Data Analytics Coursework 2

Roderick Ewles

40330977

1 Introduction

The aim or this report is to carry out Data cleaning and data mining on a set of financial data. The data set is a set of real data from a bank and it contains 1000 cases with different attributes and the bank's decision on whether or not to provide the client with a loan. The software used to carry out this analysis is open refine and weka. Initially open refine was used to clean and prepare the data and weka was used to carry out the analysis. The aim of the course work is to use up to four data analysis techniques to see if any rules can be inferred from the data. The goal is to provide 6 rules for each analysis method.

2 Data preparation

2.1 Data Cleaning

Before any analysis can be carried out the data must first be prepared. Data cleaning is the first step in this process. Data cleaning is essential as if it is not carried out any analysis performed on the data set will be less accurate. As well as this some types of dirty data, for example blank fields, will prevent certain algorithms from working on the data set.

The first task before opening the data set in open refine is to open the data set in excel, which reveals that the data set has no column headings. A new row is added and the headings are applied to the columns.

Before any data cleaning can be carried out it is important to understand the data set and the expected values for each attribute. The data set has eleven attributes, the first of which is a case number. This is quite important to the bank but is meaningless for analysis purposes as each case number is assigned randomly. The second is checking status which is the status of the customer's current account. The expected values are: no checking account, less than zero, zero to two hundred and two hundred and above. The third is credit history, which is essentially how the customer has handled debt in the past. The expected values are: all paid, no credits/all paid, existing paid, delayed previously and critical/other existing credit. The fourth is the purpose of the loan which has a range of possible acceptable values. The fifth is the credit amount; it is the numerical value of the amount of money the customer is requesting. The sixth is saving status, which is how much savings, in a savings account or bonds, the customer has. The possible values are: no known, less than one hundred, between five hundred and one thousand and one thousand and above. The seventh is employment which is the employment status of the customer. The possible values are: unemployed, less than one year, between 1 and four years, between four and seven years, and seven or more years. The eighth is personal status which describes gender and marital status. Interestingly there is only one class of female in the data set. The ninth is the age of the customer which is a numeric value. The tenth is job, which describes the job status of the customer. The possible values are: unemp/unskilled non-res, unskilled resident, skilled and highqualif/self emp/mgmt. Finally there is class, which is either good or bad relating to if the customer is approved for a loan or not. Understanding what each of these means is extremely important to the cleaning process as without this knowledge it is impossible to know what acceptable values are for each of these fields.

The data can now be cleaned in open refine. The first type oddity data is missing values. Either the whole entry can be deleted, or a reasonable value can be inserted like a mean value or the most common. Data can also be entered in the wrong format, e.g. string or numeric, this can be corrected using facets in open refine to identify them at which point they can be corrected. Meaningless values are also required to be corrected using a numeric facet, these could be the result of a typo for example 1200 may become 120000. Certain knowledge of the data is required to correctly identify and fix these. A set violation also must be corrected, these tend to occur in nominal attributes so a text facet is used. Erroneous entries could also occur which are corrected using knowledge of the data set. An example of this could be correct data in the wrong column. After these errors have been corrected it is then necessary to check that all duplicates have been removed from the data. This is important as duplicates increase the effect that entry has in analysis. Duplicates were unlikely as each entry had a case number however the process for identifying these was carried out just in case. This process involves exporting the data as a .csv and opening it in word. This is then copied into the first column of an excel spreadsheet and this is loaded in open refine. A text facet is applied and anything that occurs twice or more is a duplicate. After all of these processes have been carried out the data should be clean.

2.2 Data Conversion

The data must now be converted for two main reasons. The first is that weka requires an .arff file and the second is that certain algorithms work on certain data types. Firstly the creation of these different data sets will be described as this can be done in open refine. Following this, two techniques will be discussed for generating the correct file type.

The apriori algorithm requires all nominal data to work. This set of data is created in open refine by using a numeric facet on the numeric columns. A set of reasonable cut offs is then used to convert the data, for example old age is any age over sixty. This is done using transform in open refine which uses grel. If statements were used to change the numeric values to nominal. There structure is: if (value>=x, “a”, “b”) which means if value is greater or equal to x then change it to “a” (string) else change it to “b”. Three sets of numeric data were also created, two binary and one ranked. Some of the information in the data is lost when converting to binary as you can only have two values so personal status would become gender, although a new column could be created for marital status. One of the binary sets is a 0/ 1 set and the other is a -1/1 set. The ranked set was produced by taking the number of occurrences and adding one, then dividing by two. However this is quite hard to interpret compared to the binary sets as the model produced from a binary set.

These data sets must then be converted to a format that weka can accept. There are two methods. The first is to open the .csv in a text editor like note pad and manually format it. The file must start with"@relation x", this is followed by the attributes "@attribute age real" or "@attribute gender {male, female}" after this has been done for all attributes "@data" is added before the data. This is then saved as a .txt which is renamed to .arff. However there is a handy shortcut in weka once the data sets are in a .csv format with column headings. In weka select tools then arff viewer. Following this open the .csv to be converted, then use file save as .arff which saves a lot of time. The data is now ready to be used in weka.

3 Data Analytics

N.B. There is a table containing all of the rules found at the start of the appendix. Full output from all experiments is also given in the appendix.

3.1 Classification

Classification uses supervised learning to predict a value of a nominal target class. This is predictive method so the models the different algorithms make will have an accuracy or error rate. This describes how good the model is at predicting the correct output.

3.1.1 OneR

```
For each attribute,
  For each value of that attribute, make a rule as follows:
    count how often each class appears
    find the most frequent class
    make the rule assign that class to this attribute-value.
  Calculate the error rate of the rules.
Choose the rules with the smallest error rate.
```

Figure 1: OneR pseudocode Data Mining 4th ed Witten, I; et al

OneR is a simple yet surprisingly powerful algorithm. It tries to build a Model based off of one attribute. It does this by looking at each attribute and count how often each class appears, in this case good or bad. It then creates a rule based on this attribute for the most commonly occurring class. After doing this for each attribute it compares the error rate for each rule then picks the rule with the least errors.

OneR still performs reasonably well when compared with more sophisticated attributes. However it does have a few draw backs, the most important is that it can only predict nominal values.

The first experiment was run on the cleaned mixed numeric and nominal data (credit-g-attr10cleaned1.arff). The only column removed was case. The model constructed uses credit amount but the minimum bucket size of 6 produces an output that is too complicated to infer any rules. The accuracy is 74.3 percent however which is good. The second experiment was run with the same settings, but credit amount has been removed. The model is made using the credit history attribute and is 71.7 percent accurate. It produces the first rule, if no credits/all paid or all paid then bad, good for all other credit types. The next experiments in One R vary the bucket size to see if more rules can be discovered. The bucket size is how OneR splits up numeric data into sets as has been manually done to produce the nominal data set. After removing the credit history and case attributes and setting the bucket size to 15 more rules can be discovered. These settings still give an acceptable accuracy of 71.3 percent. The second rule is if credit amount is less than 3962 then class is good. A third is if credit amount is above 10918 then class is bad. If the same setup is applied to the nominal data set the third rule is clearer but the model is less accurate as the data has been split based on bands that have been set as reasonable.

3.1.2 J48

J48 (weka's implementation of C4.5) is an extension of ID3. The algorithm for ID3 is shown on the right, J48 uses this algorithm but has some improvements. These are: it can handle both continuous and discrete attributes, it can handle data with missing attributes and the tree is pruned after creation.

1. For each attribute: compute it's entropy with respect to the target attribute.
2. Select the attribute with the lowest entropy.
3. Divide the data into separate homogenous sets.
4. Build a tree using these sets as branches.
5. Repeat this process on each sub tree.
6. One attribute is removed at each iteration. It stops when all the data is in leaves or there are no more attributes.

