Data Analytics II Report (SET09120)

Valeri Vladimirov

School of Computing, Edinburgh Napier University, Edinburgh

40399682@napier.live.ac.uk

1 Introduction

This assignment's purpose is to prepare, clean and analyse a provided dataset with tools, such as OpenRefine and Weka. OpenRefine will be used to clean the data by fixing any mistakes found and also to transform the data from a xlsx format to an arff format. Weka will be used to analyse the data using the Classification, Association and Clustering techniques. The patterns found by the algorithms will be represented by 6 rules each. Additionally, the overall findings from these algorithms will be discussed and compared.

2 Data Preparation

Data preparation is the process of cleaning a dataset and converting it to an appropriate format. In this section I am going to explain how I have cleaned the data, what corrections I have made and why, using OpenRefine. Also, I will discuss how I converted the dataset to an .arff format and why I changed some of the numeric values to nominal in another .arff file. This was done using both OpenRefine and Weka for changing the file formats

2.1 Data Cleaning

The dataset required a few corrections, which were made using OpenRefine.
Apart from renaming the columns, all other changes were made using the numeric or text facet functionalities of OpenRefine:

- Firstly, I renamed all of the column names to the appropriate attribute names listed in the coursework specification. To ensure that the first row is not lost I added an additional row at the top of the dataset, which I then renamed. By doing this I kept all of the 1000 rows.
- I did some corrections to the purpose column, which included fixing misspelled words, wrong capital letters and unneeded single quotation marks. The changes can be seen in Table 1.
- After that I also decided to remove all other quotation marks, because some columns such as purpose and class did not have any. Also, jobs such as skilled did not have whereas the other in the column had. This was done so that columns have a similar syntax and so that the documents would look cleaner and nicer.

- Corrections were made to the credit amount column, where before making any changes I did a common transfer to number, just in case. After that I found that some values had more zero's or one's at the end or start of the number, which were removed. Changes can be seen in Table 2.
- The age column also had mistakes, which included numbers that are impossible to be an age, such as 222. Negative numbers were changed to positive. Decimal numbers were corrected. And there were values for age such as 1 and 6 which seemed impossible to have anything to do with credit, so they were increased to 21 and 26. The changes can be seen in Table 3.
- Lastly the job column had 1 correction made. I assumed that the value 'yes' refers to 'skilled', therefore I transformed it. Can be seen in Table 4.

2.2 Data Conversion

After cleaning the dataset, the next step, is to convert it to an appropriate format that will be accepted in Weka, and that is .arff. In order to do so, the following was done:

- From OpenRefine the dataset was exported as a .xlsx file.
- The .xlsx file was opened in Excel and from there saved as a .csv (Comma-Separated).
- The .csv file was opened in Weka and from there saved as an .arff, bearing in mind that the .csv extension had to be removed when naming the file.

Arff files start with a @relation, which is the name of the dataset, there is also @attribute, which represent the columns and @data from where all the data starts. This can be seen when the file is loaded into Word.

Another arff file had to be created in which some of the columns with numeric data had to be transformed to nominal data. This would be useful when using the Apriori algorithm with Association. The two columns were age and credit_amount, although Case_No is also numeric it was not changed, because it can be removed later. To transform data from numeric to nominal I used the transform functionality when editing cells by providing if/else statements written in Python (see Fig. 7). The credit_amount was split into groups of increments of 2000, which can be seen in Figure 5. The age was split into increments of 10, although the first group is of 11. This can be seen in Figure 6. After transforming from numeric to nominal the dataset was again converted to an .arff file as explained above.

3 Data Analytics

In this section I will discuss the algorithms used for finding patterns withing the dataset. Three out of the four studied techniques will be used, and 6 rules will be provided per technique. The algorithms used for this dataset are Classification, Association and Clustering.

3.1 Classification

Classification is used to predict the outcome of an experiment with a nominal target class attribute. This technique was one of the reasons I had to create another .arff file by changing some of the numeric values to nominal. For classification I have decided to use the J48 algorithm, which generates a decision tree and would allow the prediction of the target variable of a new dataset record.

