Frequent Pattern Analysis

Before starting Task, we discretized the datasets D using Equi Width Binning method of Weka. The discretized version of datasets is saved in DiscretizedD.csv file. The interval/pairs used for discretization is mapped to distinct IDs storing in a file called DiscretizationMap.csv. Furthermore the discretized datasets of DiscretizedD.csv is mapped with the IDs, is stored in MappingData.csv which is used for calculating the FP Growth Frequent Pattern. The FP Pattern is used for different task, so, it is stored in a file called FPPattern.csv. This file contains pattern with number of occurrence in every line.

Task 1: To Calculate Closed Pattern and Minimal Generator, i implemented a function name ClosedPatternMinimalGen() which provides the output for the Closed Pattern and Minimal Generator.
The output is stored in a file named ClosedAndMG.csv. Every line is formatted as:
ClosedPattern   NumberOfOccurrence   MinimalGenerator

Task 2: I implemented BitwiseMds() function for this task. Bit-set representations of the mds of patterns is stored in a file called BitSet.csv. Along with it, it can calculate intersection/difference which can be used for Jaccard Similarity in the Task 4.

Task 3: EmergingPattern() function is used for the calculation of High GrowthRate Pattern. For its calculation, BitwiseMds() value can be read to calculate support for every pattern in both the class. The Maximum GrowthRate is calculated and printed as output.

Task 4: It implemented a function named JaccardSimilarity(). PSkEps.csv file contains the first line as the maximum value for the maximization of the given function. The output format is as follows:
Pattern GrowthRate Class1 Class0
This output ,however, was calculated in Task3. The top K patterns depends on the growthrate which acts as the threshold value.
After it, jaccard similarity was calculated taking every pairs of top K patterns. This similarity was stored in the file named as PSkEPJaccard.csv. The output format is as follows:
Pattern1 Pattern2 JaccardSimilarity

The OutPut for every top 5 line of every file after taking Mammographic.txt as file, 10 as minimum support ratio and 10 as minimum growth rate is as follows:

DiscretizedMap.csv

```
Attribute Interval/Value-Pair  ID
 ATTR1 (-inf-11] 1
 ATTR1 (11-22] 2
 ATTR1 (22-33] 3
 ATTR1 (33-44] 4
```

DiscretizedD.csv

```
Attribute1 Attribute2 Attribute3 Attribute4 Attribute5

(-inf-11] (64.8-80.4] (2.8-3.4] (4.2-inf) (2.8-3.4]
(-inf-11] (49.2-64.8] (3.4-inf) (4.2-inf) (2.8-3.4]
(-inf-11] (-inf-33.6] (-inf-1.6] (-inf-1.8] (2.8-3.4]
(-inf-11] (49.2-64.8] (-inf-1.6] (4.2-inf) (2.8-3.4]
(-inf-11] (64.8-80.4] (-inf-1.6] (3.4-4.2] (2.8-3.4]
```

ClosedAndMG.csv

```
ClosedPattern   NumberOfOccurrence   MinimalGenerator
{22,12,7,16,1} 10 {22,12,7,16}
{7,1} 211 {7,1}
{12,8,24,1} 60 {12,8,24}
{9,1} 224 {9}
{14,9,1} 22 {14,9}
```

The Highest Growth Rate with pattern is:

```
Highest GrowthRate 27.3703 Emerging Pattern 14 16 1
```

PSkEPs.csv

```
14.211008823529408
Pattern  GrowthRate Class1 Class0
{14 16 1}  27.3703   0.0099  0.0422
{20 9 24}  27.3701  0.1067  0.0094
{18 12 24}  26.4265   0.0025  0.0258
{8 24 1}  26.4265   0.3325  0.3513
{14 9 24 1}  18.247  0.0347  0.0141
```

PSkEPJaccard.csv

```
Pattern1 Pattern2 JaccardSimilarity
{14 16 1}  ,  {14 16 1}  , 1.0
{14 16 1}  ,  {20 9 24}  , 0.0
{14 16 1}  ,  {18 12 24}  , 0.0
{14 16 1}  ,  {8 24 1}  , 0.0132
{14 16 1}  ,  {14 9 24 1}  , 0.0769
```

There are other file called BitSet.csv, FPPattern.csv, MappingData.csv and discretized.txt which stores the intermediate value.