Batch 10 – Data Mining 1 – Unsupervised Learning – Individual Assignment 2

- **This is an individual assignment. Honor code: 3N-b.**
- **Please upload your work on the LMS by the deadline as specified on LMS.**
- **You may use any software to work on your assignment.**
- **In your submission, you must include the output or refer to the relevant output in the excel file while answering any question or justifying your answers.**
- **An answer without a justification will not be awarded any credit even if the answer is correct.**

---

1. **Cosmetics.xlsx**

   Step 1: Download the data on cosmetics purchases (**Cosmetics.xls**) from the textbook website (http://dataminingbook.com/).

   Step 2: Using R or any other tool you are familiar with, apply association rules to these data. You may choose the threshold support and confidence in such a way that you get approximately15-20 rules with, at least, a few of them with Lift Ratio greater than 1.

   Step 3: Order those rules in decreasing order of Lift Ratio.

   Consider the result of the association rules analysis to answer the following questions.

   a) Please include (copy-paste) the output ─ first fifteen rules along with header and input parameter details ─ in your submission.
   b) What is the support of the first rule? Explain how it has been calculated for this rule.
   c) What is the confidence of the first rule? Explain how it has been calculated for this rule.
   d) What is the lift ratio of the first rule? Explain how it has been calculated.
   e) Reviewing the first fifteen rules, comment on their *redundancy* (read as constructed from same item set/tuple). How many distinct rules did you find from the first 15 rules?
   f) Interpret the first three **distinct** (i.e., excluding the redundant ones, if any, as defined above) rules, in the output, in words.
   g) Based on the distinct rules that you identified in Part (f), suggest some action that'll benefit the business owner.

2. **Airlines Network**

Step 1: Download (1) *airports.dat*, (2) *airlines.dat*, and (3) *routes.dat* from http://openflights.org/data.html or use the files that were shared with you on LMS while doing the in-class exercise on airport networks. Refer to the data descriptions provided on the site itself. You may use the R scripts shown in class (also uploaded on LMS), or any other software packages that you are familiar with to complete the exercise.

Step 2: Create a directed network graph of airline routes using *routes.dat*

Step 3: Create community-based clusters. If you are using R, choose **leading.eigenvector** approach. If you are using any other software, use the equivalent algorithm provided with the software.

Step 4: Answer the following questions.

(a) What would you call a *community* in a social-media network? Intuitive, qualitative answers are expected/acceptable.

(b) Extend the definition of community (as you suggested above in Part a) to the community of airports.

(c) How many distinct airports are there in the dataset? How many *communities* of airports got identified? List the number of airports in each cluster/community in a table.

(d) Interpret/characterize, at least, five of the communities you identified above.

Step 5: Compute the centralities (in-degree, out-degree, in-closeness, eigenvector, betweenness) of each airport. Now, run k-Means clustering to group the airports based on their centralities alone. **Take k equal to the number of communities you obtained in Part c, above.**

(e) Do you observe the groups obtained in Step 4 to be similar to or different from what you obtained in Step 5? Why?

Step 6: Now, run k-Means clustering again on the airports based on their centralities. **Go with a value of k as you find appropriate.**

(f) Interpret the clustering outcome.

Step 6: Carefully observe the centralities of the airports in the dataset.

(g) If your organization is planning on launching a new flight service on a couple of new routes, what will that be (based on the information you have in this data alone)? Explain your answer. What other information would have helped you to make a better decision?