In order to classify the dataset, I have used Weka by following these steps:

- In Weka after opening the Explorer I opened the second version of the dataset with mainly nominal values. This was decided, because the nominal values dataset produced a higher percent of correctly classified instances than the numeric with 0.7%.
- After loading the file, I went to the Classify tab in the upper left corner of Weka and selected the J48 classifier under the trees folder from the Classifier section.
- I ensured that the training set test option was selected and that also (Nom) Class was chosen.
- After that I selected start which began the calculations.

After classifying the dataset there are 3 main things to look at in the output screen. The first is under Summary, the correctly (79.7%) and incorrectly (20.3%) classified instances, which can be seen in Figure 8. From there it can be understood that the model produced was a relatively good one. Secondly the confusion matrix can be seen at the bottom of the screen and in Figure 9. It shows the performance of the model. From there it was seen that there were 650 true positive and 147 true negative values, which means these values were predicted to be true or false and they actually were. It can also be seen that there were 50 false positive and 153 false negative values, which means they were predicted to be true or false, but they were not. Finally, the J48 pruned tree can be seen, which shows interesting patterns that can be analysed.

Next to each node of the three there are two numbers. The first one is the total number of cases and the second one is the wrongly classified data for this node. When selecting the best six rules I have taken two factors into consideration. The rules should have a high number of cases (coverage) and a high accuracy. The accuracy is calculated by subtracting the wrongly classified instances from the total number of cases and then diving the result by the total number of cases. Also, the rules have been selected from various branches. Here are the six rules chosen:

Rule 1:

IF Checking_Status = no checking THEN good

This rule (394.0 / 46.0) has 394 cases and 46 wrongly classified values. It was selected because it has a high coverage of 394 and a high accuracy of 88.3%. This means that customers with no current account in the bank are likely to get the loan they have requested.

Rule 2:

IF Checking_Status = <0 AND Credit_History = critical/other existing credit

THEN good

This rule (67.0 / 18.0) has 67 cases and 18 of them are wrongly classified. It was chosen because although the number of cases is relatively small for the dataset, it is very high compared to the number of cases in other branches. It has a coverage of 67 and accuracy of 73%. It shows that clients with checking status less than 0 and other debts existing but not at this bank are likely to receive a loan.

Rule 3:

IF Checking_Status = 0<=X<200 AND Credit_Amount = 0<=X<2000

AND Purpose = radio/tv THEN good

This rule (40.0 / 9.0) has 40 cases and 9 wrongly classified. It was selected because it has a coverage of 40, which is relatively higher than others and has high accuracy 77.5%. This means that customers, which would like to get a loan for a radio/tv and have a checking status between 0 and 200 are likely to receive it.

Rule 4:

IF Checking_Status = 0<=X<200 AND Credit_Amount = 2000<=X<4000

THEN good

This rule (77.0 / 21.0) has 77 cases and 21 of them are wrongly classified. It was chosen because it has a high coverage of 77 and a relatively high accuracy of 72.7%. This means that clients with a checking status between 0 and 200 and credit amount between 2000 and 4000 are likely to receive their loan.

Rule 5:

IF Checking_Status = <0 AND Credit_History = existing paid AND

Purpose = new car THEN bad

This rule (42.0 / 15.0) has 42 cases and 15 of them are wrongly classified. It was selected because it has 42 coverage which is relatively higher than others and 64% accuracy. This means that customers that have paid existing debts back duly till now, have a checking status of less than 0 and the purpose for their loan is a new car are not likely to receive it.

Rule 6:

IF Checking_Status = >= 200 THEN good

This rule (63.0 / 14.0) has 63 cases and 14 of them wrongly classified. It was chosen because it has a coverage of 63, which is good and quite a high accuracy of 77.7%. This means that customers with a checking status of greater than or equal to 200 are likely to receive their loan.

3.2 Regression

Regression will not be used for the current dataset, because it was decided that this technique is the least powerful of the 4 data mining algorithms.