Figure 2:ID3 Algorithm

Running this on the cleaned data set with the default settings gives a tree with an accuracy of 78 percent. This produces a branch with great enough coverage and accuracy to infer rules from. The rule being: if checking status is no checking then class is good. This rule has a coverage of 394 and an accuracy of 88.3 percent. Another rule that could be inferred from this data is: Checking status between 0 and 200 and credit amount < 9283 Then class is good. This rule has a coverage of 248 and an accuracy of 64.5 percent.

If checking status is removed and J48 is run again with default settings this produces a tree with 75.6 percent accuracy. This tree produces more rules. The first is: credit history of Critical/other existing credit then class is good. This rule has a coverage of 293 and an accuracy of 82.9 percent. The second is: credit history of existing paid and credit amount<=5866 then class is good. This rule has a coverage of 465 and an accuracy of 71.6 percent. The next step is to remove credit history, which produces a tree of accuracy of 72 percent. Two more rules can be extracted from this tree. The first is: saving status of no known savings then class is good, with a coverage of 183 and an accuracy of 82.5 percent. The other is: saving status < 100 and credit amount <= 7511 then class is good, with a coverage of 561 and an accuracy of 66.7 percent. Further removing attributes results in all cases being classified as good. All of the rules that mention credit amount classify a high value as bad, in line with the OneR rule.

3.2 Association

Association rule mining attempts to find sets of items that occur together. There are two steps to this process, the first being to find the frequent item sets i.e. those item sets that meet a user defined support (coverage value). The second step is to convert these frequent item sets into rules that meet a user defined minimum confidence value. The confidence value is for the dataset, how often the rule holds true.

It is possible to figure out the item sets by using brute force and checking if every possible item set meets the minimum support. However this is highly impractical as it would be extremely time consuming. The Apriori algorithm is one solution to this problem. First it will check the individual attributes to see if they meet or exceed the support value, any that don't will not be used going forward. Following this two item combinations are made out of only those attributes that pass the support value. These two items sets will then be evaluated against the support value with any that fail being excluded. These two item sets then have another item added and are tested to see if they are frequent enough. This process continues until all the item combinations have been made or the tree of item sets dead ends as none of the combinations at the deepest level are frequent enough. This means that infrequent sets are pruned and their children are not considered as they will also not be frequent enough which drastically cuts down the number of cases to be considered.

The next part is to restructure these sets into rules which are then tested to see if they meet a user defined confidence value. For example if there is a set that considers when apples and oranges are bought together. Assuming these items have made it past the first stage a rule could be made like: if apples are bought then oranges are bought. The confidence is then calculated by dividing the support value of apples and oranges by the support value for apples. If this is above the confidence threshold the rule passes.

As mentioned earlier the apriori algorithm requires nominal data only. Using the method described in section two the nominal data set was created with the following cut offs applied to the numeric attributes: credit amount: 0-2000 low, 2000-9000 medium, >9000 high;
Age: 0-40 young, 40-60 middle aged, >60 old aged.

This data set was then used in weka to try and find rules. The first experiment was run in apriori with default settings. One rule that could be made using this model is: if no checking and skilled job then class is good, with a coverage of 264 and an accuracy of 90.1 percent. For the next experiment the confidence threshold was dropped to 0.85 and the number of rules was increased to 50. The list of rules has a lot of rules with checking status as no checking. One or the rules is if no checking then good with a coverage of 394 and an accuracy of 88.3 percent. Because of this the checking status attribute was removed and the experiment was run again with the same settings. This produces two more rules: if critical/other existing credit and low credit amount then class is good, coverage: 132, accuracy: 87.1 percent; and if critical/other existing credit and male single then class is good, coverage: 181, accuracy: 86.2 percent. As the experiment produced only nine rules the confidence threshold was dropped further to 0.8 for the next experiment. This produced several more rules: if critical/other existing credit and job is skilled then class is good, coverage: 185, accuracy: 83.7 percent; if female and jobs skilled and class good then young, coverage: 130, accuracy: 81.5 percent; if critical/other existing credit and age is young and job is skilled then class is good, coverage: 127, accuracy: 81.1 percent.

A final experiment was carried out with a confidence threshold of 0.75 to see if any other interesting rules with good coverage could be found. One of note is: if amount is medium and class is bad then saving status is <100, coverage: 150, accuracy 75.3 percent.

3.3 Clustering

Clustering is not used to predict a class but rather tries to tries to split the data up into natural groups. This can be done using several algorithms. The goal of clustering is to group items in such a way that they bear a closer resemblance to other members of their cluster than those in other clusters.

3.3.1 K means

K means clustering is a classic clustering technique. The user first defines the number of clusters. Then k points are chosen at random as cluster centres. All instances are assigned to their closest cluster according to their Euclidian distance from the k point. Following this the centroid is calculated. These centroids become the new centre values for each cluster. Instances are assigned to clusters again based on Euclidian distance, this process continues until the centroids do not move. Unfortunately there is a drawback to the algorithm which is that the algorithm minimises the distance of instances to the cluster centre but this minimum is a local minimum and there is no guarantee it is a global minimum. One strategy to find or approach a global minimum is to run k means several times changing the seed and picking the one with the minimum distance.

The algorithm was run several times as if there are not enough clusters, all clusters are classed as good, which makes sense as the majority of cases are classed as good. Therefore a number of clusters that would cluster some cases as bad was found through repeated experimentation. In experiment one five clusters are used with a seed of fifty. This gave a fourth cluster with the following attributes (the full output is shown in the appendix, only most interesting attributes discussed here): if single male and 3248 credit amount and savings <100 and checking account <0 then class is bad 21% coverage. If the experiment is run again with four clusters the following rule can be derived: if female and checking<0 and savings <100 and age 30.7 then case is bad 31% coverage. Interestingly both of these clusters have purpose listed as new car and job as skilled and credit history as existing paid. A third experiment was then run with the same settings ignoring all attributes except the credit amount, personal status, age and class. There are two female clusters leading to a rule: a younger woman trying to get a large loan is unlikely to be classed as good 20% coverage, where a middle aged woman trying to borrow less will be approved 23%. Another experiment was run using 5 clusters (seed 10) ignoring all attributes except: Credit history, Credit amount, saving status, Personal status. Comparing cluster 0 to cluster 4, most of the attributes are the same or better except for the checking status leading to the rule: if checking <0 and purpose is new car then class is bad, coverage 20 percent. Another test was run with the same settings ignoring: Checking status, Credit history, Credit amount and saving status. This leads to two more rules: if purpose radio/TV and employed for 7 or more years and male single and age 34 and job skilled then good, coverage: 33 percent; if purpose new car and employed between one and four years and female and age 28 and job skilled then class bad, coverage: 26 percent.

3.3.2 EM

Following this the EM algorithm was used briefly for a couple of reasons. EM works a bit differently to k means. It starts with initial guesses for the attributes in a cluster and works out the probability of each instance being in that cluster. This is soft clustering as it does not assign the instance to a cluster like k means. After this it will use those probabilities estimate the attributes. This process is then iterated until convergence is achieved, this is determined by how much the log likelihood changes over each iteration. EM is therefore has two very important steps, expectation and maximisation (hence EM). Expectation estimates the cluster to which each instance belongs and maximisation estimates the attributes of a cluster based on those instances.

Setting the number of clusters to -1 allows EM to generate the number of clusters that fit best (this is the default setting). When this is done EM produces 10 clusters (same variables ignored as the last k means experiment). This can be used to extract the rule: young female applicants requesting a small amount are more likely to be approved. A second experiment was then carried out using EM considering all attributes. This lead to two rules: if checking <0 and existing paid and purpose radio/TV and amount 1926 and savings <100 and employed between one and four years and age 26 and skilled and female then good, coverage 30 percent; if no checking and existing paid and purpose radio/TV and amount 1809 and savings <100 and employed for seven or more years and single male and age 42 and skilled then good, coverage 34 percent. EM was run again with the same settings ignoring the following attributes: purpose, employment status, personal status, age and job. This produces two more rules: if no checking and existing paid and amount is 2984.7 and savings <100 then good, coverage: 25 percent; if no checking and existing paid and amount 1338.7 and savings <100 then bad, coverage 32 percent. Another experiment was run using the same attributes ignoring: Checking status, Credit history, Credit amount, Saving status and Age. This produces the rule: if purpose new car and employed for 7 or more years and single male and skilled then good, coverage: 32 percent. A final experiment was carried out on the data with the same settings to see if there was any pattern between purpose personal status and age. An interesting rule can be inferred from this data: if purpose is new car and single male and age is 51.5 then good, coverage: 21 percent.

