Mining Frequent Patterns in Uncertain Transactional Databases

What is frequent pattern mining?

Frequent pattern mining aims to discover all interesting patterns in a transactional database that have **support** no less than the user-specified **minimum support** (**minSup**) constraint. The **minSup** controls the minimum number of transactions that a pattern must appear in a database.

What is the uncertain transactional database?

A transactional database is a collection of transactions, where each transaction contains a transaction-identifier and a set of items with their respective uncertain value.
A hypothetical transactional database containing the items ***a, b, c, d, e, f, and g*** as shown below

tid	Transactions
1	a(0.4) b(0.5) c(0.2) g(0.1)
2	b(0.2) c(0.3) d(0.4) e(0.2)
3	a(0.3) b(0.1) c(0.3) d(0.4)
4	a(0.2) c(0.6) d(0.2) f(0.1)
5	a(0.3) b(0.2) c(0.4) d(0.5) g(0.3)
6	c(0.2) d(0.7) e(0.34) f(0.2)
7	a(0.6) b(0.4) c(0.3) d(0.2)
8	a(0.2) e(0.2) f(0.2)
9	a(0.1) b(0.3) c(0.2) d(0.4)
10	b(0.3) c(0.2) d(0.1) e(0.6)

Note: Duplicate items must not exist in a transaction.

Acceptable format of uncertain transactional databases in PAMI

Each row in a transactional database must contain only items with their respective uncertain values.

a(0.4) b(0.5) c(0.2) g(0.1)
b(0.2) c(0.3) d(0.4) e(0.2)
a(0.3) b(0.1) c(0.3) d(0.4)
a(0.2) c(0.6) d(0.2) f(0.1)
a(0.3) b(0.2) c(0.4) d(0.5) g(0.3)
c(0.2) d(0.7) e(0.34) f(0.2)
a(0.6) b(0.4) c(0.3) d(0.2)
a(0.2) e(0.2) f(0.2)
a(0.1) b(0.3) c(0.2) d(0.4)
b(0.3) c(0.2) d(0.1) e(0.6)

What is the input to uncertain frequent pattern mining algorithms

Algorithms to mine the uncertain frequent patterns requires uncertain database and minSup (specified by user).

- Transactional database in following formats:

  > - In string format (`/Users/Likhitha/Downlaods/sampleInputFile.txt')
  > - In URL format (`https://www.u-aizu.ac.jp/~udayrage/datasets/transactionalDatabases/transactional_T10
  > - In DataFrame format (dataframe variable with heading `Transactions` which contains only items and `uncertain` which contains uncertain values of each item in transaction respectively)

- minSup should be mentioned in **count (beween 0 to length of database)** or __percentage (multiplied with length of database)

What is the output of uncertain frequent pattern mining algorithms

The output of these algorithms is in two ways:

> - Saves the patterns in user specified output file.
> - Returns the patterns in dataframe variable.

How to run the frequent pattern algorithm in terminal

- Download the code from github.
- Navigate to PAMI folder where you downloaded the file.
- Go to uncertainFrequentPattern/basic folder

Execute the following command on terminal.

python3 algorithmName.py `path of Sample input file` `path of output file` minSup `seperator`

Sample command to execute the PUFGrowth code in uncertainFrequentPattern/basic folder

python3 `PUFGrowth.py` `/Users/Donwloads/inputFile.txt` `/Users/Downloads/outputFile.txt` 0.05 `' '`

How to implement the code by importing PAMI package

Import the PAMI package executing: **pip3 install PAMI**

Run the below sample code by making simple changes

- Replace sampleInputFile name or path in place of iFile and sampleOutputFile name or path in place of oFile
- Specify the minSup (like 10 or 0.1) in place of minSup
- Specify the seperator of input file after minSup. (If no seperator is specified the default tab seperator is considered for input file)

```
import PAMI.uncertainFrequentPattern.basic.PUFGrowth as alg
obj = alg.PUFGrowth(iFile, minSup, ',')
obj.startMine()
obj.savePatterns(oFile) (to store the patterns in file).
Df = obj.getPatternsAsDataFrame() (to store the patterns in dataframe)
obj.printStats()
```

What is the output of frequent pattern mining algorithms

Returns the pattern and support respectively with minSup=0.5 for above sample database.

The output in file format:

```
f 0.5
e 1.3399999999999999
b 2.0
b a 0.56
b c 0.51
a 2.0999999999999996
a c 0.6100000000000001
c 2.7
d 2.9000000000000004
c d 0.8600000000000001
```

The output in DataFrame format:

	Patterns	Support
0	f	0.50
1	e	1.34
2	b	2.00
3	b a	0.56
4	b c	0.51
5	a	2.09
6	a c	0.61
7	c	2.70
8	d	2.90
9	c d	0.86