3.3 Association

Association is a data mining technique which finds frequent patterns in the data. The Apriori algorithm will be used since it is a simple algorithm and was suggested in the lectures. It is a breadth first search method and is used to find frequent item sets in a transaction database by exploring first the item sets of the same size. In order to use association, the following should be done:

- As in classification, the Weka Explorer should be opened, and the second version of the dataset with nominal values should be loaded. Since Association, more precisely the Apriori algorithm, works only with nominal values the Case_No attribute should be removed from the dataset in the Preprocess tab. This is done by selecting the attribute and clicking on the remove button.
- Then I went to the Associate tab, choose the Apriori algorithm in Associator. After choosing, I clicked on it and went to the numRules field and typed 6. This is done because the coursework assignment requires only 6 rules per technique. And now the 6 best rules will be selected. Finally, I clicked Start in order to begin the process.

From the output the six best rules can be seen (see Fig. 10), which all have a confidence of above 0.9 and a high coverage. The coverage of a rule can be seen on the left side of the "==>" arrow, on the right is how many can be classified as good. The accuracy, also known as confidence, can be seen in the angle brackets, next to "conf:".

Rule 1:

IF Checking_Status=no checking AND Purpose=radio/tv 127

THEN Class=good 120

<conf:(0.94)> lift:(1.35) lev:(0.03) [31] conv:(4.76)

This rule has an accuracy of 94% and a coverage of 127. It shows that from 127 customers that do not have an account in the bank and the purpose for their loan is a radio or tv, 120 of them are likely to receive their loan.

Rule 2:

IF Checking_Status=no checking AND Credit_History=critical/other existing credit

153 THEN Class=good 143

<conf:(0.93)> lift:(1.34) lev:(0.04) [35] conv:(4.17)

This rule has an accuracy of 93% and a coverage of 153. It shows that from 153 customers that do not have an account in the bank and have other debts existing (but not at this bank), 143 of them are likely to receive their loan.

Rule 3:

IF Checking_Status=no checking AND Employment = >=7 115 THEN

Class=good 107

<conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.83)

This rule has an accuracy of 93% and a coverage of 115. It can be seen that 107 out of 115 customers that do not have an account in the bank and have been employed for 7 or more years are likely to receive a loan.

Rule 4:

IF Checking_Status=no checking AND Personal_Status=male single

AND Job=skilled 151 THEN Class=good 139

<conf:(0.92)> lift:(1.32) lev:(0.03) [33] conv:(3.48)

This rule has an accuracy of 92% and a coverage of 151. It can be seen that 139 out of 151 customers that do not have an account in the bank and are skilled single male employees are likely to receive a loan.

Rule 5:

IF Checking_Status=no checking AND Credit_Amount=2000<=X<4000 131

THEN Class=good 120

<conf:(0.92)> lift:(1.31) lev:(0.03) [28] conv:(3.27)

This rule has an accuracy of 92% and a coverage of 131. 120 out of the 131 customers that do not have an account in the bank and have a credit amount of between 2000 and 4000 are likely to receive the loan they have requested.

Rule 6:

IF Checking_Status=no checking AND Credit_Amount=0<=X<2000

AND Job=skilled 116 THEN Class=good 106

<conf:(0.91)> lift:(1.31) lev:(0.02) [24] conv:(3.16)

This rule has an accuracy of 91% and a coverage of 116. 106 out of the 116 customers that do not have an account in the bank and have a credit amount of between 0 and 2000 and are classified as skilled employees are likely to receive the loan they have requested.

3.4 Clustering

Clustering is a data mining technique which groups objects which are similar to each other into the same clusters. For the current task, the k-Means algorithm will be used. In order to perform clustering the following should be done:

- After opening the Weka explorer, the first version of the dataset should be loaded and the Case_No attribute should be removed by selecting it and clicking the Remove button.
- Then you should go to the Cluster tab and choose a clusterer, which should be SimpleKMeans. Then by clicking on the method a new window will be shown and next to numClusters 6 should be written. This means that it will produce 6 clusters, which is perfect for the current case, because we need 6 rules. After that is done, Start should be clicked, and the output will be seen.

After clustering the dataset at the bottom of the output (see Fig. 11) the number of objects in each cluster can be seen and their percentage from the total amount of instances in the dataset (1000), which can also be treated as the coverage. Above the Clustered Instances the actual starting points can be seen (see Fig. 12). In Figure 13 the 6 clusters can be seen starting from 0 to 5 and each of them can be treated as a rule. An example rule (Cluster 3) would look like:

- If the customer is a skilled male single and has no current account in the bank, all existing debts are paid, with a purpose for the loan a radio/tv, a mean credit amount of around 3179, without any known savings, a work experience of 7 years or above and a mean age of around 36 years old then he will likely receive his loan. This rule has a coverage of 214.