Please note as EM outputs probability of an instance belonging to a cluster the rules are inferred from the most probable member of a cluster.

Conclusion

An initial examination of the data was carried out with One R, this found three important rules, two which make sense and one that is unexpected. The two rules describing credit amount make sense i.e. low amounts get approved and very high amounts do not. However the credit history rule where those with good credit are classed as bad is strange as it is the opposite of what is expected. This was investigated further with other algorithms.

J48 gave a very clear rule of no checking then good. This is unexpected as well, as if they are not already a bank customer it would be expected that they are a greater risk but the data does not show this. The next rule makes sense, if they have some money in a checking amount and don't ask for too much they are approved. The next rule is again unexpected as being approved with critical or other existing credit would seem like a warning sign not to approve a loan. The next three rules make sense as by and large the conditions are positive and they are approved.

Considering the J48 trend of bad credit being approved, Apriori generates rules that make sense. The checking attribute dominated the output with the already discovered rule of no checking then good. Any combination of bad credit with a low credit amount, skilled young or male was approved. Encouragingly, young skilled female customers were approved. Another interesting and logical rule discovered was if the customer has little savings and asks for a medium credit amount they are declined.

Kmeans revealed some interesting rules. The first of these is a pair of rules, young women are likely to be declined where middle aged women are more likely to be approved. The algorithm highlighted the fact that a checking account in arrears is likely to be declined, particularly if the loans purpose is a new car. It also showed that young women trying to get a loan for a new car are often declined, where men approaching middle age seeking a loan for a radio or TV are approved.

EM revealed some interesting rules, including one that contradicts a k means rule which was young women are approved. This may be due to the fact that the algorithm works differently to kmeans. It also revealed that individuals with a job who have paid off previous loans who have a low amount of savings and are asking for a small amount are approved if they are male or female. Then a pair of contradictory rules was discovered, in that all of the conditions are the same but the higher amount was approved and the lower amount declined. Again this may be due to how the algorithm functions. Finally a rule was discovered that skilled single men who have had a job for at least 7 years are approved a loan for the purpose of a new car. This lead to another experiment to see if there was any connection between personal status, purpose and age. This experiment found a "mid-life crisis" rule: single men aged 51 seeking a loan for the purpose of a new car are approved.

The most concerning rules discovered were the rules discovered using apriori showing customers with poor credit history get approved. Although, this may be due to the fact that these customers are more profitable to the bank as long as they do not go bankrupt. These customers would be more profitable as the bank will charge customers who miss payments.

4 Appendix

Algorithm	Condition	Class
OneR	No credit/all paid or all paid	Bad
OneR	Credit amount <3962	Good
OneR	Credit Amount >10918	Bad
J48	No checking	Good
J48	Checking status 0<=x<200 & Credit amount <9283	Good
J48	Critical/other existing credit	Good
J48	Credit History existing paid & credit amount <=5866	Good
J48	No known savings	Good
J48	Savings <100 & credit amount <= 7511	Good
Apriori	No checking & skilled	Good
Apriori	No checking	Good
Apriori	Critical/other existing credit & credit amount low	Good
Apriori	Critical/other existing credit & single male	Good
Apriori	Critical/other existing credit & skilled	Good
Apriori	Female & job skilled & class good	Young
Apriori	Critical/other existing credit & young & skilled	Good
Apriori	Credit amount medium & class bad	Savings <100
Kmeans	Single male & amount 3248 & savings <100 & checking <0 & new car	Bad
Kmeans	Female & checking <0 & savings <100 & age 30.7 & new car	Bad
Kmeans	Female & young	Bad
Kmeans	Female & middle aged	Good
Kmeans	Checking <0 & new car	Bad
Kmeans	Radio/TV & employed >=7 & single male & age 34 & skilled	Good
Kmeans	New car & employed 1<=x<4 & female & age 28 & skilled	Bad
EM	Female & young	Good
EM	Checking <0 & existing paid & radio/TV & amount 1926 & savings <100 & employed 1<=x<4 & age 26 & skilled & female	Good
EM	No checking & existing Paid & radio/TV & amount 1809 & savings <100 & employed >=7 & single male & age 42 & skilled	Good
EM	No checking & existing paid & amount 2984.7 & savings <100	Good
EM	No checking & existing paid & amount 1338.7 & savings <100	Bad
EM	New car & employed >=7 & single male & skilled	Good
EM	New car & single male & age 51	Good

Bibliography

Berthold, M. R., 2010. *Guide to intelligent Data Analysis.* 1st ed. London: Springer-Verlag London .

Hand, D., Heikki, M. & P, S., 2001. *Principles of Data Mining.* 1st ed. Massachusetts: Massachusetts Institute of Technology.

Witten, I. & Frank, E., 2017. *Data Mining.* 4th ed. Cambridge US: Elsevier.

OneR experiment 1

=== Run information ===

Scheme: weka.classifiers.rules.OneR -B 6

Relation: credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 10

Checking_status

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Credit_amount:

< 883.0 -> good

< 922.0 -> bad

< 938.0 -> good

< 979.5 -> bad

< 1206.5 -> good

< 1223.5 -> bad

< 1267.5 -> good

< 1286.0 -> bad

< 1325.5	-> good
< 1345.5	-> bad
< 1821.5	-> good
< 1865.5	-> bad
< 3913.5	-> good
< 3969.0	-> bad
< 4049.5	-> good
< 4329.5	-> bad
< 4726.0	-> good
< 5024.0	-> bad
< 6322.5	-> good
< 6564.0	-> bad
< 6750.0	-> good
< 6917.5	-> bad
< 7760.5	-> good
< 8109.5	-> bad
< 9340.5	-> good
< 10331.5	-> bad
< 11307.0	-> good
>= 11307.0	-> bad

(743/1000 instances correct)

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances	743	74.3	%
Incorrectly Classified Instances	257	25.7	%

Kappa statistic 0.2993

Mean absolute error 0.257

Root mean squared error 0.507

Relative absolute error 61.1672 %

Root relative squared error 110.6259 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.911	0.650	0.766	0.911	0.832	0.321	0.631	0.760	good
	0.350	0.089	0.629	0.350	0.450	0.321	0.631	0.415	bad
Weighted Avg.	0.743	0.482	0.725	0.743	0.718	0.321	0.631	0.657	

=== Confusion Matrix ===

a b <-- classified as

638 62 | a = good

195 105 | b = bad

OneR experiment 2

=== Run information ===

Scheme: weka.classifiers.rules.OneR -B 6

Relation: credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R4

Instances: 1000

Attributes: 9

Checking_status

Credit_history

Purpose

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Credit_history:

critical/other existing credit -> good

existing paid -> good

delayed previously -> good

no credits/all paid -> bad

all paid -> bad

(717/1000 instances correct)

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances 717 71.7 %

Incorrectly Classified Instances 283 28.3 %

Kappa statistic 0.1567

Mean absolute error 0.283

Root mean squared error 0.532

Relative absolute error 67.3553 %

Root relative squared error 116.087 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.949	0.823	0.729	0.949	0.824	0.202	0.563	0.727	good
	0.177	0.051	0.596	0.177	0.272	0.202	0.563	0.352	bad
Weighted Avg.	0.717	0.592	0.689	0.717	0.659	0.202	0.563	0.615	

=== Confusion Matrix ===

a b <-- classified as

664 36 | a = good

247 53 | b = bad

OneR experiment 3

=== Run information ===

Scheme: weka.classifiers.rules.OneR -B 15

Relation: credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R2

Instances: 1000

Attributes: 9

Checking_status

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Credit_amount:

< 3962.0 -> good

< 4329.5 -> bad

< 10918.0 -> good

>= 10918.0 -> bad

(713/1000 instances correct)

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances 713 71.3 %

Incorrectly Classified Instances 287 28.7 %

Kappa statistic 0.1131

Mean absolute error 0.287

Root mean squared error 0.5357

Relative absolute error 68.3074 %

Root relative squared error 116.9045 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.967	0.880	0.719	0.967	0.825	0.169	0.544	0.719	good
	0.120	0.033	0.610	0.120	0.201	0.169	0.544	0.337	bad
Weighted Avg.	0.713	0.626	0.687	0.713	0.638	0.169	0.544	0.604	

=== Confusion Matrix ===

a b <-- classified as

677 23 | a = good

264 36 | b = bad

OneR nominal data experiment

=== Run information ===

Scheme: weka.classifiers.rules.OneR -B 15

Relation: credit-g-attr10cleaned-nominal-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R2

Instances: 1000

Attributes: 9

Checking_status

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Credit_amount:

very low -> good

high -> good

low -> good

very high -> bad

(707/1000 instances correct)

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 707 70.7 %

Incorrectly Classified Instances 293 29.3 %

Kappa statistic 0.0716

Mean absolute error 0.293

Root mean squared error 0.5413

Relative absolute error 69.7325 %

Root relative squared error 118.1201 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.977	0.923	0.712	0.977	0.824	0.127	0.527	0.711	good
	0.077	0.023	0.590	0.077	0.136	0.127	0.527	0.322	bad
Weighted Avg.	0.707	0.653	0.675	0.707	0.617	0.127	0.527	0.595	

=== Confusion Matrix ===

a b <-- classified as

684 16 | a = good

277 23 | b = bad

J48 experiment 1

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 10

Checking_status

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

Checking_status = <0

| Credit_history = critical/other existing credit: good (67.0/18.0)

| Credit_history = existing paid

| | Purpose = radio/tv

| | | Employment = >=7: good (6.0/1.0)

| | | Employment = 1<=X<4

| | | | Job = skilled: bad (11.0/3.0)

| | | | Job = unskilled resident: good (4.0/2.0)

| | | | Job = high qualif/self emp/mgmt: good (2.0)

| | | | Job = unemp/unskilled non res: bad (0.0)

| | | Employment = 4<=X<7: good (6.0/3.0)

| | | Employment = unemployed: bad (1.0)

| | | Employment = <1

| | | | Age <= 30: good (5.0/1.0)

| | | | Age > 30: bad (4.0)

| | Purpose = education: bad (7.0/2.0)

| | Purpose = furniture/equipment

| | | Employment = >=7

| | | | Credit_amount <= 1680: good (3.0)

| | | | Credit_amount > 1680: bad (6.0/1.0)

| | | Employment = 1<=X<4

| | | | Age <= 23: bad (5.0/1.0)

| | | | Age > 23: good (13.0/1.0)

| | | Employment = 4<=X<7: good (5.0/2.0)

| | | Employment = unemployed: bad (5.0/2.0)

| | | Employment = <1: good (8.0/2.0)

| | Purpose = new car: bad (42.0/15.0)

| | Purpose = used car

| | | Credit_amount <= 6850: good (10.0/1.0)

| | | Credit_amount > 6850: bad (4.0)

| | Purpose = business

| | | Age <= 23: good (2.0)

| | | Age > 23: bad (2.0)

| | Purpose = domestic appliance: bad (5.0/1.0)

| | Purpose = repairs: bad (1.0)

| | Purpose = other: good (2.0)

| | Purpose = retraining: bad (1.0)

| Credit_history = delayed previously: bad (12.0/3.0)

| Credit_history = no credits/all paid: bad (13.0/3.0)

| Credit_history = all paid: bad (22.0/6.0)

Checking_status = 0<=X<200

| Credit_amount <= 9283: good (248.0/88.0)

| Credit_amount > 9283: bad (21.0/4.0)

Checking_status = no checking: good (394.0/46.0)

Checking_status = >=200: good (63.0/14.0)

Number of Leaves : 34

Size of the tree : 46

Time taken to build model: 0.06 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances 780 78 %

Incorrectly Classified Instances 220 22 %

Kappa statistic 0.3969

Mean absolute error 0.3183

Root mean squared error 0.3989

Relative absolute error 75.7613 %

Root relative squared error 87.0575 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.941	0.597	0.786	0.941	0.857	0.429	0.776	0.856	good
	0.403	0.059	0.747	0.403	0.524	0.429	0.776	0.607	bad
Weighted Avg.	0.780	0.435	0.775	0.780	0.757	0.429	0.776	0.781	

=== Confusion Matrix ===

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

J48 experiment 2

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 9

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

Credit_history = critical/other existing credit: good (293.0/50.0)

Credit_history = existing paid

| Credit_amount <= 5866: good (465.0/132.0)

| Credit_amount > 5866

| | Credit_amount <= 10722

| | | Job = skilled: bad (27.0/12.0)

| | | Job = unskilled resident: good (6.0/1.0)

| | | Job = high qualif/self emp/mgmt

| | | | Purpose = radio/tv: bad (2.0)

| | | | Purpose = education: good (1.0)

| | | | Purpose = furniture/equipment: bad (2.0)

| | | | Purpose = new car: bad (5.0/2.0)

| | | | Purpose = used car: good (7.0/1.0)

| | | | Purpose = business: good (0.0)

| | | | Purpose = domestic appliance: good (0.0)

| | | | Purpose = repairs: good (0.0)

| | | | Purpose = other: good (0.0)

| | | | Purpose = retraining: good (0.0)

| | | Job = unemp/unskilled non res: good (1.0)

| | Credit_amount > 10722: bad (14.0/1.0)

Credit_history = delayed previously: good (88.0/28.0)

Credit_history = no credits/all paid

| Personal_status = male single

| | Saving_status = no known savings: good (3.0)

| | Saving_status = <100: bad (19.0/4.0)

| | Saving_status = 500<=X<1000: bad (0.0)

| | Saving_status = >=1000: bad (0.0)

| | Saving_status = 100<=X<500: good (2.0/1.0)

| Personal_status = female div/dep/mar: bad (12.0/3.0)

| Personal_status = male div/sep: good (2.0)

| Personal_status = male mar/wid: good (2.0)

Credit_history = all paid

| Purpose = radio/tv

| | Age <= 31: bad (4.0/1.0)

| | Age > 31: good (5.0)

| Purpose = education: bad (3.0)

| Purpose = furniture/equipment

| | Saving_status = no known savings: bad (0.0)

| | Saving_status = <100: bad (4.0)

| | Saving_status = 500<=X<1000: good (3.0/1.0)

| | Saving_status = >=1000: good (1.0)

| | Saving_status = 100<=X<500: bad (0.0)

| Purpose = new car: bad (12.0/3.0)

| Purpose = used car

| | Personal_status = male single: good (2.0)

| | Personal_status = female div/dep/mar: bad (3.0/1.0)

| | Personal_status = male div/sep: good (0.0)

| | Personal_status = male mar/wid: good (0.0)

| Purpose = business: bad (7.0/3.0)

| Purpose = domestic appliance: good (1.0)

| Purpose = repairs: bad (0.0)

| Purpose = other: bad (2.0)

| Purpose = retraining: good (2.0)

Number of Leaves : 43

Size of the tree : 54

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances 756 75.6 %

Incorrectly Classified Instances 244 24.4 %

Kappa statistic 0.2956

Mean absolute error 0.3551

Root mean squared error 0.4213

Relative absolute error 84.5054 %

Root relative squared error 91.9442 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.957	0.713	0.758	0.957	0.846	0.349	0.688	0.800	good
	0.287	0.043	0.741	0.287	0.413	0.349	0.688	0.522	bad
Weighted Avg.	0.756	0.512	0.753	0.756	0.716	0.349	0.688	0.717	

