DATA ANALYTICS

Coursework II

Stefan Hristov
40284739

Stefan Hristov
40284739

Table of Contents

Preparation and Cleaning

The original data was provided in the form of an excel sheet. The file was directly loaded into OpenRefine. Number of actions were taken to prepare and clean the data for further analysis.

- At loading the data in OpenRefine the option Parse next was deselected so we keep the first line of data instead of using it for column headers.
- The columns were renamed to the names provided in the specification sheet.
- Text facet was made from the second column (checking_status) and the single quotation marks were removed from the data.
- The same process was done to columns 3, 6, 7, 8 and 10.
- With a text facet on column 4, single quotation marks were removed as well. “ather” was changed to “other”. “busines” and “busness” were changed to “business”. “Eduction” was changed to “education”. “Radio/Tv” was changed to “radio/tv”.
- Numeric facet was created on column 9. All occurrences of negative age were changed to positive. Decimal age was also changed to integer.
- Case number 54 – Age changed from 1 to 21 based on the information of employment between 1 and 4 years, single and credit for a used car. All of this suggests a young person.
- Case number 26 – Age changed from 6 to 26 based on the context. Single and the reason for the credit is furniture. This could suggest a young person, in their middle 20s.
- Case number 68 – Age changed from 222 to 22. Input mistake.
- Case number 80 – Age changed from 222 to 22. Input mistake.
- Case number 174 – Age changed from 333 to 33. Input mistake.
- Case number 432 – Credit amount changed from 111 328 000 to 111 328. Based on being highly qualified and age 29 with reason for credit “other”, it was concluded it could be for an apartment. The amount of 111 328 makes sense for an apartment.
- Case number 560 – Credit amount changed from 19 280 000 to 1928. The purpose of the credit is for furniture so 1928 makes most sense.
- Case number 595 – Credit amount changed from 13 580 000 to 135 800. Highly skilled 40 years old person with more than 7 years work could want to buy a villa or some expensive toy for a hobby. The amount must be 135 800 instead of 13 580 or 1358 since they credit was denied.
- Case number 648 – Credit amount changed from 13 860 000 to 13 860. With purpose of a new car for a 26 years old female with under 4 years of experience 138 600 credit doesn’t make sense, therefore the amount must be 13 860.
- Case number 660 – Credit amount changed from 63 610 000 to 6361. With other critical credit and good for this credit for equipment, only 6361 makes sense.
- Case number 452 – Credit amount changed from 5 180 000 to 518. The purpose is a tv.
- Case number 514 – Credit amount changed from 5 850 000 to 585. The purpose is a tv.

- Case number 444 – Credit amount changed from 7 190 000 to 7190. In Germany the university education is free so 71 900 doesn't make sense. Any courses that cost under 1000 (719) would have probably be granted considering 7+ years of experience with purpose of education. Therefore, the amount must be 7190.
- In column 10, there were 2 occurrences of job "yes". They were changed to skilled since is the most common group.

Further the data was exported to an excel format file. Then columns and credit amount and age were changed to categorical values. Credit amount were combined in categories as under 1000, under 2500, under 5 000, under 10 000 and over 10 000. Age was combined in 18 to 23, under 30, under 40, under 50, under 60 and over 60. The data was exported to an excel format file as well. Both files were then exported from Excel to csv files. The csv files were made into arff files. At this point the files were ready to be used for analysis in Weka.

Analysis

The specification suggests the use of minimum 3 techniques and 6 rules for each of them.

Classification

For the classification part the J48 algorithm was used since it produces a pruned tree and it was evaluated to be the best choice for this assignment. The dataset used for this was the categorical data (the one that age and credit amount were categorized) because it produces a higher percent of overall accuracy of 80.3%. The algorithm was ran using a training set as instructed in the lectures.

Since classification is for prediction and the context is loaning money from a bank, accuracy and specific cases are valued over general rules. Therefore, as long as there was a reasonable coverage, rules produced deep in the tree were preferred. Some rules are more general since they have good coverage and good accuracy.

