MEIC 2018/19 Ciência de Dados

Lab 6: Pattern Mining

Part I

1. **Load, analyze and prepare categorical, real-valued and transactional data**
   a. Load *vote*, *marketing*, *zoo*, *supermarket*, *ionosphere*, *page_blocks* and *dermatology* datasets
   b. *vote*, *marketing* and *zoo* datasets either have integer or nominal attributes:
      i. denormalize them to produce binary data (without discretization)
      ii. produce a transactional dataset from the denormalized data
      iii. analyze their properties: visualize data, compute item frequencies
   c. Map *supermarket* into a transactional dataset and analyze item distributions
   d. Prepare the real-valued *ionosphere* and *page_blocks* datasets:
      i. Discretize them using:
         1. equal-width discretization
         2. equal-frequency discretization
         3. [*optional*] cutting-off breakpoints of a Gaussian distribution after attribute normalization
      ii. Compare differences depending on the applied discretization strategy and number of symbols (suggested: 3 and 5)
   e. Map the *dermatology* dataset (composed of both categorical and numeric attributes) into a transactional dataset using previous principles

2. **Perform pattern mining** on transactional data
   a. Discover the set of frequent itemsets on the *vote*, *marketing*, *zoo* and *dermatology* data by iteratively decreasing the support threshold until a reasonable number of patterns are produced
   b. Identify the set of closed itemsets from the outputted patterns
   c. Discover the set of rules using *vote*, *marketing*, *zoo* and *dermatology* data using multiple confidence thresholds (suggested 70% and 90%)
   d. Critically analyze the produced results according to:
      i. the number and interestingness criteria (inc. confidence and lift) of the discovered rules
      ii. the time required to discover the rules

a. Analyze the *supermarket* dataset:
   - i. how the discovered association rules vary with the inputted support and confidence thresholds?
   - ii. can support and confidence be dynamically parameterized? How?

e. Mine the real-valued *ionosphere* and *page_blocks* datasets:
   - i. Analyze the differences associated with the outputted association rules in accordance with the discretization strategy and number of symbols

3. Evaluating patterns

Select the association rules produced by two of the listed datasets and retrieve indicators of interestingness, such as χ^2(chiSquared), support, confidence, conviction, cosine, coverage, leverage, lift, and odds-ratio for the selected association rules. Critically analyze the gathered rules.

Part II

4. Load and analyze sequential databases

a. Load *example*, *sign*, *FIFA*. and *msnbc* sequence databases

b. Analyze their properties: sequence length and frequency of items

5. Sequential pattern mining

a. apply a sequential pattern mining algorithm (e.g. CSPADE in R and PrefixSpan in Python) to discover sequential patterns in *example*, *sign*, *FIFA*. and *msnbc* databases by incrementally decreasing the support until a tractable number of sequential patterns are discovered

b. analyze the pattern and coverage transactions of the produced patterns

c. identify the closed sequences (sequences that are not subsequence of any other outputted sequence)

d. advanced aspects
   - i. generate association sequence rules of varying confidence (use R)
   - ii. select top-k relevant sequences (use Python)

Part III

6. **Pencil-and-paper** exercise. Consider the following dataset where $a_{ij} \in [-1,1]$

$y1$	$y2$	$y3$	$y4$	$y5$
0.9	-0.3	-0.2	0.1	A
-0.3	0.3	0.4	0.4	B
0.6	0.6	-0.1	-0.1	A
0.6	-0.7	-0.2	-0.1	C
-0.2	0.2	-0.7	-0.4	A

a. Map the above dataset into two transactional datasets:

   i. discretize numeric attributes using 4 equal-width intervals

   ii. rediscretize rows along $y1$-$y5$ attributes using 2 bins of equal-frequency

b. Given the 2nd transactional dataset, apply the *Apriori* algorithm:

   i. extract frequent itemsets with support 0.60

   ii. generate association rules with confidence 0.60

   iii. select 2 rules and critically compare their confidence, lift and conviction

Bonus: show results from **exercises 3** and **5.a** to have your mark

Resources (packages):

- ARM/FIM
  - R: arules
  - Python: dplyr, mlxtend.frequent_patterns, orangecontrib.associate, pymining
- SPM
  - R: arulesSequences
  - Python: pymining, prefixspan