=== Confusion Matrix ===

a b <-- classified as

670 30 | a = good

214 86 | b = bad

J48 experiment 3

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 8

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

Saving_status = no known savings: good (183.0/32.0)

Saving_status = <100

| Credit_amount <= 7511: good (561.0/187.0)

| Credit_amount > 7511: bad (42.0/12.0)

Saving_status = 500<=X<1000: good (63.0/11.0)

Saving_status = >=1000: good (48.0/6.0)

Saving_status = 100<=X<500

| Credit_amount <= 6204: good (89.0/26.0)

| Credit_amount > 6204: bad (14.0/6.0)

Number of Leaves :   7

Size of the tree :   10

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances   720   72   %

Incorrectly Classified Instances   280   28   %

Kappa statistic   0.1315

Mean absolute error   0.3916

Root mean squared error   0.4425

Relative absolute error   93.205 %

Root relative squared error 96.561 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.974	0.873	0.722	0.974	0.830	0.201	0.630	0.772	good
	0.127	0.026	0.679	0.127	0.213	0.201	0.630	0.399	bad
Weighted Avg.	0.720	0.619	0.709	0.720	0.645	0.201	0.630	0.660	

=== Confusion Matrix ===

```
  a   b   <-- classified as
 682  18 |   a = good
 262  38 |   b = bad
```

Apriori experiment 1

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Copy of credit-g-attr10cleanednominal2-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 10

Checking_status

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (100 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 27

Size of set of large itemsets L(2): 151

Size of set of large itemsets L(3): 175

Size of set of large itemsets L(4): 65

Size of set of large itemsets L(5): 1

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 150 ==> Class=good 139 <conf:(0.93)> lift:(1.32) lev:(0.03) [34] conv:(3.75)

5. Checking_status=no checking Credit_amount=low Job=skilled 114 ==> Class=good 105 <conf:(0.92)> lift:(1.32) lev:(0.03) [25] conv:(3.42)

6. Checking_status=no checking Credit_amount=low 170 ==> Class=good 156 <conf:(0.92)> lift:(1.31) lev:(0.04) [37] conv:(3.4)

7. Credit_history=existing paid Employment=<1 111 ==> Age=young 101 <conf:(0.91)> lift:(1.3) lev:(0.02) [23] conv:(3.03)

8. Checking_status=no checking Credit_history=existing paid Job=skilled 128 ==> Class=good 116 <conf:(0.91)> lift:(1.29) lev:(0.03) [26] conv:(2.95)

9. Checking_status=no checking Job=skilled 264 ==> Class=good 238 <conf:(0.9)> lift:(1.29) lev:(0.05) [53] conv:(2.93)

Apriori experiment 2

=== Run information ===

Scheme: weka.associations.Apriori -N 50 -T 0 -C 0.85 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Copy of credit-g-attr10cleanednominal2-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 10

Checking_status

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (100 instances)

Minimum metric <confidence>: 0.85

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 27

Size of set of large itemsets L(2): 151

Size of set of large itemsets L(3): 175

Size of set of large itemsets L(4): 65

Size of set of large itemsets L(5): 1

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 150 ==> Class=good 139 <conf:(0.93)> lift:(1.32) lev:(0.03) [34] conv:(3.75)

5. Checking_status=no checking Credit_amount=low Job=skilled 114 ==> Class=good 105 <conf:(0.92)> lift:(1.32) lev:(0.03) [25] conv:(3.42)

6. Checking_status=no checking Credit_amount=low 170 ==> Class=good 156 <conf:(0.92)> lift:(1.31) lev:(0.04) [37] conv:(3.4)

7. Credit_history=existing paid Employment=<1 111 ==> Age=young 101 <conf:(0.91)> lift:(1.3) lev:(0.02) [23] conv:(3.03)

8. Checking_status=no checking Credit_history=existing paid Job=skilled 128 ==> Class=good 116 <conf:(0.91)> lift:(1.29) lev:(0.03) [26] conv:(2.95)

9. Checking_status=no checking Job=skilled 264 ==> Class=good 238 <conf:(0.9)> lift:(1.29) lev:(0.05) [53] conv:(2.93)

10. Checking_status=no checking Personal_status=male single 232 ==> Class=good 208 <conf:(0.9)> lift:(1.28) lev:(0.05) [45] conv:(2.78)

11. Checking_status=no checking Personal_status=male single Age=young 144 ==> Class=good 129 <conf:(0.9)> lift:(1.28) lev:(0.03) [28] conv:(2.7)

12. Checking_status=no checking Credit_amount=medium Job=skilled 141 ==> Class=good 126 <conf:(0.89)> lift:(1.28) lev:(0.03) [27] conv:(2.64)

13. Checking_status=no checking Credit_amount=low Age=young 112 ==> Class=good 100 <conf:(0.89)> lift:(1.28) lev:(0.02) [21] conv:(2.58)

14. Checking_status=no checking 394 ==> Class=good 348 <conf:(0.88)> lift:(1.26) lev:(0.07) [72] conv:(2.51)

15. Checking_status=no checking Age=young Job=skilled 188 ==> Class=good 166 <conf:(0.88)> lift:(1.26) lev:(0.03) [34] conv:(2.45)

16. Checking_status=no checking Saving_status=<100 Job=skilled 125 ==> Class=good 110 <conf:(0.88)> lift:(1.26) lev:(0.02) [22] conv:(2.34)

17. Saving_status=<100 Personal_status=female div/dep/mar Job=skilled 123 ==> Age=young 108 <conf:(0.88)> lift:(1.25) lev:(0.02) [21] conv:(2.31)

18. Checking_status=no checking Credit_history=existing paid 187 ==> Class=good 164 <conf:(0.88)> lift:(1.25) lev:(0.03) [33] conv:(2.34)

19. Checking_status=no checking Credit_amount=medium Age=young 143 ==> Class=good 125 <conf:(0.87)> lift:(1.25) lev:(0.02) [24] conv:(2.26)

20. Checking_status=no checking Credit_amount=medium Personal_status=male single 135 ==> Class=good 118 <conf:(0.87)> lift:(1.25) lev:(0.02) [23] conv:(2.25)

21. Checking_status=no checking Credit_history=existing paid Age=young 134 ==> Class=good 117 <conf:(0.87)> lift:(1.25) lev:(0.02) [23] conv:(2.23)

22. Checking_status=no checking Age=young 267 ==> Class=good 233 <conf:(0.87)> lift:(1.25) lev:(0.05) [46] conv:(2.29)

23. Credit_history=critical/other existing credit Credit_amount=low 132 ==> Class=good 115 <conf:(0.87)> lift:(1.24) lev:(0.02) [22] conv:(2.2)

24. Saving_status=no known savings Personal_status=male single 116 ==> Class=good 101 <conf:(0.87)> lift:(1.24) lev:(0.02) [19] conv:(2.17)

25. Checking_status=no checking Credit_amount=medium 206 ==> Class=good 179 <conf:(0.87)> lift:(1.24) lev:(0.03) [34] conv:(2.21)

26. Saving_status=<100 Employment=<1 120 ==> Age=young 104 <conf:(0.87)> lift:(1.24) lev:(0.02) [20] conv:(2.12)

27. Checking_status=no checking Saving_status=<100 191 ==> Class=good 165 <conf:(0.86)> lift:(1.23) lev:(0.03) [31] conv:(2.12)

28. Credit_history=critical/other existing credit Personal_status=male single 181 ==> Class=good 156 <conf:(0.86)> lift:(1.23) lev:(0.03) [29] conv:(2.09)

29. Employment=<1 172 ==> Age=young 148 <conf:(0.86)> lift:(1.23) lev:(0.03) [27] conv:(2.06)

30. Checking_status=no checking Employment=1<=X<4 139 ==> Class=good 119 <conf:(0.86)> lift:(1.22) lev:(0.02) [21] conv:(1.99)