All of the other rules can be seen in the table in Figure 13. After analysing the output, the following conclusions can be made.

- The only cluster classified as bad is the 4^{th} one. It shows that skilled and single male customers with a checking status of less than 0, existing debts paid, purpose for the loan a new car, a mean credit amount of around 4543, less than 100 savings, employed for 7 or more years and a mean age of 36 aren't likely to receive their loan.
- In all of the clusters the savings are either less than 100 or none.
- The Job type hasn't had any effect on whether the customer will receive the loan since no matter if they are skilled, unskilled, or highly qualified they have all been classified as good (apart from cluster 4).
- Also, the years of employment don't matter, since no matter if the customer is unemployed or has work experience from 1 year to 7 or more, they are all classified as good (apart from cluster 4).
- It can be seen that for all of the clusters the mean credit amount is in the range of 2000 to 5000.

4. Summary

After analysing the results from the 3 algorithms (Larry Alton 2017) there are some key conclusions that could be made. The first one is that from the 18 rules, 9 of them in which customers have no current account in the bank are classified as good. This would mean that the bank will give out loans to these people in order to attract them to the bank, so they can create an account. Also, all 5 people who have applied for a loan for a radio/tv are likely to receive their loan, probably because the loan wouldn't be of such a great amount as compared to a new car. The only 2 badly classified rules are the same (one from clustering and one from classification). They state that if the customer wants a loan for a new car and the status of the current existing account is less than 0 and the existing debts are paid they will likely not receive their desired loan. They would probably have been classified as good if the customer didn't have an existing account in the bank as in Cluster 0 in Figure 13, since then the bank would want to attract the customer and give him a loan.

Overall, from the 3 algorithms used and 18 rules created only 2 of them were classified as bad, meaning that the results from the data mining were mostly good and more precisely the algorithms used were efficient. Personally, I think that Classification (TutorialsPoint) could be considered the most efficient technique, because it is predictive compared to Clustering (GeeksForGeeks 2019) and Association (Quora), which are both descriptive. Although Association is the technique which has the rules with the highest accuracy (confidence) and none badly classified rules, classification is preferred, because it shows a lot more rules, which are viewed as a decision tree. Classification gives a lot more variety when selecting the rules thanks to the decision tree. All of the rules can be seen even those with 0 accuracy and coverage, showing all of the possible branches, whereas in Clustering and Association the exact number of rules (clusters) should be selected. Clustering is the technique which in my opinion was the least effective for making discoveries and gaining insights into the data, because it gave the least accurate results.

To sum up, for me the most useful technique was Classification providing a lot of variety. Association was also useful, because of the very accurate results and no bad rules. Clustering was decided to be the least effective and least accurate technique from those used. Regression wasn't even used because it was decided beforehand that it would be the least useful. However, that is only for the current dataset. Which data mining techniques to be used depends on the data and in another dataset they could be different to those used here.

5. Appendix

Purpose	Frequency	Correction	Explanation
ather	1	other	misspelled
busines	3	business	misspelled
busness	3	business	misspelled

Eduction	1	education	misspelled and wrong capital letter
Radio/Tv	2	radio/tv	wrong capital letters
‘new car’	234	new car	unneeded single quotation marks
‘used car’	103	used car	unneeded single quotation marks
‘domestic appliance’	12	domestic appliance	unneeded single quotation marks

Table. 1. Showing the corrections made for the purpose column.

Credit Amount	Correction	Explanation
111328000	1328	Removed 0’s and 1’s
19280000	1928	Removed 0’s
13580000	1358	Removed 0’s
13860000	1386	Removed 0’s
63610000	6361	Removed 0’s
7190000	719	Removed 0’s
5180000	518	Removed 0’s
5850000	585	Removed 0’s

Table. 2. Showing the corrections made for the credit amount column.

