

Coursework Report

Maria Luque Anguita

40280156@live.napier.ac.uk

Edinburgh Napier University – Data Analytics (SET09120)

1. Introduction

The aim of this coursework is to demonstrate an understanding of OpenRefine and Weka to conduct an exploratory data mining of data provided and produce a report about my discoveries. OpenRefine is to be used to clean the data provided and Weka to provide visualisation tools and algorithms for the data analysis and predictive modelling. Weka is a popular suite of machine learning software written in Java used for data processing, clustering, classification, regression, visualisation and feature selection.

2. Data Preparation

Data preparation consists of cleaning, structuring and integrating data to make it ready for analysis. This produces a neat and well-prepared dataset which results in accurate, meaningful and clean data visualisation that enhances the overall machine intelligence. [1] For this I used Google's OpenRefine tools.

2.1. Data Cleaning

The data provided came with many errors which were to be fixed before starting to analyse them. For this, OpenRefine provides many useful tools that let you see the dataset errors and allows you to edit them easily. Below is a table of the changes I made to the dataset before mining the data:

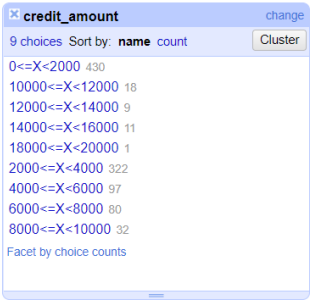
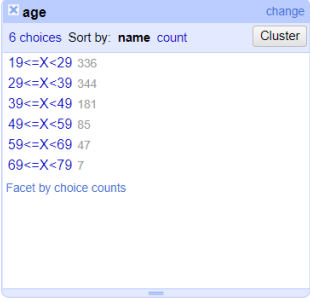
1. Renamed all the columns into their attribute names (checking_status, credit_history, purpose...)
2. Deleted the first column case_no since it would not be useful for the analysis of the data

Attribute	Before	After	Description
purpose			
	ather	other	Misspelt
	busines	business	Misspelt
	business	business	Misspelt
	Eduction	education	Lower case and misspelt
	Radio/Tv	radio/tv	Lower case so they are all equal
	'used car'	used car	Removed "
	'new car'	new car	Removed "
	'domestic appliance'	domestic appliance	Removed "
credit_amount			
	11132800 in case_no 432	1328	Compared value to others in purpose and seems unreasonable. Removed zeros and first number ones.

	19280000 in case_no 560	1928	Seems like the last 4 zeros were a mistake so I deleted them.
	13580000 in case_no 595	1358	Removed zeros, same as before.
	13860000 in case_no 648	1386	Removed zeros.
	13860000 in 648	1386	Removed zeros.
	63610000 in 660	6361	Removed zeros.
	7190000 in 444	719	Removed zeros.
	5180000 in 452	518	Removed zeros.
	5850000 in 524	585	Removed zeros.
personal_status			
	'female div/dep/mar'	'female div/sep/mar'	Misspelt dep → sep
age			
	222	22	Remove digit.
	333	33	Remove digit.
	-29	29	Change to positive.
	-34	34	Change to positive.
	-35	35	Change to positive.
	0.24	24	Change to a normal age.
	0.35	35	Change to a normal age.
	0.44	44	Change to a normal age.
	1	21	Seems unreasonable. By checking the clients requesting a loan for the same purpose looks like this person could be 21.
	6	26	Seems unreasonable. By checking the clients requesting a loan for the same purpose looks like this person could be 26.
job			
	yes	skilled	Assumed that: yes, the person is skilled.

2.2. Data conversion

For the nominal data set I changed all the numerical values to increments of x so the data was divided in groups. Data is transformed into nominal because it is easier for associations, since nominal values don't have a ranking, but the data is separated in bulks.

Change	Description
	Changed credit amount to nominal in increments of 2000 so it can be analysed by nominal algorithms.
	Changed age to nominal in increments of 10.

3. Data Analytics

Data mining is a process to discover interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from large amount of data. During data mining, the data is explained, and future rules are calculated by data analysis. [2]

Overfitting might happen if we supply too much data because the model will be created perfectly but for that data. We want to use the data to predict future unknowns not a perfect model of what the current data looks like.

3.1. Classification

Classification is the process of identifying to which set of the different categories a new observation belongs to, by using a dataset or data whose category class is known. It answers the question: "How likely is person X to get the loan?" where the X describes the different attributes of that exact person, so another person with similar attributes will be predicted the same result and this allows us to define rules.