Rule 1

```
checking_status = <0

|   credit_history = existing paid

|   |   saving_status = <100

|   |   |   purpose = radio/tv

|   |   |   |   employment = 1<=X<4

|   |   |   |   |   job = skilled: bad (10.0/2.0)
```

This rule has coverage of 1% and accuracy of 80% and is on 6th level in the tree.

Rule 2

checking_status = 0<=X<200

| saving_status = <100

| | credit_amount = 5000<=X<10000

| | | personal_status = male single: good (15.0/3.0)

This rule has coverage of 1.5% and accuracy of 80% and is on 4th level in the tree.

Rule 3

checking_status = <0

| credit_history = existing paid

| | saving_status = <100

| | | purpose = new car: bad (31.0/9.0)

This rule has coverage of 3.1% and accuracy of 70.9% and is on 4th level in the tree.

Rule 4

checking_status = 0<=X<200

| saving_status = <100

| | credit_amount = X>=10000: bad (11.0)

This rule has coverage of 1.1% and accuracy of 100% and is on 3rd level in the tree.

Rule 5

checking_status = 0<=X<200

| saving_status = no known savings: good (45.0/7.0)

This rule has coverage of 4.5% and accuracy of 84.4% and is on 2nd level in the tree.

Rule 6

checking_status = no checking: good (394.0/46.0)

This rule has coverage of 39.4% and accuracy of 88.3% and is on 1st level in the tree.

Association

For association, the Apriori algorithm was used as suggested in the lectures. The minimum support was set to 10% and the minimum confidence was set to 90%. The 6 best rules with regard to confidence were chosen. For all of the rules not only the confidence is strong, but the lift is really high as well.

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)
```

This rule has confidence of 94% and lift of 1.35.

Rule 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)
```

This rule has confidence of 93% and lift of 1.34.

Rule 3

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

This rule has confidence of 93% and lift of 1.33.

Rule 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)
```

This rule has confidence of 92% and lift of 1.32.

Rule 5

```
checking_status=no checking credit_amount=1000<=X<2500 job=skilled 125 ==> class=good
115    <conf:(0.92)> lift:(1.31) lev:(0.03) [27] conv:(3.41)
```

This rule has confidence of 92% and lift 1.31.

Rule 6

checking_status=no checking age=30<X<=40 142 ==> class=good 130 <conf:(0.92)> lift:(1.31) lev:(0.03) [30] conv:(3.28)

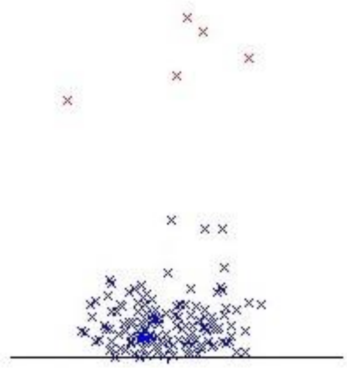
This rule has confidence of 92% and lift 1.31.

Clustering

For clustering, simple K means is the algorithm used. The number of clusters to be produced is set to 10 in order to produce slightly more accurate clusters (it splits the whole data in the selected number of clusters). The accuracy is evaluated with the help of a chart. The top 6 were chosen as rules.

Rule 1

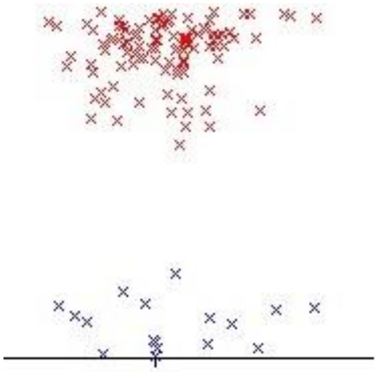
checking_status	no checking
credit_history	existing paid
purpose	radio/tv
credit_amount	2500<=X<5000
saving_status	no known savings
employment	>=7
personal_status	male single
age	30<x<=40
job	skilled
class	good



This rule has 144 occurrences which is 14.4% with very high accuracy.

Rule 2

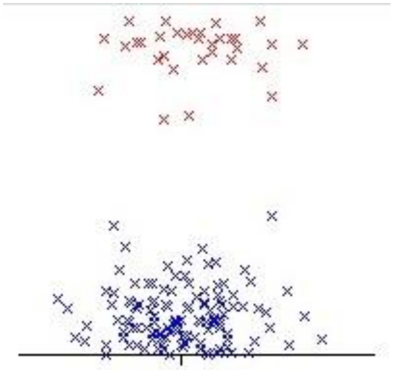
checking_status	<0
credit_history	existing paid
purpose	new car
credit_amount	2500<=X<5000
saving_status	<100
employment	>=7
personal_status	male single
age	30<x<=40
job	skilled
class	bad



This rule has 122 occurrences which is 12.2% with decent accuracy.

Rule 3

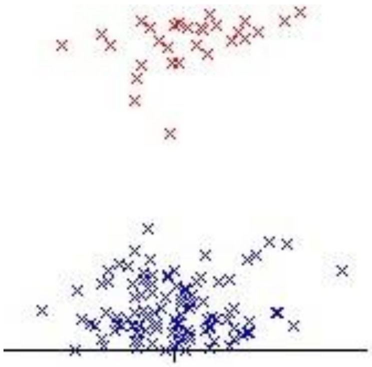
checking_status	0<=X<200
credit_history	existing paid
purpose	radio/tv
credit_amount	1000<=X<2500
saving_status	<100
employment	1<=X<4
personal_status	male single
age	30<X<=40
job	unskilled resident
class	good



This rule has 156 occurrences which is 15.6% with decent accuracy.

Rule 4

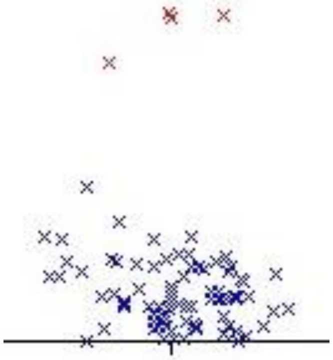
checking_status	<0
credit_history	existing paid
purpose	new car
credit_amount	2500<=X<5000
saving_status	<100
employment	>=7
personal_status	male single
age	30<X<=40
job	skilled
class	bad



This rule has 136 occurrences which is 13.6% with decent accuracy.

Rule 5

checking_status	no checking
credit_history	existing paid
purpose	used car
credit_amount	1000<=X<2500
saving_status	<100
employment	4<=X<7
personal_status	male single
age	23<X<=30
job	skilled
class	good



This rule has 85 occurrences which is 8.5% with high accuracy.

Rule 6

checking_status	no checking
credit_history	existing paid
purpose	radio/tv
credit_amount	1000<=X<2500
saving_status	100<=X<500
employment	>=7
personal_status	male single
age	40<X<=50
job	skilled
class	good

This rule has 42 occurrences which is 4.2% with high accuracy.

Conclusion

From the used methodologies and the data provided, perhaps the classification is the most useful. Classification provides information used for predictions, where association and clustering are used for describing the data. Of course, analysing the data still helps deducting predictions.

From the results from above, it seems clustering has the lowest value since it is not as precise as the association or classification.

Overall, the data seems non-authentic and therefore neither of the methodologies (since recommended to be based on the training set) produced accurate rules. The rules create on the training set to be used to describe the training set seems subjective.

Furthermore, the data seems it has been manipulated beforehand, perhaps clustered already to some extent. There are many un-explained decisions and there is no clear predicament to whether someone will be classified as good or bad.

Additionally, perhaps it was going to be more helpful if data of such size has been evaluated manually and deductions were made based on relation between specific properties with applied heuristics. Perhaps the algorithms proposed do not perfectly fit the dataset.

In conclusion, the most useful rules perhaps are the classification rules, but they should only be used as guide line since they have relative low coverage and not high enough accuracy to be used reliably in context of money loaning. The final decision should be taken from a trained person provided with further information for the specific client.