Age	Frequency	Correction	Explanation
222	2	22	Impossible for age (big)
333	1	33	Impossible for age (big)
-34	1	34	Cannot have negative age
-35	1	35	Cannot have negative age
-29	1	29	Cannot have negative age
0.44	1	44	Cannot have decimal age
0.24	1	24	Cannot have decimal age
0.35	1	35	Cannot have decimal age
1	1	21	Impossible for age (small)
6	1	26	Impossible for age (small)

Table. 3. Showing the corrections made for the age column.

Job	Frequency	Correction	Explanation

yes	2	skilled	Cannot have ‘yes’ as a job, so it was assumed it is ‘skilled’

Table. 4. Showing the corrections made for the job column.

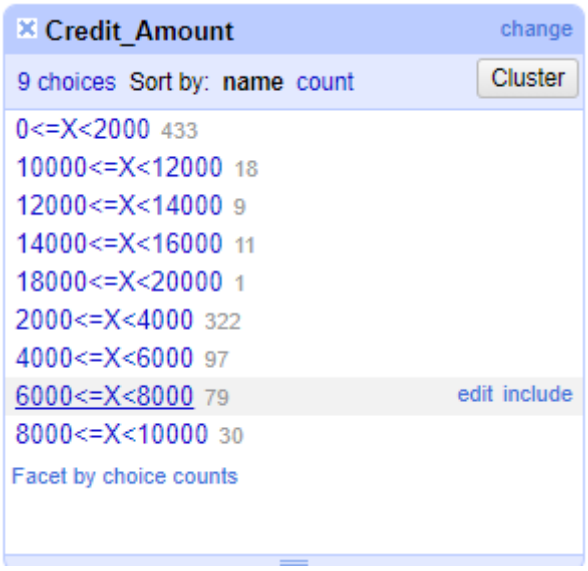

Figure. 5. Showing the change of numeric to nominal data for Credit_Amount.

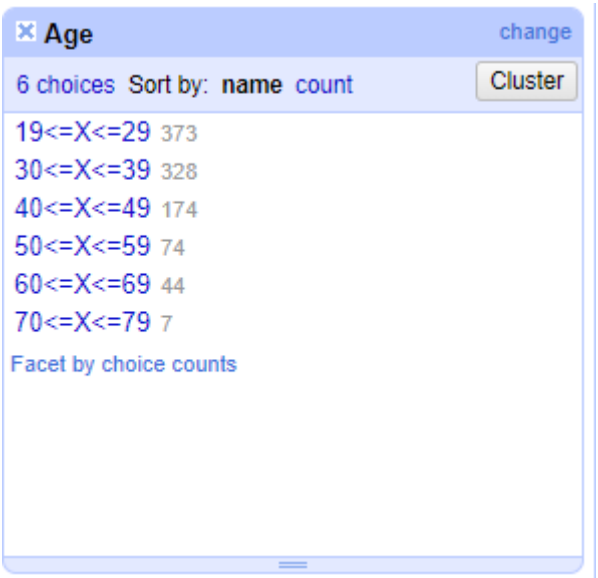

Figure. 6. Showing the change of numeric to nominal data for age.

```
# Code for the age
if value >= 19 and value <= 29:
 return "19<=X<=29"
elif value >= 30 and value <= 39:
 return "30<=X<=39"
elif value >= 40 and value <= 49:
 return "40<=X<=49"
elif value >= 50 and value <= 59:
 return "50<=X<=59"
elif value >= 60 and value <= 69:
 return "60<=X<=69"
elif value >= 70 and value <= 79:
 return "70<=X<=79"

 # Code for the credit amount
 if value >= 0 and value < 2000:
 return "0<=X<2000"
elif value >= 2000 and value < 4000:
 return "2000<=X<4000"
elif value >= 4000 and value < 6000:
 return "4000<=X<6000"
elif value >= 6000 and value < 8000:
 return "6000<=X<8000"
elif value >= 8000 and value < 10000:
 return "8000<=X<10000"
elif value >= 10000 and value < 12000:
 return "10000<=X<12000"
elif value >= 12000 and value < 14000:
 return "12000<=X<14000"
elif value >= 14000 and value < 16000:
 return "14000<=X<16000"
elif value >= 16000 and value < 18000:
 return "16000<=X<18000"
elif value >= 18000 and value < 20000:
 return "18000<=X<20000"
```

Figure. 7. Showing the code snippet for transforming the data.

```
=== Summary ===

Correctly Classified Instances         797               79.7    %
Incorrectly Classified Instances       203               20.3    %
```

Figure. 8. Showing the percentage of correctly and incorrectly classified instances.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 650  50 |   a = good
 153 147 |   b = bad
```

Figure. 9. Showing the confusion matrix of the model.