31. Personal_status=female div/dep/mar Job=skilled 196 ==> Age=young 167 <conf:(0.85)> lift:(1.22) lev:(0.03) [29] conv:(1.96)

Apriori experiment 3

=== Run information ===

Scheme: weka.associations.Apriori -N 50 -T 0 -C 0.85 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Copy of credit-g-attr10cleanednominal2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 9

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (100 instances)

Minimum metric <confidence>: 0.85

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 24

Size of set of large itemsets L(2): 119

Size of set of large itemsets L(3): 131

Size of set of large itemsets L(4): 52

Size of set of large itemsets L(5): 1

Best rules found:

1. Credit_history=existing paid Employment=<1 111 ==> Age=young 101 <conf:(0.91)> lift:(1.3) lev:(0.02) [23] conv:(3.03)

2. Saving_status=<100 Personal_status=female div/dep/mar Job=skilled 123 ==> Age=young 108 <conf:(0.88)> lift:(1.25) lev:(0.02) [21] conv:(2.31)

3. Credit_history=critical/other existing credit Credit_amount=low 132 ==> Class=good 115 <conf:(0.87)> lift:(1.24) lev:(0.02) [22] conv:(2.2)

4. Saving_status=no known savings Personal_status=male single 116 ==> Class=good 101 <conf:(0.87)> lift:(1.24) lev:(0.02) [19] conv:(2.17)

5. Saving_status=<100 Employment=<1 120 ==> Age=young 104 <conf:(0.87)> lift:(1.24) lev:(0.02) [20] conv:(2.12)

6. Credit_history=critical/other existing credit Personal_status=male single 181 ==> Class=good 156 <conf:(0.86)> lift:(1.23) lev:(0.03) [29] conv:(2.09)

7. Employment=<1 172 ==> Age=young 148 <conf:(0.86)> lift:(1.23) lev:(0.03) [27] conv:(2.06)

8. Personal_status=female div/dep/mar Job=skilled 196 ==> Age=young 167 <conf:(0.85)> lift:(1.22) lev:(0.03) [29] conv:(1.96)

Apriori experiment 4

=== Run information ===

Scheme: weka.associations.Apriori -N 50 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Copy of credit-g-attr10cleanednominal2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 9

Credit_history

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (100 instances)

Minimum metric <confidence>: 0.8

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 24

Size of set of large itemsets L(2): 119

Size of set of large itemsets L(3): 131

Size of set of large itemsets L(4): 52

Size of set of large itemsets L(5): 1

Best rules found:

1. Credit_history=existing paid Employment=<1 111 ==> Age=young 101 <conf:(0.91)> lift:(1.3) lev:(0.02) [23] conv:(3.03)

2. Saving_status=<100 Personal_status=female div/dep/mar Job=skilled 123 ==> Age=young 108 <conf:(0.88)> lift:(1.25) lev:(0.02) [21] conv:(2.31)

3. Credit_history=critical/other existing credit Credit_amount=low 132 ==> Class=good 115 <conf:(0.87)> lift:(1.24) lev:(0.02) [22] conv:(2.2)

4. Saving_status=no known savings Personal_status=male single 116 ==> Class=good 101 <conf:(0.87)> lift:(1.24) lev:(0.02) [19] conv:(2.17)

5. Saving_status=<100 Employment=<1 120 ==> Age=young 104 <conf:(0.87)> lift:(1.24) lev:(0.02) [20] conv:(2.12)

6. Credit_history=critical/other existing credit Personal_status=male single 181 ==> Class=good 156 <conf:(0.86)> lift:(1.23) lev:(0.03) [29] conv:(2.09)

7. Employment=<1 172 ==> Age=young 148 <conf:(0.86)> lift:(1.23) lev:(0.03) [27] conv:(2.06)

8. Personal_status=female div/dep/mar Job=skilled 196 ==> Age=young 167 <conf:(0.85)> lift:(1.22) lev:(0.03) [29] conv:(1.96)

9. Credit_history=existing paid Personal_status=female div/dep/mar Job=skilled 122 ==> Age=young 103 <conf:(0.84)> lift:(1.21) lev:(0.02) [17] conv:(1.83)

10. Employment=4<=X<7 Job=skilled 119 ==> Age=young 100 <conf:(0.84)> lift:(1.2) lev:(0.02) [16] conv:(1.79)

11. Credit_amount=medium Employment=1<=X<4 Job=skilled 131 ==> Age=young 110 <conf:(0.84)> lift:(1.2) lev:(0.02) [18] conv:(1.79)

12. Credit_history=critical/other existing credit Job=skilled 185 ==> Class=good 155 <conf:(0.84)> lift:(1.2) lev:(0.03) [25] conv:(1.79)

13. Credit_amount=medium Employment=1<=X<4 Class=good 121 ==> Age=young 101 <conf:(0.83)> lift:(1.19) lev:(0.02) [16] conv:(1.73)

14. Credit_history=critical/other existing credit 293 ==> Class=good 243 <conf:(0.83)> lift:(1.18) lev:(0.04) [37] conv:(1.72)

15. Saving_status=<100 Employment=1<=X<4 Job=skilled 140 ==> Age=young 116 <conf:(0.83)> lift:(1.18) lev:(0.02) [18] conv:(1.68)

16. Saving_status=no known savings 183 ==> Class=good 151 <conf:(0.83)> lift:(1.18) lev:(0.02) [22] conv:(1.66)

17. Employment=1<=X<4 Job=skilled Class=good 159 ==> Age=young 131 <conf:(0.82)> lift:(1.18) lev:(0.02) [19] conv:(1.64)

18. Credit_history=existing paid Employment=1<=X<4 Job=skilled 127 ==> Age=young 104 <conf:(0.82)> lift:(1.17) lev:(0.02) [15] conv:(1.59)

19. Employment=1<=X<4 Job=skilled 229 ==> Age=young 187 <conf:(0.82)> lift:(1.17) lev:(0.03) [26] conv:(1.6)

20. Personal_status=female div/dep/mar Job=skilled Class=good 130 ==> Age=young 106 <conf:(0.82)> lift:(1.16) lev:(0.01) [15] conv:(1.56)

21. Credit_amount=medium Employment=1<=X<4 177 ==> Age=young 144 <conf:(0.81)> lift:(1.16) lev:(0.02) [20] conv:(1.56)

22. Credit_history=critical/other existing credit Credit_amount=medium 150 ==> Class=good 122 <conf:(0.81)> lift:(1.16) lev:(0.02) [17] conv:(1.55)

23. Employment=1<=X<4 Personal_status=male single Class=good 128 ==> Age=young 104 <conf:(0.81)> lift:(1.16) lev:(0.01) [14] conv:(1.54)

24. Credit_history=critical/other existing credit Age=young Job=skilled 127 ==> Class=good 103 <conf:(0.81)> lift:(1.16) lev:(0.01) [14] conv:(1.52)

25. Employment=4<=X<7 174 ==> Age=young 141 <conf:(0.81)> lift:(1.16) lev:(0.02) [19] conv:(1.54)

26. Credit_history=existing paid Saving_status=<100 Job=skilled 192 ==> Age=young 155 <conf:(0.81)> lift:(1.15) lev:(0.02) [20] conv:(1.52)

27. Credit_history=existing paid Credit_amount=low Saving_status=<100 155 ==> Age=young 125 <conf:(0.81)> lift:(1.15) lev:(0.02) [16] conv:(1.5)

28. Purpose=radio/tv Saving_status=<100 169 ==> Age=young 136 <conf:(0.8)> lift:(1.15) lev:(0.02) [17] conv:(1.49)

29. Purpose=radio/tv Personal_status=male single 146 ==> Class=good 117 <conf:(0.8)> lift:(1.14) lev:(0.01) [14] conv:(1.46)

30. Credit_history=existing paid Personal_status=female div/dep/mar 186 ==> Age=young 149 <conf:(0.8)> lift:(1.14) lev:(0.02) [18] conv:(1.47)

