An example for Frequent Pattern Mining using the Eunomia package

11/12/2020

Contents

0.0.1 Connect to the database

```
### Define database parameters
cdmdatabaseschema = "main"
resultsdatabaseschema = "main"
fpm_inputFile = "fpm_testing.txt"
fpm_outputFile_SPAM = "fpm_testingResults_SPAM.txt"
fpm_outputFile_SPADE = "fpm_testingResults_SPADE.txt"
fpm_outputFile_prefixSpan = "fpm_testingResults_prefixSpan.txt"
fpm_outputFile_Clasp = "fpm_testingResults_Clasp.txt"
fpm_outputFile_CMClasp = "fpm_testingResults_CMClasp.txt"
fpm_outputFile_MaxSP = "fpm_testingResults_MaxSP.txt"
fpm_outputFile_VMSP = "fpm_testingResults_VMSP.txt"
fpm_outputFile_VGEN = "fpm_testingResults_VGEN.txt"
fpm_outputFile_RuleGrowth = "fpm_testingResults_RuleGrowth.txt"
fpm_outputFile_ERMiner =  "fpm_testingResults_ERMiner.txt"

connectionDetails <- Eunomia::getEunomiaConnectionDetails()
connection <- connect(connectionDetails)
```

```
## Connecting using SQLite driver
```

```
#on.exit(DatabaseConnector::disconnect(connection)) #Close db connection on error or exit
```

0.0.2 Define cohort

```
# Define cohort
cohort <- "SELECT person_id AS subject_id,
  condition_start_date AS cohort_start_date
INTO #diagnoses
FROM @cdm.condition_occurrence
WHERE condition_concept_id IN (
    SELECT descendant_concept_id
    FROM @cdm.concept_ancestor
    WHERE ancestor_concept_id = 4329847 -- Myocardial infarction
)
  AND condition_concept_id NOT IN (
    SELECT descendant_concept_id
    FROM @cdm.concept_ancestor
    WHERE ancestor_concept_id = 314666 -- Old myocardial infarction
);
INSERT INTO @cdm.cohort (subject_id, cohort_start_date, cohort_definition_id)
SELECT subject_id,
  cohort_start_date,
  CAST (1 AS INT) AS cohort_definition_id
FROM #diagnoses
INNER JOIN @cdm.visit_occurrence
  ON subject_id = person_id
    AND cohort_start_date >= visit_start_date
    AND cohort_start_date <= visit_end_date
WHERE visit_concept_id IN (9201, 9203, 262); -- Inpatient or ER;"

renderTranslateExecuteSql(connection, cohort, cdm = cdmdatabaseschema)
```

```
##   |
## Executing SQL took 0.0163 secs
```

```
sql <- "ALTER TABLE #diagnoses ADD cohort_definition_id INT NOT NULL DEFAULT(1)"

# Execute the script to receive the data
renderTranslateExecuteSql(connection, sql)
```

```
##   |
## Executing SQL took 0.000735 secs
```

```
querySql(connection, "SELECT count(*) FROM diagnoses;")
```

```
##   COUNT(*)
## 1       67
```

0.0.3 Get the data and close the connection

```r
# Define covariate settings
TemporalcovariateSettings_eunomia <- createTemporalCovariateSettings(useConditionOccurrence = TRUE,
                                                                     temporalStartDays = seq(-(60*365), -1, by = 1) ,
                                                                     temporalEndDays = seq(-(60*365)+1, 0, by = 1))

# Extract covariates
TemporalcovariateData_eunomia <- getDbCovariateData(connection = connection,
                         cdmDatabaseSchema = cdmdatabaseschema,
                         cohortDatabaseSchema = resultsdatabaseschema,
                         cohortTable = "diagnoses",
                         rowIdField = "subject_id",
                         covariateSettings = TemporalcovariateSettings_eunomia,
                         cohortTableIsTemp = TRUE)
```

```
## Sending temp tables to server
## Constructing features on server
##   |
## Executing SQL took 35.8 secs
## Fetching data from server
## Fetching data took 0.151 secs
```

```r
disconnect(connection)
```

0.0.3.1 Frequent pattern mining

0.1 Prepare the data

```r
testData <- getInputFileForFrequentPatterns(covariateDataObject = TemporalcovariateData_eunomia, fileTo
```

```
## Extracting temporal data...

## Extracting covariate names...

## Generating input file for frequent pattern mining...

## Input data has been created succesfully and saved in fpm_testing.txt
```