Pruning is a technique used in machine learning that reduces the size of decision trees by removing sections of the tree that don't provide much information. This reduces the complexity of the final classifier and improves the accuracy of the final prediction. For classification I used a pruned tree, as the default settings were set to do so.

J48 was used for this, it is an algorithm used to generate a decision tree which is generated by C4.5 and allows you to predict the target variable of a new dataset record.

The results in Weka showed an accuracy of 80.3% so we could say that the model produced is a good one.

The confusion matrix shown is a table used to describe the performance of a classification model for which the true values are known. It shows the false positives and false negatives. A false positive happens when the algorithm predicts a yes and the actual answer was a no. A false negative happens when the algorithm predicts no but the actual answer is yes. For example:

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN=50	FP=10	60
Actual: YES	FN=5	TP=100	105
	55	110	

For the dataset provided, the confusion matrix was:

=== Confusion Matrix ===

```

  a    b  <-- classified as
659  41 |   a = good
179 121 |   b = bad

```

This means that there were 41 false positives and 179 false negatives, which were incorrectly classified.

Rule 1:

IF `checking_status` = <0 AND `credit_history` = existing paid AND `purpose` = radio/tv AND `employment` = >=7 THEN good (6.0/1.0)

If a customer has no current credit but with a credit history that has been paid and the purpose for the loan is for a radio or tv and they have been working for more than 7 years, then they will get the loan. Out of the 6 cases that apply to this rule, 1 has been incorrectly classified, which means that if the machine learning algorithm was to learn from this data, it would predict 1 wrong out of the 6 cases.

Rule 2:

IF `checking_status` = <0 AND `credit_history` = existing paid AND `purpose` = furniture/equipment AND `employment` = 1<=X<4 AND `age` > 23 THEN good (13.0/1.0)

If clients have a current account of less than zero but they have a credit history of all credits have been paid and their purpose is to buy furniture/equipment and they have been employed for at least 1 year but less than 4 and they're above the age of 23 then they are very likely to get the loan. Out of 13 cases only 1 has been incorrectly classified.

Rule 3:

IF `checking_status` = <0 AND `credit_history` = existing paid AND `purpose` = new car THEN bad (42.0/15.0)

For most clients in the model, if they have existing paid credits but nothing in the current account and they want to buy a new car then they are most likely to not get the loan. Out of all the 42 cases, 15 were incorrectly classified.

Rule 4:

IF `checking_status` = no checking THEN good (394.0/46.00)

As we can see in this rule, 88% of the clients who had no checking status got the loan. However, 46 of out the 394 were incorrectly classified.

Rule 5:

IF `checking_status` = < 0 AND `credit_history` = existing paid AND (`purpose` = education OR new car OR repairs OR retraining OR education) THEN bad

As we can see when the clients have a checking status of less than 0 and existing paid credit history but their purpose is either education / a new car / repairs / retraining or education they are not likely to get the loan. However, if the purpose is other there are more chances to get the loan.

For this rule I grouped several other rules since they all had the same class and attributes before the purpose.

Rule 6:

IF `checking_status` = <0 AND `credit_history` = existing paid AND `purpose` = furniture/equipment AND `employment` = 1<=X<4 AND `age` > 23 THEN good (13.0/1.0)

As we can see if the client is over 23 years old, has been employed for at least 1 year but less than 4, wants to buy furniture and has existing paid history, even if their `checking_status` is less than 0 then they will be likely to get the loan. However, for the clients who have the same attributes but are less than 23 years old then they will not likely get the loan (5.0/1.0).

General overview:

As we can see in 5 of the 6 rules the `credit_history` = existing paid and `checking_status` = < 0 therefore we could say it is a common pattern. Purpose then plays an important part in the final decision along with other attributes.

3.2. Regression

Regression allows you to predict a numerical value for a given set of input values. It is the easiest to perform and the least powerful method of data mining. Hence, I decided to leave this one.

3.3. Association

The association rules are found by using the Apriori algorithm, which finds frequent itemsets and associations between the sets. The Apriori reaches good performance by decreasing the size of candidate sets. For this part I left the default settings that came with the Apriori algorithm but used my normalised dataset described in section 2.2.

Rule 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)

If there is no checking status and the purpose is radio/tv then they are likely to get the loan. There is a confidence of 0.94 which means that this rule is very accurate and is true for 94% of the cases that have those attributes.

