IFN645 Large Scale Data Mining

Semester 2, 2021

Major Assignment

Due date: Monday, 25 October 2021, 11:59pm
Weighting: 40%
Team/Individual: You can work on this assignment individually or in a team of 2.

Required to be submitted on Blackboard:

One single zip file which contains the following files/folders:

1. A report (pdf file) containing your answers to the questions specified below and a statement of completeness stating which tasks have been completed.
2. Three separate folders each containing one Java project for one of the tasks.

Introduction

This assignment is intended to allow you to display your knowledge and understanding developed in lectures and practicals. The purpose of this assignment is to give you (1) an understanding that various methods can be applied to different types of datasets such as text and transactions, and (2) the benefits of applying data analytics techniques to a data domain.

There are three questions in this assignment. You need to complete all the three questions.

Task Specification

1. **Task 1: Association mining in Java (13 Marks)**
   A bank has conducted a marketing campaign via phone calls to promote their new products including a term-deposit product. After the campaign, the bank wants to analyse the data collected in the campaign in order to get a better understanding to their customers.

   - Dataset
     Download the datasets **`bank.arff`**, **`bank_no.arff`**, and **`bank_yes.arff`** from the Blackboard.
     `bank.arff` is the entire dataset collected in the campaign which consists of 45,211 records. The other two datasets are subsets of the entire dataset containing customers in each class. Each record in the datasets is about one customer described by 11 attributes. To help you understand the data, the following table provides you with the description to each of the attributes.

Attribute	Description
age	Customer age
job	Customer job type
marital	Customer marital status
education	Customer education status
default_credit	Whether customer has credit in default or not

balance	Customer bank account balance
housing	Whether customer has housing/home loan or not
loan	Whether customer has personal loan or not
call_duration	Phone contact duration in seconds, e.g., 500-1k indicates that the phone talk with the customer lasts 500 to 1000 seconds.
past_marketing	Outcome of last marketing to this customer, e.g., failure, indicates that the last marketing was unsuccessful to this customer.
subscribed	After this marketing campaign, whether the customer subscribed the term-deposit product or not. This attribute divides customers into two classes, i.e., yes-class and no-class.

In the context of this task, an item is an attribute with a specific value, e.g., age=40s, and a pattern is a set of items, e.g., {default_credit=no, loan=no, age=21-30s} is a size-3 pattern. The patterns generated from the bank dataset describe the characteristics or behaviour of the customers in that dataset.

After the marketing campaign, the bank wants to obtain some information from the dataset. Specifically, they want to know the popular patterns in each customer class and also want to know the associations between patterns and each class, i.e., the association rules that represent the implications from customer attributes to one of the classes. As a data analyst, you are required to develop a Java program that can generate patterns and association rules from the datasets to answer the following questions.

- Questions

  1) Generate frequent patterns from the entire dataset (i.e., **bank.arff**) using two frequent pattern mining algorithms and compare their performance in terms of time efficiency. You may use 3 different minimum supports to do the comparison. Show your comparison result and choose the algorithm with better performance.

  2) Using the chosen algorithm in question_1), generate the top 5 most frequent size-3 patterns from the yes-class dataset (i.e., **bank_yes.arff**) and no-class dataset (i.e., **bank_no.arff**) separately, then compare the generated patterns from the two datasets and identify the same or different characteristics between the customers in the two classes.

  3) Generate the top 5 most frequent maximum patterns from yes-class and no class datasets separately, identify any similarity or differences between the two classes in terms of the maximum patterns.

  4) Use three algorithms to generate frequent closed patterns from the entire dataset and compare their time efficiency.

5) Using the entire dataset, generate the ~~top 10 most frequent~~ association rules with subscribed=yes as the consequent and also the top 10 most frequent association rules with subscribed=no as the consequent. You can specify some appropriate minimum support and minimum confidence.
    a. List the rules generated for each class.
    b. Observe the rules and identify any redundant rules in each set of the rules. If there exist redundant rules, list them and state why you think they are redundant.

2. Task 2: Classification in Weka and Java (13 marks)

In Task 1, you have developed a Java program to generate patterns and association rules from the bank datasets to describe the behaviour of the customers in the dataset. In this task, you are required to write a Java program to classify the customers in the entire bank dataset **`bank.arff`**. Before writing your Java program, you are required to analyse the dataset in Weka to select 3 classification algorithms based on their classification performance in Weka. For evaluation, you can use the default 10-fold cross-validation.

2.1. Data analysis in Weka

For the following questions, you need to

- Use the **`AttributeSelectedClassifier`** in Weka to analyse the data.
- Describe your working process.
- Provide evidence to justify your decision. The evidence can be tables to show performance comparison, screenshots, or some outputs from Weka.