31. Saving_status=<100 Employment=1<=X<4 210 ==> Age=young 168 <conf:(0.8)> lift:(1.14) lev:(0.02) [21] conv:(1.47)

32. Credit_history=existing paid Credit_amount=low Job=skilled 150 ==> Age=young 120 <conf:(0.8)> lift:(1.14) lev:(0.01) [15] conv:(1.45)

Apriori experiment 5

=== Run information ===

Scheme: weka.associations.Apriori -N 50 -T 0 -C 0.75 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: Copy of credit-g-attr10cleanednominal2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1

Instances: 1000

Attributes: 8

Purpose

Credit_amount

Saving_status

Employment

Personal_status

Age

Job

Class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (100 instances)

Minimum metric <confidence>: 0.75

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 22

Size of set of large itemsets L(2): 94

Size of set of large itemsets L(3): 85

Size of set of large itemsets L(4): 29

Size of set of large itemsets L(5): 1

Best rules found:

1. Saving_status=<100 Personal_status=female div/dep/mar Job=skilled 123 ==> Age=young 108 <conf:(0.88)> lift:(1.25) lev:(0.02) [21] conv:(2.31)

2. Saving_status=no known savings Personal_status=male single 116 ==> Class=good 101 <conf:(0.87)> lift:(1.24) lev:(0.02) [19] conv:(2.17)

3. Saving_status=<100 Employment=<1 120 ==> Age=young 104 <conf:(0.87)> lift:(1.24) lev:(0.02) [20] conv:(2.12)

4. Employment=<1 172 ==> Age=young 148 <conf:(0.86)> lift:(1.23) lev:(0.03) [27] conv:(2.06)

5. Personal_status=female div/dep/mar Job=skilled 196 ==> Age=young 167 <conf:(0.85)> lift:(1.22) lev:(0.03) [29] conv:(1.96)

6. Employment=4<=X<7 Job=skilled 119 ==> Age=young 100 <conf:(0.84)> lift:(1.2) lev:(0.02) [16] conv:(1.79)

7. Credit_amount=medium Employment=1<=X<4 Job=skilled 131 ==> Age=young 110 <conf:(0.84)> lift:(1.2) lev:(0.02) [18] conv:(1.79)

8. Credit_amount=medium Employment=1<=X<4 Class=good 121 ==> Age=young 101 <conf:(0.83)> lift:(1.19) lev:(0.02) [16] conv:(1.73)

9. Saving_status=<100 Employment=1<=X<4 Job=skilled 140 ==> Age=young 116 <conf:(0.83)> lift:(1.18) lev:(0.02) [18] conv:(1.68)

10. Saving_status=no known savings 183 ==> Class=good 151 <conf:(0.83)> lift:(1.18) lev:(0.02) [22] conv:(1.66)

11. Employment=1<=X<4 Job=skilled Class=good 159 ==> Age=young 131 <conf:(0.82)> lift:(1.18) lev:(0.02) [19] conv:(1.64)

12. Employment=1<=X<4 Job=skilled 229 ==> Age=young 187 <conf:(0.82)> lift:(1.17) lev:(0.03) [26] conv:(1.6)

13. Personal_status=female div/dep/mar Job=skilled Class=good 130 ==> Age=young 106 <conf:(0.82)> lift:(1.16) lev:(0.01) [15] conv:(1.56)

14. Credit_amount=medium Employment=1<=X<4 177 ==> Age=young 144 <conf:(0.81)> lift:(1.16) lev:(0.02) [20] conv:(1.56)

15. Employment=1<=X<4 Personal_status=male single Class=good 128 ==> Age=young 104 <conf:(0.81)> lift:(1.16) lev:(0.01) [14] conv:(1.54)

16. Employment=4<=X<7 174 ==> Age=young 141 <conf:(0.81)> lift:(1.16) lev:(0.02) [19] conv:(1.54)

17. Purpose=radio/tv Saving_status=<100 169 ==> Age=young 136 <conf:(0.8)> lift:(1.15) lev:(0.02) [17] conv:(1.49)

18. Purpose=radio/tv Personal_status=male single 146 ==> Class=good 117 <conf:(0.8)> lift:(1.14) lev:(0.01) [14] conv:(1.46)

19. Saving_status=<100 Employment=1<=X<4 210 ==> Age=young 168 <conf:(0.8)> lift:(1.14) lev:(0.02) [21] conv:(1.47)

20. Saving_status=<100 Personal_status=female div/dep/mar 194 ==> Age=young 155 <conf:(0.8)> lift:(1.14) lev:(0.02) [19] conv:(1.46)

21. Purpose=radio/tv Credit_amount=low 151 ==> Class=good 120 <conf:(0.79)> lift:(1.14) lev:(0.01) [14] conv:(1.42)

22. Saving_status=<100 Employment=1<=X<4 Class=good 140 ==> Age=young 111 <conf:(0.79)> lift:(1.13) lev:(0.01) [12] conv:(1.4)

23. Credit_amount=medium Personal_status=female div/dep/mar 141 ==> Age=young 111 <conf:(0.79)> lift:(1.12) lev:(0.01) [12] conv:(1.36)

24. Credit_amount=low Saving_status=<100 Job=skilled 153 ==> Age=young 120 <conf:(0.78)> lift:(1.12) lev:(0.01) [12] conv:(1.35)

25. Job=skilled Class=bad 185 ==> Age=young 145 <conf:(0.78)> lift:(1.12) lev:(0.02) [15] conv:(1.35)

26. Employment=1<=X<4 339 ==> Age=young 265 <conf:(0.78)> lift:(1.12) lev:(0.03) [27] conv:(1.36)

27. Credit_amount=medium Personal_status=male single Age=young 214 ==> Class=good 167 <conf:(0.78)> lift:(1.11) lev:(0.02) [17] conv:(1.34)

28. Credit_amount=medium Personal_status=male single Age=young Job=skilled 150 ==> Class=good 117 <conf:(0.78)> lift:(1.11) lev:(0.01) [12] conv:(1.32)

29. Employment=1<=X<4 Class=good 235 ==> Age=young 183 <conf:(0.78)> lift:(1.11) lev:(0.02) [18] conv:(1.33)

30. Purpose=radio/tv 280 ==> Class=good 218 <conf:(0.78)> lift:(1.11) lev:(0.02) [21] conv:(1.33)

31. Purpose=radio/tv Job=skilled 194 ==> Class=good 151 <conf:(0.78)> lift:(1.11) lev:(0.02) [15] conv:(1.32)

32. Employment=>=7 Job=skilled 162 ==> Class=good 126 <conf:(0.78)> lift:(1.11) lev:(0.01) [12] conv:(1.31)

33. Employment=4<=X<7 Class=good 135 ==> Age=young 105 <conf:(0.78)> lift:(1.11) lev:(0.01) [10] conv:(1.31)

34. Personal_status=female div/dep/mar 310 ==> Age=young 241 <conf:(0.78)> lift:(1.11) lev:(0.02) [23] conv:(1.33)

35. Employment=1<=X<4 Personal_status=male single Age=young 134 ==> Class=good 104 <conf:(0.78)> lift:(1.11) lev:(0.01) [10] conv:(1.3)

36. Employment=4<=X<7 174 ==> Class=good 135 <conf:(0.78)> lift:(1.11) lev:(0.01) [13] conv:(1.31)

37. Saving_status=<100 Job=skilled Class=bad 134 ==> Age=young 103 <conf:(0.77)> lift:(1.1) lev:(0.01) [9] conv:(1.26)

38. Purpose=radio/tv Job=skilled 194 ==> Age=young 149 <conf:(0.77)> lift:(1.1) lev:(0.01) [13] conv:(1.27)

39. Credit_amount=low Personal_status=female div/dep/mar 159 ==> Age=young 122 <conf:(0.77)> lift:(1.1) lev:(0.01) [10] conv:(1.26)

40. Employment=1<=X<4 Personal_status=male single 175 ==> Age=young 134 <conf:(0.77)> lift:(1.09) lev:(0.01) [11] conv:(1.25)

