP-G72_UP
G1_UP-G38_Down
G1_UP-G54_UP
G1_UP-G70_Down
G1_UP-G10_Down
G1_UP-G67_UP
G13_Down-G6_UP
G28_Down-G6_UP
G59_UP-G6_UP
G38_Down-G6_UP
G32_Down-G6_UP

3. BODY HAS 1 OF G1_UP OR HEAD HAS 2 OF G6_UP: **7**

G1_UP-G59_UP
G1_UP-G72_UP
G1_UP-G38_Down
G1_UP-G54_UP
G1_UP-G70_Down
G1_UP-G10_Down
G1_UP-G67_UP

4. HEAD HAS 1 OF G1_UP AND BODY HAS 0 OF DISEASE: **7**

G59_UP-G1_UP
G72_UP-G1_UP
G38_Down-G1_UP
G54_UP-G1_UP
G70_Down-G1_UP
G10_Down-G1_UP
G67_UP-G1_UP

5. HEAD HAS 1 OF DISEASE OR RULE HAS 1 OF (G72_UP, G96_DOWN): **24**

G72_UP-G1_UP
G1_UP-G72_UP
G72_UP-G13_Down
G13_Down-G72_UP
G72_UP-G59_UP
G59_UP-G72_UP
G96_Down-G59_UP
G59_UP-G96_Down
G82_Down-G72_UP
G72_UP-G82_Down
G97_Down-G72_UP
G72_UP-G97_Down
G38_Down-G72_UP
G72_UP-G38_Down
G32_Down-G72_UP
G72_UP-G32_Down
G96_Down-G82_Down
G82_Down-G96_Down
G72_UP,G82_Down-G59_UP
G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G82_Down-G59_UP,G72_UP
G72_UP-G59_UP,G82_Down
G59_UP-G72_UP,G82_Down

6. BODY HAS 1 of (G59_UP, G96_DOWN) AND SIZE OF RULE >=3: **6**

G59_UP,G82_Down-G72_UP
G59_UP,G72_UP-G82_Down
G59_UP-G72_UP,G82_Down
G72_UP,G96_Down-G59_UP
G59_UP,G72_UP-G96_Down
G96_Down-G59_UP,G72_UP
G59_UP-G72_UP,G96_Down