1) Select 4 classification algorithms based on attribute analysis. You can choose the 4 algorithms from **`NaiveBayes`**, **`NaiveBayesMultinomial`**, **`IBk`**, **`PART`**, **`OneR`**, **`ZerR`**, **`J48`**, **`Randomforest`**. This question is not to choose attributes. This question ~~is to choose algorithms~~ based on their classification accuracy performance. You can use any evaluator, search method, or ranker to do it. Justify your decision.
2) Select 3 algorithms from the 4 algorithms chosen in question 1) of Section 2.1 based on cost analysis. You can assume that class “subscribed = y” is more important to you and you want to minimize the classification error to this class. Justify your decision.
3) For each of the 3 algorithms chosen in question 2) of Section 2.1, determine the number of attributes to select in order to achieve a relatively better classification performance by using that algorithm. Justify your decision.
   There are 11 attributes in this dataset. For saving your time, you may choose a number between 4 to 8. This question is to determine the **number**, not to choose the attributes.

2.2. Java program

For this part, you are required to develop a Java program to classify customers in the entire bank dataset. Your program should satisfy the following requirements:

1) Perform the classification task using the 4 algorithms chosen in question_1) of Section 2.1 by taking cost into consideration. Your program should display classification accuracy (i.e., correctly classified instances) and total cost for each classifier. Basically, your program should produce the same results as the results obtained in question_2) of Section 2.1 using Weka.
   Hint:You can use Weka class **`CostSensitiveClassifier`** (https://weka.sourceforge.io/doc.dev/weka/classifiers/meta/CostSensitiveClassifier.html) to do the classification.

2) Perform the classification task using the 3 algorithms chosen in question_2) of Section 2.1 with the number of attributes determined in question-3) of Section 2.1 for each of the three algorithms. Your program should display correctly classified instances values and accuracy. The accuracy results should be the same as the results obtained in question_3) of Section 2.1 using Weka.

3. Task 3: Text classification in Weka and Java (12 marks)

Download dataset **`News.arff`**. This is a text dataset consisting of news documents. These news documents are categorised into four classes: computer, politics, science, and sports. In this task, you are required first to classify the news documents using Weka to determine some parameters in the filter, then develop a Java program to classify the news in this dataset.

3.1. Attribute selection in Weka

In this part, you need to use a filter in Weka to extract attributes from the documents. You can choose 100 attributes and use J48 classifier to do the classification. For the parameters in the filter, you can use their default values or you can set up some values of your choice without tuning them (i.e., you can randomly choose some values). But you are required to tune 3 or 4 parameters that you think are important for determining attributes to represent the documents.

1) Briefly describe your working process in Weka to determine the values for the parameters in the filter. Provide evidence to show your working.
2) Which parameters in the filter that you want to tune? What are the chosen values for these parameters? Justify your decision with evidence.

3.2. Java program

For this part, you are required to develop a Java program to classify the documents in the news dataset.

1) Perform the classification task using 4 classification algorithms, **`IBk`**, **`SMO`**, **`J48,`** and the method **`HoeffdingTree`** in Weka, and use the filter with the parameter settings determined in section 3.1.
2) Your program should display the correctly classified instances results, accuracy, and the time taken by each algorithm.
3) Which classifier performs the best in terms of time efficiency? Describe why this algorithm is faster than others.

Submission Requirements

- Your report should just include responses to the questions described above. It should be structured well and easy to read. Your responses should be well justified by evidence.
- Your Java programs should be commented appropriately and should minimize repeated code using a refactoring manner. Should display outputs.
- Your work must be your own or your team of 2 students.
- Submit one single zip file which contains your report and three Java projects.
- If you are in a team of 2, each team member should submit an identical submission. One of the submissions will be marked. Your submission file name should be: **student1ID_student2ID.zip**.
- If you complete the assignment individually, your submission file name should be: **studentID.zip**.

General Marking Scheme

Task 1 (13 marks)	
Questions	7
Java program Correctly accomplish the required tasks using the right algorithms Produce the required outputs Coding quality	6
Task 2 (13 marks)	
Data analysis in Weka	7
Clearly describe the approach for generating answers for each question Provide evidence to support your decision	
Java program	6
Correctly accomplish the required tasks using the right algorithms Produce the required outputs Coding quality	
Task 3 (12 marks)	
Attribute selection in Weka	6
Clearly describe your approach for generating answers for each question Provide evidence to support your decision	
Java program	6
Correctly accomplish the required tasks using the right algorithms Produce the required outputs Coding quality	
Report presentation (2 marks)	