0.2 Run SPAM

```r
spam_frequentPatterns <- runFrequentPatterns(algorithm = "SPAM",
                                             inputFile = fpm_inputFile,
                                             outputFile = fpm_outputFile_SPAM,
                                             minsup = 0.5,
                                             showID = TRUE)
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/I
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(spam_frequentPatterns)
```

```
##                     Sequence Count   Support
## 1 Streptococcal sore throat     34 0.5074627
## 2            Osteoarthritis     61 0.9104478
## 3          Acute bronchitis     60 0.8955224
## 4 Coronary arteriosclerosis     66 0.9850746
## 5   Acute viral pharyngitis     64 0.9552239
## 6     Myocardial infarction     67 1.0000000
```

0.3 Run SPADE

```
spade_frequentPatterns <- runFrequentPatterns(algorithm = "SPADE",
                                        inputFile = fpm_inputFile,
                                        outputFile = fpm_outputFile_SPADE,
                                        minsup = 0.5,
                                        showID = TRUE)
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/I
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(spade_frequentPatterns)
```

```
##                     Sequence Count   Support
## 1 Streptococcal sore throat     34 0.5074627
## 2            Osteoarthritis     61 0.9104478
## 3          Acute bronchitis     60 0.8955224
## 4 Coronary arteriosclerosis     66 0.9850746
## 5   Acute viral pharyngitis     64 0.9552239
## 6     Myocardial infarction     67 1.0000000
```

0.4 Run prefixSpan

```
pS_frequentPatterns <- runFrequentPatterns(algorithm = "prefixSpan",
                                        inputFile = fpm_inputFile,
                                        outputFile = fpm_outputFile_prefixSpan,
                                        minsup = 0.5,
                                        showID = TRUE)
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/I
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(pS_frequentPatterns)
```

```
##                                                                           Sequence Count   Support
## 1                                                       Streptococcal sore throat    34 0.5074627
## 2                          Streptococcal sore throat => Streptococcal sore throat    34 0.5074627
## 3 Streptococcal sore throat => Streptococcal sore throat => Myocardial infarction    34 0.5074627
## 4                              Streptococcal sore throat => Myocardial infarction    34 0.5074627
## 5                                                                  Osteoarthritis    61 0.9104478
## 6                                     Osteoarthritis => Coronary arteriosclerosis    41 0.6119403
```

0.5 Run Clasp

```
clasp_frequentPatterns <- runFrequentPatterns(algorithm = "Clasp",
                                              inputFile = fpm_inputFile,
                                              outputFile = fpm_outputFile_Clasp,
                                              minsup = 0.50,
                                              showID = TRUE )
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Res
```

```
head(clasp_frequentPatterns)
```

```
##
## 1                 Viral sinusitis => Viral sinusitis => Acute bronchitis => Acute bronchitis => Viral
## 2   Viral sinusitis => Viral sinusitis => Viral sinusitis => Viral sinusitis => Viral sinusitis => Vi
## 3                         Viral sinusitis => Viral sinusitis => Acute viral pharyngitis => Acute vira
## 4 Viral sinusitis => Viral sinusitis => Acute bronchitis => Acute bronchitis => Acute viral pharyngit
## 5                                                                               Viral sinusitis => Vi
## 6         Acute viral pharyngitis => Acute viral pharyngitis => Acute viral pharyngitis => Acute vira
##   Count   Support
## 1    45 0.6716418
## 2    46 0.6865672
## 3    54 0.8059701
## 4    38 0.5671642
## 5    62 0.9253731
## 6    46 0.6865672
```

0.6 Run CM-Clasp

```
cmclasp_frequentPatterns <- runFrequentPatterns(algorithm = "CM-Clasp",
                                              inputFile = fpm_inputFile,
                                              outputFile = fpm_outputFile_CMClasp,
                                              minsup = 0.50,
                                              showID = TRUE )
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Res
```

```
head(cmclasp_frequentPatterns)
```

```
##
## 1                 Viral sinusitis => Viral sinusitis => Acute bronchitis => Acute bronchitis => Viral
## 2   Viral sinusitis => Viral sinusitis => Viral sinusitis => Viral sinusitis => Viral sinusitis => Vi
## 3                         Viral sinusitis => Viral sinusitis => Acute viral pharyngitis => Acute vira
## 4 Viral sinusitis => Viral sinusitis => Acute bronchitis => Acute bronchitis => Acute viral pharyngi
## 5                                                                               Viral sinusitis => Vi
## 6         Acute viral pharyngitis => Acute viral pharyngitis => Acute viral pharyngitis => Acute vira
##   Count   Support
## 1    45 0.6716418
## 2    46 0.6865672
## 3    54 0.8059701
## 4    38 0.5671642
## 5    62 0.9253731
## 6    46 0.6865672
```