Rule 2:

`checking_status=no` checking `employment=>=7` 115 ==> `class=good` 107 <conf:(0.93)>
lift:(1.33) lev:(0.03) [26] conv:(3.83)

Clients that don't have information about their checking status but have been working for more than 7 years are more likely to get the loan they request. This rule also has a confidence of 0.93. A confidence of 0.9 means that the rule is 90% efficient therefore if the value is 0.90 or above it means it is a very good model.

Rule 3:

`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)

Single males that have a skilled job and no checking status are very likely to get the loan.

Rule 4:

`checking_status=no` checking `credit_history=existing` paid `job=skilled` 130 ==>
`class=good` 117 <conf:(0.9)> lift:(1.29) lev:(0.03) [26] conv:(2.79)

For most clients, if they have existing paid credits and a skilled job then they are 90% likely to get the loan, as the confidence shows. This rule is the one that has the lowest confidence, but it is still 0.90 so it is still very efficient.

Rule 5:

`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)

Clients that have a skilled job and are requesting a credit of less than 2000 are likely to get the loan even if there is no checking status.

Rule 6:

`checking_status=no` checking `age=29<=X<39` 151 ==> `class=good` 137 <conf:(0.91)>
lift:(1.3) lev:(0.03) [31] conv:(3.02)

Clients aged between 29 and 39 and that have no checking status are likely to get the loan they requested. This is true for 91% of the clients.

3.4. Clustering

Clustering allows a user to make groups of data to determine patterns from the data. Weka expects me to know how many clusters I want to create, I decided to make 6 since we need to find 6 rules. Apart from that, the rest of the settings were set by default.

I used the SimpleKMeans algorithm which uses the Euclidian distance to measure the distances between instances and clusters and decide to which cluster the instance should go. The seed value in the settings defines the initial assignment of instances to the clusters. The initial clusters are randomly selected and then modified to produce the most different possible clusters.

The 6 clusters made where the following:

	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
checking status	no checking	no checking	0<=X<200	0<=X<200	no checking	no checking
credit history	existing paid	critical / other existing credit	existing paid	existing paid	existing paid	critical / other existing credit
purpose	furniture / equipment	new car	new car	radio/tv	radio/tv	used car
credit amount	2593	2543	4043	2406	2915	4975
saving status	<100	<100	<100	<100	<100	no known savings
employment	1<=X<4	1<=X<4	<1	1<=X<4	4<=X<7	>=7
personal status	female div/sep/mar	male single	female div/sep/mar	male single	male single	male single
age	36	40	30	34	31	44
job	unskilled resident	skilled	skilled	unskilled resident	skilled	high qualif/ self emp/mgmt
class	good	good	bad	good	good	good

Clustered Instances

0	76 (10%)
1	167 (21%)
2	165 (21%)
3	107 (13%)
4	193 (24%)
5	92 (11%)

As we can see most of the data was equally separated. Cluster 4 is the biggest one. The whole dataset was 1000 instances. However, for clustering I divided the data into two sets, one of 800 instances (the data used here) and another one consisting of the remaining 200 instances to test that the results were consistent. So if we add the number of instances in each cluster we can see it adds up to 800.

As we can see in Figure 1 of the Appendix, cluster 2 is the most likely to not get a loan. From the cluster 2 attributes we can deduce that people who have a checking status between 0 and 200, have existing paid credits, want to buy a new car, want a mean credit amount of 4043 but have less than 100 in savings and have been working for less than 1 year in a skilled job at the mean age of 30 will **not** get a loan.

By analysing the table, we can see that the checking status is not a deciding factor for the loan since most of the clusters have no status of the existing account but 2 of them have very little ($0 \leq X < 200$) so it doesn't seem like they take those attributes very much in account.

Another detail we can observe from the table is that for 5 out of 6 clusters, their savings account is < 100 and the other cluster doesn't even have savings accounts. From this we can also deduce that saving money is not taken into account for the final decision.

We can also see that most of the clients are male singles, while if they are feminine div/sep/mar 50% get the loan and the other 50% doesn't.

4. Conclusion

Overall, I could conclude that the most effective technique for making the most realistic rules is Association, since it finds associations between items with no particular focus on a target one.

Classification was good to gain an insight into the data but not mainly used to find rules.

Both can be used for prediction and are used for supervised learning, which means that the output of the data is known. However, association can also be used for unsupervised learning (when the output of the data provided to the machine learning algorithm is not known).

