LE/EECS 4412 – Data Mining

Winter 2022 – Section M

Course Project

Submission Deadline: Sunday March 27, 2022, before 23:59

Part 4: Phase 2

Objectives

The purpose of this activity is to learn how to apply data mining techniques related to classification, association and clustering.

Submission Requirements

- Please submit your results/output in a PDF file. Submit your code in a text (.txt) file.
- The name of the files must be your YorkU student number(s) such as `100131001-100131002.pdf`.
- File must be uploaded to the Part4 submission link provided. The name of submission on the link must also be the same as your file name above.
- **If you use any tools such as Python, R, Excel etc., then upload all related material (code, scripts, worksheets, CSVs) as a single file zip file to the submission link. Submission of this material is compulsory to get the grade in this part.** Name this file as `100131001-100131002.zip`.
- Your diagrams, if applicable, can be hand drawn or digital. In both cases, the images must be clear.
- If the files do not open properly or the content is not clear, then you will be awarded zero.
- **Deadline is March 27, 2022, before 23:59. Late submission is permitted with 10% penalty per day only up to 7 days after the original deadline.**
- **Your submissions will be verified using Turnitin (or some other suitable tool) for originality. 40% or more similarity will be awarded zero in the assignment and reported to the department. We may report similarity less than 40% if it is of significant nature.**

Warning: Please add the necessary headings and labels to your report so that the TAs can understand the different parts properly. Anything that we cannot understand will be awarded zero.

General Instructions

- **Please add the statement given in Appendix A in the start of your report. Each team member must add her own statement.**
- **Please make sure that your document is easy to understand; clearly add the task number, part number, captions and foot notes wherever required. If the TA can't locate the answer, then it is your responsibility.**
- **Make sure that images/screenshots are clear. If the image is big then split it into multiple parts. Clearly write their purpose. You can add multiple images even if the question statement doesn't say so to make sure that your answer is easy to comprehend.**
- **Highlight the significant parts of each image so that TA can easily identify the required information.**
- **Add necessary explanation to make sure that TA can understand different parts of your document.**

Task 1: Association Analysis

- Select a suitable table/object type from your data set for association analysis. If you have too many data rows in this table, then only keep 1000 rows.
  - Briefly explain/justify why you think it is more suitable than the others.
  - submit this data in the form of CSV file; name it as `100131001-100131002—T1Old.csv`
- Discretize your dimensions using a suitable technique:
  - For each dimension use a different discretization technique
  - Add first 10 rows of data in a tabular form in your report as sample
  - submit the complete modified data (after discretization) in the form of CSV file; name it as `100131001-100131002—T1Disc.csv`.
- Using some suitable tool/technique:
  - Find the 10 most frequent item-sets. The length/size of these item-sets must be in the range log(n)—n/2 where n is the number of items in your table. Compute the support for these item-sets.
  - From your 10 most frequent items-sets, design 10 association rules, 5 with most confidence and 5 with least confidence. Specify the confidence for each of these rules.
  - For each of your top five rules, compute any 5 different measures of interest from Chapter05 and add the results to your report.
  - Is it possible to apply Z-statistic on any of your top 5 rules? Explain why or why not.

Task 2: Clustering Analysis

- Select a suitable table/object type from your data set for clustering analysis. If you have too many data rows in this table, then only keep 1000 rows.
  - Perform any necessary preprocessing steps. Add explanation to your report.
  - submit the original data in the form of CSV file; name it as `100131001-100131002—T2Org.csv`
  - submit the modified data in the form of CSV file; name it as `100131001-100131002—T2Mod.csv`
- Using some suitable tool/technique:
  - Perform the k-means clustering of your data with k=3, 4, 5 and Euclidian distance. Which value of k produces the least SSE (sum of squared errors)?
  - Add a class attribute to your data and assign class labels to your data rows based on the k-means clustering. Pick your best k from the previous step.
  - submit the class label column/dimension in the form of CSV file; name it as `100131001-100131002—T2Class.csv`

Task 3: Classification Revisited

- Use your data from Task 2, along with the class labels, and using some suitable tool/technique:
  - Create a naïve bayes classifier for your data set.
  - Use 3-fold cross validation approach to evaluate your classifier. For your training set in each run, describe at least three measures from chapter04.
  - Draw the ROC curve based on test data set.

Task 4: Progress on your Objectives

- Is/are any of your initial five questions is/are or may be addressed at this point? Describe.
- Do you need to restructure or reformulate your questions? If yes, then provide the updated set of problems. If no, then explain why not.

Selection of tools

- To perform the tasks described in this part, you are free to use any languages and tools of your choice and comfort.
- Any source codes and tool specific documents must be provided with the submission document.
- Any installation and usage instructions must also be provided in the document.

Appendix A

The following statement must be added in the beginning of your report. Each team member must submit her own independent statement. The signature can be electronic, or you can add a scan of statement with handwritten signature.

I ***student_name*** student ID # ***student_id*** acknowledge that I have contributed at least 30% time and effort to the preparation of this report and work discussed herein.

Student_Signature