Itemsets and Association Rules Mining
Practical work

Yannick Toussaint

4 octobre 2020

To carry this project, you will be by group of 3. The dataset is large enough for the groups to propose different analysis of the same dataset... So do not copy from one group to an other !

Your mission : *You just have been recruted as a journalist and you are in charge of analysing a survey that has just been published (2016 !). Depending the readers of your journal, you have to decide what you will focus on....* If you are unscrupulous about the results you can publish, you still have scientific values and want to use a rigorous data mining process.

The project aims at analysing and understanding the content of a dataset using Pattern Mining and Association Rules techniques. The following dataset is a survey on people living in France, more specifically in Grand-Est. The project is very open... You are relatively free to choose what you want to focus on. You may have very general observation or very specific ones. You also have to adapt the work to the characteristics of your computer, you may adapt tresholds and/or dataset size to be able to get results. Please, be precise to explicit which constraints you applied.

Thus, you will extract itemsets and association rules from a real dataset. The goal of the project is to be able to observe certain frequent cooccurrences among attributes. You are free to decide which attributes you will study but you will have to justify and explain how you built your observations.

1 The dataset

The data set comes from Insee which is the national institute for statistics and economic studies. Each line describe a person/familly. The national collection of the dataset contains 3,3 million of individuals (see `https://www.insee.fr/fr/statistiques/4171523?sommaire=4171558#consulter-sommaire`). The Grand-Est Region dataset contains 1,474,560 records. Each individual is described by 57 multivaluated attributes which are either a symbolic or numerical values. Some attributes have more that 120 different values. Thus, the dataset should be simplified. The pdf file gives the meaning of the attributes.

I could suggest to reduce the list of attributes to the following, but feel free to make your own list if you prefer (justify) : REGION, AGER20, ANARR, ANEMR, ASCEN, BAIN, BATI, CATL, CHOS, CLIM, CMBL, COUPLE, CS1, CUIS, DEROU, DIPL_15, EAU, EGOUL, ELEC, EMPL, ETUD, GARL, ILTUU, IMMI, INAT, INFAM, INPER, MOCO, MODV, NA38, NATC, NATNC, NBPI, NPERR, RECH, SANI, SEXE, SFM, STOCD, TACT, VOIT, WC.

2 Experiments

Question 2.1
Write a program in python to prepare the data for SPMF. Your program should start with a list of attributes to keep and build a dataset where :

- One line = one individual,
- Each multivaluated attribute is transformed into single-valued attributes which are then associated to a number,
- All the attributes on a line are sorted from the lower to the higher value.

Of course, test your program on 10 lines from the dataset to make it faster...

Question 2.2
Write the program that decode the results from SPMF

Then, run SPMF to extract itemsets using the FPG algorithm. As the algorithm requires a lot of CPU and memory space, you may perform several tests, starting with a high support treshold to extract a *small* number of itemsets, and decreasing the treshold value to get a some more itemsets. SPMF is a java program. you may increased the default size of the java virtual machine calling SPMF with the following command : *java -Xms256m -Xmx2g ./SPMF* where Xms is the intial size of the virtual machine and Xmx is the maximum VM size.

The files that are produced in the following steps may be very big (up to millions of lines). It could be a good idea to use command lines to extract from these files the itemsets or association rules you want to observe.

Question 2.3
Choose 10 (rather different) itemsets of your choice. Justify the way you choose them. Comment each of them relatively to the domain. In order to comment and interprete an itemset, you may need to look/search other itemsets. Explain how you build you interpretation.

Question 2.4
Extract Asssociation Rules. As for itemsets, start with a high support and a high confidence and descrease the values up to the limit of your machine. Select about 10 attributes you want to focus on and look at the association rules involving these attributes. Explain how you select some rules, how you interprete these rules and formulate observation concerning the result of the survey.

3 Output of your work

Each tandem should send me a zip file with the following :

- A small report that answers the questions, justifies your choices, explains how you works, the difficulties you met and how you solved them. Insert a lot of examples in the report (itemsets, association rules...) to explain your analysis.
- The programs you developped with comments inside the program file. You should also write a readme to explain how to lunch the programs. I should be able to run your programs from the directory when it will be saved (on my machine). So please, manage the paths so that they do not depend on the machine ! Some of the program are use for preprocessing the data, some are for post-processing (filtering...)... Make it clear !
- The resulting files from SPMF
- A file gathering the different commands you used to analyse the results and what the commands can be used for.
- Any information you find useful to communicate