COMP7103 Assignment 2

Due date: Nov 16, 2022 11:59pm

Question 1 Classification [50 marks]

The ***Sports articles for objectivity analysis Data Set*** is the result of extracting features from 1000 sports articles and manually labelled as either "**subjective**" or "**objective**" using Amazon Mechanical Turk. A copy of the dataset and the description is also available on Moodle.

Link to the dataset: https://archive.ics.uci.edu/ml/datasets/Sports+articles+for+objectivity+analysis

Use Weka and/or any other tools, perform the following.

- Prepare a training dataset using the data from `TEXT0405` to `TEXT0904`. Check that your training dataset consists of an equal number of "**subjective**" and "**objective**" records.
- Add one derived attribute "**site**", which equals to the domain of the URL. For instance, "**site**" will be "*msn.foxsports.com*" for record `TEXT0001`.
- Find a **classification model** for the class label "**Label**" in the dataset. Note that you are expected to try different ideas and refine your model in the data mining process. In particular, you must experiment and document your process of data reduction (see parts a and b below).

Answer the following questions.

a) Briefly describe **what you did** in constructing the model, including but not limited to data preprocessing, attribute selection, parameter tuning, training and testing data construction, model building, and evaluation, etc. You may omit some of the above-mentioned items if you did not do them. You must clearly describe what your **final classification model** is, which will be used to answer part c) of the question.

b) Summarize your data reduction process, with **justification** through model evaluation.

c) Using your final classification model in part a), demonstrate, with explanation, how records `TEXT0001` and `TEXT1000` are classified by the model.

Question 2 Association analysis [35 marks]

Table 1 shows a summary of the weather data of Hong Kong from Sep 1, 2022 to Sep 12, 2022. Attributes ***Pressure*** and ***Temperature*** are rounded and grouped into consecutive ranges. ***Rainfall*** is binarized to "*Yes*" / "*No*".

Date (Record ID)	Pressure	Temperature	Rainfall
1	1008..1012	29..29	Yes
2	1003..1007	30..30	No
3	1003..1007	30..30	No
4	1003..1007	31..31	No
5	1003..1007	31..31	No
6	1008..1012	31..31	No
7	1013..1017	28..28	Yes
8	1013..1017	30..30	Yes
9	1013..1017	30..30	No
10	1008..1012	29..29	Yes
11	1008..1012	29..29	No
12	1003..1007	31..31	No

Table 1 *Weather data for association rule mining*

Perform Quantitative Association Rule (QAR) mining by following the steps below. Show your steps.

a) For attributes ***Pressure*** and ***Temperature***, merge the intervals with low support counts with maxsup = 7 (maximum support after merging is at most 7). Consider merging intervals with the lowest support count first.
b) With the merged intervals, transform the dataset in the same way as illustrated in *p.34 of Chapter 6 notes*. For attribute ***Rainfall***, treat ***Rainfall: Yes*** and ***Rainfall: No*** as two attributes.
c) Use the Apriori algorithm to find all frequent itemsets with a support threshold of 20%.
d) For each frequent itemset, generate rules with confidence of at least 75%.
e) Check if there are rules that could be combined, with confidence of at least 75%.

Question 3 Cluster analysis [15 marks]

*Cluster analysis will be covered near the end of the course

Table 2 shows the location of 6 data objects. Perform hierarchical clustering using **Group Average** as the proximity measure. Use **Manhattan distance** as the distance measure. Show your steps and draw the corresponding dendrogram.

Object	Location
$\boldsymbol{O_1}$	$(0, 3)$
$\boldsymbol{O_2}$	$(9, 33)$
$\boldsymbol{O_3}$	$(21, 9)$
$\boldsymbol{O_4}$	$(21, 15)$
$\boldsymbol{O_5}$	$(33, 15)$
$\boldsymbol{O_6}$	$(42, 24)$

Table 2 *Locations of 6 data object*