```
Best rules found:

1. Checking_Status=no checking Purpose=radio/tv 127 ==> Class=good 120    <conf:(0.94)> lift:(1.35) lev:(0.03) [31] conv:(4.76)
2. Checking_Status=no checking Credit_History=critical/other existing credit 153 ==> Class=good 143    <conf:(0.93)> lift:(1.34) lev:(0.04) [35] conv:(4.17)
3. Checking_Status=no checking Employment=>=7 115 ==> Class=good 107    <conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.83)
4. Checking_Status=no checking Personal_Status=male single Job=skilled 151 ==> Class=good 139    <conf:(0.92)> lift:(1.32) lev:(0.03) [33] conv:(3.48)
5. Checking_Status=no checking Credit_Amount=2000<=X<4000 131 ==> Class=good 120    <conf:(0.92)> lift:(1.31) lev:(0.03) [28] conv:(3.27)
6. Checking_Status=no checking Credit_Amount=0<=X<2000 Job=skilled 116 ==> Class=good 106    <conf:(0.91)> lift:(1.31) lev:(0.02) [24] conv:(3.16)
```

Figure. 10. Showing the 6 best rules for Association

```
Clustered Instances

0       174 ( 17%)
1        81 (  8%)
2       111 ( 11%)
3       214 ( 21%)
4       183 ( 18%)
5       237 ( 24%)
```

Figure. 11. Showing the number of objects in each cluster and their percentage.

```
Cluster 0: 'no checking','critical/other existing credit','new car',7855,<100,1<=X<4,'female div/dep/mar',25,skilled,bad
Cluster 1: <0,'critical/other existing credit','used car',6615,<100,unemployed,'male single',75,'high qualif/self emp/mgmt',good
Cluster 2: 0<=X<200,'existing paid',radio/tv,1155,<100,>=7,'male mar/wid',40,'unskilled resident',good
Cluster 3: 0<=X<200,'existing paid',retraining,754,'no known savings',>=7,'male single',38,skilled,good
Cluster 4: <0,'critical/other existing credit','new car',3966,<100,>=7,'female div/dep/mar',33,skilled,bad
Cluster 5: 0<=X<200,'existing paid',radio/tv,753,<100,1<=X<4,'female div/dep/mar',64,skilled,good
```

Figure. 12. Showing the 6 initial starting points.

#	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Checking_Status	no checking	<0	0<=X<200	no checking	<0	0<=X<200
Credit_History	critical/other existing credit	critical/other existing credit	existing paid	existing paid	existing paid	existing paid
Purpose	new car	used car	radio/tv	radio/tv	new car	radio/tv
Credit_Amount	2894.046	4889.5926	2098.2252	3179.9813	4543.7049	2602.1941
Saving_Status	<100	<100	<100	no known savings	<100	<100
Employment	1<=X<4	unemployed	>=7	>=7	>=7	1<=X<4
Personal_Status	Female div/dep/mar	male single	male single	male single	Male single	Female div/dep/mar
Age	34.3046	43.9012	36.2973	40.5327	36.0109	28.3038
Job	skilled	high qualif/self emp/mgmt	unskilled resident	skilled	skilled	skilled
Class	good	good	good	good	bad	good

Figure. 13. Showing the 6 clusters.

6. References

Larry Alton 2017, *Most important Data Mining Techniques:* https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

Quora, *Clustering and Association difference:* https://www.quora.com/What-is-the-difference-between-clustering-and-association-rule-mining

GeeksForGeeks 2019,*Classification and Clustering difference:* https://www.geeksforgeeks.org/ml-classification-vs-clustering/

TutorialsPoint, *What is Classification:* https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm