IFN645: Case Study 2

Mining from Retail and Media Data Sets

Due date: 26th May 2019; Weighting: 25%

Introduction

The purpose of this assignment is to give you an understanding that data mining methods can be applied to various types of data sets such as record data, transactional data, text data and weblogs, and show you the benefits of applying mining techniques to data domains of any kind. This assignment is divided into four parts: Clustering, Association mining, Text Mining, and Web Mining. Same as case study 1, you can do this assignment either using SAS Enterprise Miner or Python.

Instructions

1. The assignment is due on 26th May. It is a firm deadline.
2. You should submit the assignment via Blackboard Assignment.
3. The datasets required for this assignment can be found on BlackBoard with the file named as **casestudy2-data.zip**. It includes four datasets:
    a. online_shoppers_intention to perform clustering
    b. online-retail to perform association mining
    c. bbc (text-files-to-mine) to perform text mining
    d. web_log_data to perform web mining
4. Submit the **team contract** with Peer Appraisal. Failure to submit this would incur the penalty.
5. The assignment will be **marked in the practical class in week 13**. The tutor will check the code/diagrams, plots and results, along with the assignment report, to assign you marks. The entire team should be present to show the project result to the tutor and answer questions to receive marks. We will ask questions to each student and will assign 15% of total marks as per individual performance.
6. Name the case-study report as **casestudy2.doc**. The word file should include a cover page with the Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Python users can submit the jupyter notebook instead of the word document. Combine this file with your **team contract** and name the compressed file as **casestudy2.zip.** Submit the zip file on **Blackboard (under assessment panel Assignment 2)**.

7. A report should be submitted via online submission answering each question of the case study. There is no need of including an introduction, summary, conclusion or references in the report. The report should just include responses to the questions set in the case-study. Some answers may require screenshots. Use them as needed. You can even include your own table detailing those results outcomes. While you may like to go into extreme detail about, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through. For SAS users, remember to include the final diagram of the project showing all nodes connected in your diagram.
8. This is a group assignment. The team size is three. You can continue the same group as in case study 1. If you have formed a new group after assignment 1, please notify the teaching team. In this case, you need to register your new team at Blackboard. Remember to delete the details of your old team at Blackboard.
9. The group is to be ARRANGED and MANAGED by you. As in real life, the performance of the individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.
10. Of course, the work your group hand in must be your own; no collaboration or borrowing from other groups is permitted.
11. Read the Assessment Policies on Blackboard or QUT Website.

Distribution of Marks (Total: 25 marks)

The data mining models would be examined in the practical class in week 13. Your group should be prepared to show the final diagrams and results panels to your marker. The marker will also ask each of the members related questions to test your understanding of the topic and will assign you the team and individual marks.

In data mining, there is hardly ever a single solution. Also many times, there is no correct or wrong solution. You may find that your project partner may have a different solution like yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

The marks are distributed as follows.

Clustering Pre-processing and K-means analysis (10 marks)

Association Mining and it's data Pre-processing (5 marks)

Text Mining (5 marks)

Web Mining (5 marks)

Part 1: Clustering the Online Retail Data

The “online_shoppers_intention” data set contains 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period. This was done to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset consists of 10 numerical and 8 categorical attributes[1].

Attribute Information: The following six attributes, "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration", represent the number of different types of pages visited by the visitor in a session and the total time spent in each of these page categories.

- The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

- The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.
- The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that was the last in the session.
- The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with the transaction.

- The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date.
- For example, for Valentine’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

The dataset also includes “Month of the Year”, “Operating System”, “Browser”, “Region”, “Traffic Type”, “Visitor Type” as returning or new visitor and a Boolean value indicating whether the date of the visit is “Weekend”.

The last attribute “Revenue” indicates whether the online browsing ended with shopping or it did not end with shopping.

Task: Initially, the company wants to learn the user characteristics in terms of users’ time spent on the website. Could you help the company to understand those characteristics by profiling the customers using the k-means analysis? (Hint: First Clustering Model)

---

[1] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018).

Next, the company wants to know its customers in-depth and would like to include information such as where the users come from and when they access the website. (Hint: Refined Clustering Model)

In summary, the task is to conduct k-mean clustering on this data set and describe the **minimum number of effective clusters**. Answer the followings in relation to this data and analysis.

Task 1. Data Preparation for Clustering.

1. Can you identify data quality issues in this dataset such as unusual data types, missing values, etc?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice. Answer this question for each of the tasks 2 and 3.

Task 2. The first clustering model

1. Build a clustering model to profile the customers based on the time they spend on the website. Answer the followings:
    a. What is the optimal number of clusters?
    b. For the model with the optimal number of clusters, list the variables that were found important in determining the clusters?
    c. Explain the cluster results.
2. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?
3. Interpret the cluster analysis outcome. In other words, characterize the nature of each cluster by giving it a descriptive label and a brief description.

Task 3. Refining the clustering model

1. Add more information such as where the users come from and when they access the website, to the clustering analysis that you have conducted in the previous task. Answer the followings:
    a. What is the optimal number of clusters?
    b. Whether this model has different variable importance than the previous model (Task 2.1)?
    c. Explain the cluster results.
2. What is the effect of using the standardization method on the model above? Does the variable normalization process enable a better clustering solution?
3. Can you highlight the differences between the two clustering solutions (Tasks 2.1 & 3.1) focusing on cluster interpretation? In other words, explain what is the effect of adding other variables in the clustering analysis?

Decision Making: Finally, could you detail - how the outcome of clustering analysis can be used by decision makers?