0.7 Run VMSP

```
vmsp_frequentPatterns <- runFrequentPatterns(algorithm = "VMSP",
                                              inputFile = fpm_inputFile,
                                              outputFile = fpm_outputFile_VMSP,
                                              minsup = 0.50,
                                              showID = TRUE )
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(vmsp_frequentPatterns)
```

```
##
## 1
## 2                                                                             Streptococcal sore thr
## 3                         Acute bronchitis => Acute viral pharyngitis => Acute viral pharyngitis => (
## 4                                Acute bronchitis => Acute bronchitis => Acute viral pharyngitis => (
## 5 Viral sinusitis => Viral sinusitis => Coronary arteriosclerosis => Coronary arteriosclerosis => Vi
## 6                       Osteoarthritis => Osteoarthritis => Viral sinusitis => Viral sinusitis => Vi
##   Count   Support
## 1    67 1.0000000
## 2    34 0.5074627
## 3    34 0.5074627
## 4    34 0.5074627
## 5    38 0.5671642
## 6    34 0.5074627
```

0.8 Run VGEN

```
vgen_frequentPatterns <- runFrequentPatterns(algorithm = "VGEN",
                                              inputFile = fpm_inputFile,
                                              outputFile = fpm_outputFile_VGEN,
                                              minsup = 0.50,
                                              showID = TRUE )
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(vgen_frequentPatterns)
```

```
##                     Sequence Count   Support
## 1                   #SUP: 67    67 1.0000000
## 2   Acute viral pharyngitis     64 0.9552239
## 3 Streptococcal sore throat     34 0.5074627
## 4 Coronary arteriosclerosis     66 0.9850746
## 5            Osteoarthritis     61 0.9104478
## 6          Acute bronchitis     60 0.8955224
```

0.9 Run RuleGrowth

```
ruleGrowth_frequentPatterns <- runFrequentPatterns(algorithm = "RuleGrowth",
                                              inputFile = fpm_inputFile,
                                              outputFile = fpm_outputFile_RuleGrowth,
                                              minsup = 0.50,
                                              minconf = 0.50,
                                              showID = FALSE #Does not retrieve IDs
                                              )
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(ruleGrowth_frequentPatterns)
```

```
##                                                                       Sequence Count   Support Confid
## 1                                         Osteoarthritis ==> Acute_bronchitis    40 0.5970149  0.655
## 2                 Osteoarthritis,Acute_viral_pharyngitis ==> Acute_bronchitis    39 0.5820896  0.661
## 3 Osteoarthritis,Acute_viral_pharyngitis,Viral_sinusitis ==> Acute_bronchitis    39 0.5820896  0.661
## 4                         Osteoarthritis,Viral_sinusitis ==> Acute_bronchitis    40 0.5970149  0.655
## 5                                         Acute_bronchitis ==> Osteoarthritis    51 0.7611940  0.850
## 6                 Acute_bronchitis,Acute_viral_pharyngitis ==> Osteoarthritis    42 0.6268657  0.724
```

0.10 Run RuleGrowth

```
erminer_frequentPatterns <- runFrequentPatterns(algorithm = "ERMiner",
                                              inputFile = fpm_inputFile,
                                              outputFile = fpm_outputFile_ERMiner,
                                              minsup = 0.50,
                                              minconf = 0.5,
                                              showID = TRUE #Does not retrieve IDs
                                              )
```

```
## Analysing 67 sequence IDs...Running frequent pattern algorithm...[1] "java -jar /Library/Frameworks/
## The command line that has been running is: java -jar /Library/Frameworks/R.framework/Versions/4.0/Re
```

```
head(erminer_frequentPatterns)
```

```
##                                                  Sequence Count   Support Confidence
## 1          Osteoarthritis ==> Coronary_arteriosclerosis    41 0.6119403  0.6721311
## 2        Acute_bronchitis ==> Coronary_arteriosclerosis    53 0.7910448  0.8833333
## 3 Acute_viral_pharyngitis ==> Coronary_arteriosclerosis    57 0.8507463  0.8906250
## 4            Acute_viral_pharyngitis ==> Osteoarthritis    50 0.7462687  0.7812500
## 5            Osteoarthritis ==> Acute_viral_pharyngitis    42 0.6268657  0.6885246
## 6          Acute_viral_pharyngitis ==> Acute_bronchitis    55 0.8208955  0.8593750
```