Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Knowledge Discovery in Databases

SS 2016

Exercise 5: Hierarchical Association Rules, Frequent Itemset Mining, Clustering

Regarding tutorials on 25.05.-27.05.2016.

Exercise 5-1 Hierarchical Association Rules

Let A, A_1, A_2, B be items. A is a generalization of A_1 and A_2. Prove or contradict (by giving a counterexample) the following assumptions:

(a) $support(A \Rightarrow B) = support(A_1 \Rightarrow B) + support(A_2 \Rightarrow B)$

(b) If $support(A_1 \Rightarrow B) > minSup$, then $support(A \Rightarrow B) > minSup$.

(c) If $support(A \Rightarrow B) > minSup$, then $support(A_1 \Rightarrow B) > minSup$.

Exercise 5-2 Frequent Itemset Mining on Bakery Receipts

In this exercise, we want to perform frequent itemset mining on a real world dataset. The *Extended Bakery* [1] dataset contains 75.000 receipts collected from a bakery chain with several stores distributed along the West Coast in the US. Several smaller subsets containing fewer transactions are also available. The receipts can be retrieved from the csv-file `75000-out1.csv`. Each row in the file corresponds to a receipt and starts with an ID, followed by the list of bought items. For each of the 50 distinct item IDs, the name, category, price and type are provided in an additional csv-file `goods_description.csv`. Both files can be downloaded from the course website.

For frequent itemset mining, you are free to write your own algorithm implementations, but you can also just use an existing library. For instance, if you decide to work with *python*, you can use the *PyFIM* [2] library.

(a) Start with loading the receipts and the dictionary from the bakery dataset files.

(b) Perform frequent itemset mining with different $minSup$ values. If possible, also try different algorithms, such as the Apriori and the FP-Growth algorithm. Discuss your results.

(c) Also perform association rule mining, again with varying parameter values for $minSup$ and $minConf$. Can you find some interesting rules?

(d) If you used *PyFIM* with the *arules* function in the previous exercise, you can include different evaluation measures in the result. Explore some of the further evaluation measures presented in the lecture.

[1] `https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery`
[2] `http://www.borgelt.net/pyfim.html`

(e) The file `goods_description.csv` provides the additional attributes 'category' and 'type' for each item, which induce a simple generalization hierarchy of bakery products. Compared to your results from (c), can you find more interesting hierarchical association rules? (Hint: In *PyFIM*, it is also possible to use strings as item identifiers.)

Exercise 5-3 k**-Means**

In this exercise, we will implement a k-means algorithm.

(a) Load the dataset `blobs.csv` from the course website and visualize it.

(b) Implement a function `kmeans(data,k)`.

(c) Optional: Visualize intermediate results after each iteration.

(d) Apply your method to the dataset from (a) using different values for k and plot the results.

(e) Load the dataset `mouse.csv` from the course website and visualize it. Apply your method to the mouse dataset as well and discuss the differences.