Part 2: Association Mining the Online Retail Data

A UK-based online retail store is interested in determining the associations between gift items sold online. They have collected the transactions of over 4200 products made over a year. The transactional dataset contains transactions (~500K) occurring between 01/12/2010 and 09/12/2011 for an online retail company that mainly sells unique all-occasion gifts. There are a total of 4207 unique products represented in the data set. Many customers of the company are wholesalers[2].

The online_retail data description for Association Mining:

Name	Description
InvoiceNo	Invoice number, a 6-digit integral number uniquely assigned to each transaction. Unique transaction identification number for a given sale. A sale may include several products and thus the same transaction id may occur over several rows.
StockCode	Product (item) code, a 5-digit integral number uniquely assigned to each distinct product.
InvoiceDate	Invoice Date and time, the day and time when each transaction was generated.
Description	Product (item) name
Quantity	The quantities of each product (item) per transaction
UnitPrice	Unit price. Product price per unit in sterling
CustomerID	Customer number, a 5-digit integral number uniquely assigned to each customer.
Country	Country name, the name of the country where each customer resides

The **task** is to conduct Association analysis on this data set. Answer the followings in relation to this data and analysis.

Task 4. Association Mining

1. Can you identify data quality issues in this dataset for performing association analysis?
2. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
3. Conduct association mining and answer the following:
    a. What is the highest lift value for the resulting rules? Which rule has this value?

[2] Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

   b. What is the highest confidence value for the resulting rules? Which rule has this value?
   c. Discuss and explain the results. Interpret them to discuss the rule-set obtained.
4. The store is particularly interested in products that individuals purchase when they buy “HERB MARKER CHIVES”.
   a. How many rules are in the subset?
   b. Based on the rules, what are the other products these individuals are most likely to purchase?
5. Can you perform sequence analysis on this dataset? If yes, present your results. If not, rationalise why?
6. How the outcome of this study can be used by decision makers?

Part 3: Text Mining (Clustering) the News Stories

Task 5 Text Mining

A leading news corporation is planning to start an online personalised news story service. They have a collection of individual stories in the form of a compressed single file (text-files-to-mie.zip). Perform text mining on this dataset to determine clusters of stories based on similar topics.

Answer the followings in relation to this data and analysis.

1. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
2. Can you identify data quality issues in order to perform text mining?
3. Based on the ZIPF plot, list the top 10 terms that will be least useful for clustering purpose.
4. Did you disregard any frequent terms in addition to items automatically selected by the Filter tool?
5. Justify the term weighting option selected.
6. What is the number of input features available to execute clustering? (Hint: how the original text data is converted into a feature set that can be mined for knowledge discovery?)
7. State how many clusters are generated? Name each cluster meaningfully according to the terms that appear in the clusters?
8. Identify the first fifteen high frequent terms (that are not stopwords or noise) in the start list?
9. Describe how these clusters can be useful in the online personalised news story service planned by the news corporation.

Part 4: Web Mining the Log Data for a Website

Task 6 Web Mining

For an e-commerce business, the website structure and site plan were established with the efficiency and usability in mind, but its effectiveness had not been verified. Only basic statistics have been produced through simple report and query techniques; however, they provide no means for sophisticated web site analysis and predictions. Your task is to determine the patterns of user-browsing the website and analyse those patterns to provide the results and recommendations to the website owner.

You have been provided with a log file in CSV format (web_log_data). This was originally a text file and was processed with the steps required for web usage mining as explained in the lecture. The processing steps were: (1) removing unproductive items from the log file such as graphics, sound, etc; and (2) identifying users and sessions based on IP address, date and time. The goal of user session identification is to divide the page access of each user into individual sessions.

The dataset consists of the IP address, timestamp, request, step, session id, and user id. Your task is to **apply two data mining operations**, such as classification or clustering or association mining/sequence mining, to the pre-processed data set. Answer the followings in relation to this data and the analyses that you have chosen.

1. For each data mining operation:
    a. The rationale for selecting the specific operation/method.
    b. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
    c. Can you identify data quality issues in order to perform web mining?
    d. Discuss the results obtained. Discuss also the applicability of findings of the method. Should include a high-level managerial kind of discussion on the findings, should not be just interpretation of results as shown in the results panel.

Assignment Criteria Sheet:

Criteria	Score
Non Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components. Questions (both written and orally) were poorly answered.	1-5
Has demonstrated a task with a working model having a data source, and diagram with substantial but incorrect implementation of at least one of the seven components. Questions (both written and orally) were poorly answered.	6-11
Has implemented all tasks with at least two being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	12
Has implemented all four parts: One mining task is fundamentally correct, with substantially correct workflow diagrams which may contain minor errors. Response to questions shows a fundamental understanding of terms and concepts.	13-15
Has fundamentally correct implementation of all four parts i.e. selection of correct variables in all four data, correct allocations, understanding, and explanation of clusters, findings association rules, in all applications. Demonstrate competent application of tools. Many questions have been reasonably answered. Demonstrate a good understanding of the methods and terms used in clustering and association mining, during written and verbal analyses. Some minor errors are allowed. A written report is required to be of a reasonable standard. Response to questions shows basic knowledge on the topic.	16-18
Has implemented all of the requirements above with very few errors. A strong focus on the application of tools and evaluation and interpretation of results is evident. Response to questions shows in-depth knowledge on the topic.	19-21
All of the criteria above are met, extensive model generation and analysis have been conducted to produce exceptional outcomes and have applied principles learned in lectures to enhance the results. Response to questions shows extensive knowledge on the topic.	22-25