41. Credit_amount=medium Employment=1<=X<4 Age=young 144 ==> Job=skilled 110 <conf:(0.76)> lift:(1.22) lev:(0.02) [19] conv:(1.53)

42. Saving_status=<100 Job=skilled 364 ==> Age=young 278 <conf:(0.76)> lift:(1.09) lev:(0.02) [23] conv:(1.26)

43. Credit_amount=medium Job=skilled Class=good 247 ==> Age=young 188 <conf:(0.76)> lift:(1.09) lev:(0.02) [15] conv:(1.23)

44. Saving_status=<100 Job=skilled Class=good 230 ==> Age=young 175 <conf:(0.76)> lift:(1.09) lev:(0.01) [13] conv:(1.23)

45. Employment=>=7 Job=skilled 162 ==> Personal_status=male single 123 <conf:(0.76)> lift:(1.39) lev:(0.03) [34] conv:(1.83)

46. Credit_amount=low Personal_status=male single 196 ==> Class=good 148 <conf:(0.76)> lift:(1.08) lev:(0.01) [10] conv:(1.2)

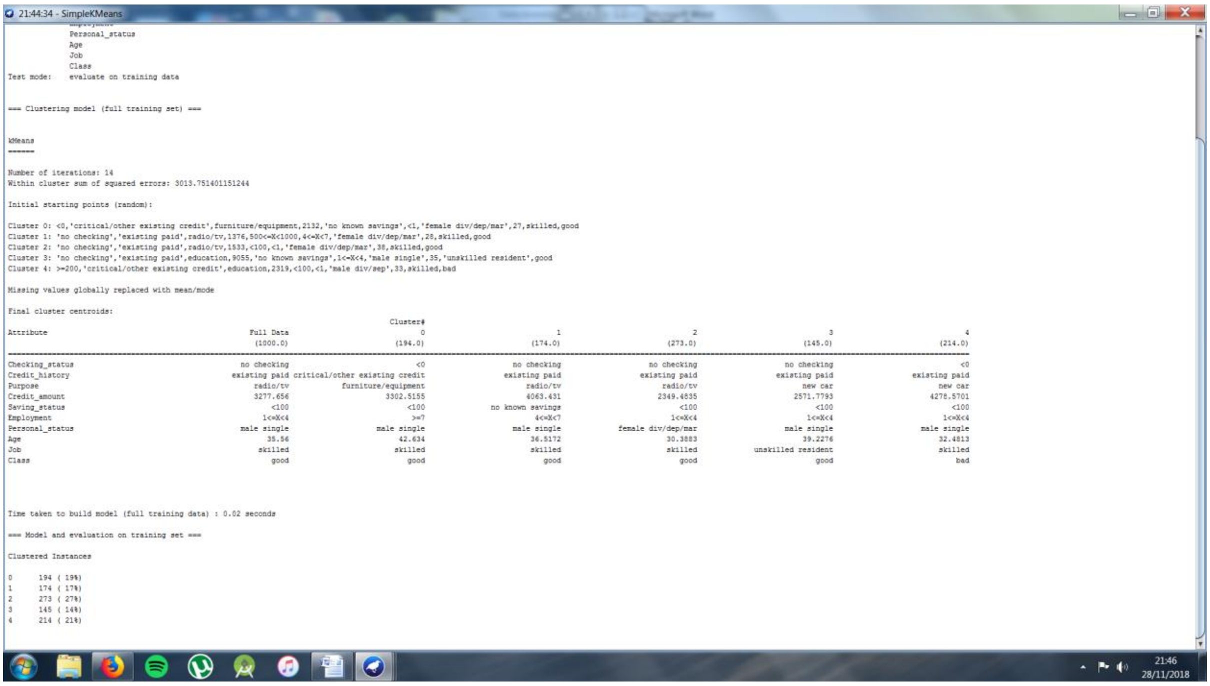
47. Credit_amount=medium Class=bad 150 ==> Saving_status=<100 113 <conf:(0.75)> lift:(1.25) lev:(0.02) [22] conv:(1.57)

48. Purpose=radio/tv Saving_status=<100 169 ==> Class=good 127 <conf:(0.75)> lift:(1.07) lev:(0.01) [8] conv:(1.18)

49. Purpose=furniture/equipment 181 ==> Age=young 136 <conf:(0.75)> lift:(1.07) lev:(0.01) [9] conv:(1.18)

50. Employment=>=7 Personal_status=male single 181 ==> Class=good 136 <conf:(0.75)> lift:(1.07) lev:(0.01) [9] conv:(1.18)

K means experiment 1 seed 50 5 clusters



```
21:44:34 - SimpleKMeans
              Personal_status
              Age
              Job
              Class
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 14
Within cluster sum of squared errors: 3013.751401151244

Initial starting points (random):

Cluster 0: <0,'critical/other existing credit',furniture/equipment,2132,'no known savings',<1,'female div/dep/mar',27,skilled,good
Cluster 1: 'no checking','existing paid',radio/tv,1376,500<=X<1000,4<=X<7,'female div/dep/mar',28,skilled,good
Cluster 2: 'no checking','existing paid',radio/tv,1533,<100,<1,'female div/dep/mar',38,skilled,good
Cluster 3: 'no checking','existing paid',education,9055,'no known savings',1<=X<4,'male single',35,'unskilled resident',good
Cluster 4: >=200,'critical/other existing credit',education,2319,<100,<1,'male div/sep',33,skilled,bad

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                      Full Data                       0                 1                 2                  3              4
                                (1000.0)                 (194.0)           (174.0)           (273.0)            (145.0)        (214.0)
=======================================================================================================================================
Checking_status              no checking                      <0       no checking       no checking        no checking             <0
Credit_history             existing paid critical/other existing credit  existing paid  existing paid   existing paid  existing paid
Purpose                         radio/tv     furniture/equipment          radio/tv          radio/tv            new car        new car
Credit_amount                   3277.656               3302.5155          4063.431         2349.4835          2571.7793      4278.5701
Saving_status                       <100                    <100  no known savings              <100               <100           <100
Employment                        1<=X<4                     >=7            4<=X<7            1<=X<4             1<=X<4         1<=X<4
Personal_status              male single             male single       male single female div/dep/mar       male single    male single
Age                                35.56                  42.634           36.5172           30.3883            39.2276        32.4813
Job                              skilled                 skilled           skilled           skilled unskilled resident        skilled
Class                               good                    good              good              good               good            bad




Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       194 ( 19%)
1       174 ( 17%)
2       273 ( 27%)
3       145 ( 14%)
4       214 ( 21%)
```

K means experiment 2 4 clusters

22:01:09 - SimpleKMeans

```
             Saving_status
             Employment
             Personal_status
             Age
             Job
             Class
Test mode:   evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 11
Within cluster sum of squared errors: 3227.832822640441

Initial starting points (random):

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

Missing values globally replaced with mean/mode

Final cluster centroids:
                                              Cluster#
Attribute                 Full Data                  0                              1                  2                  3
                           (1000.0)            (306.0)                        (148.0)            (259.0)            (287.0)
=============================================================================================================================
Checking_status         no checking                 <0                             <0           0<=X<200        no checking
Credit_history        existing paid      existing paid critical/other existing credit      existing paid      existing paid
Purpose                    radio/tv            new car                       used car           radio/tv           radio/tv
Credit_amount              3277.656          3486.0588                      4697.8311            2474.39               3048
Saving_status                  <100               <100                           <100               <100   no known savings
Employment                   1<=X<4             1<=X<4                            >=7             1<=X<4                >=7
Personal_status         male single female div/dep/mar                    male single        male single        male single
Age                           35.56            30.6569                        43.8041            33.0116            38.8362
Job                         skilled            skilled      high qualif/self emp/mgmt unskilled resident            skilled
Class                          good                bad                           good               good               good




Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       306 ( 31%)
1       148 ( 15%)
2       259 ( 26%)
3       287 ( 29%)
```

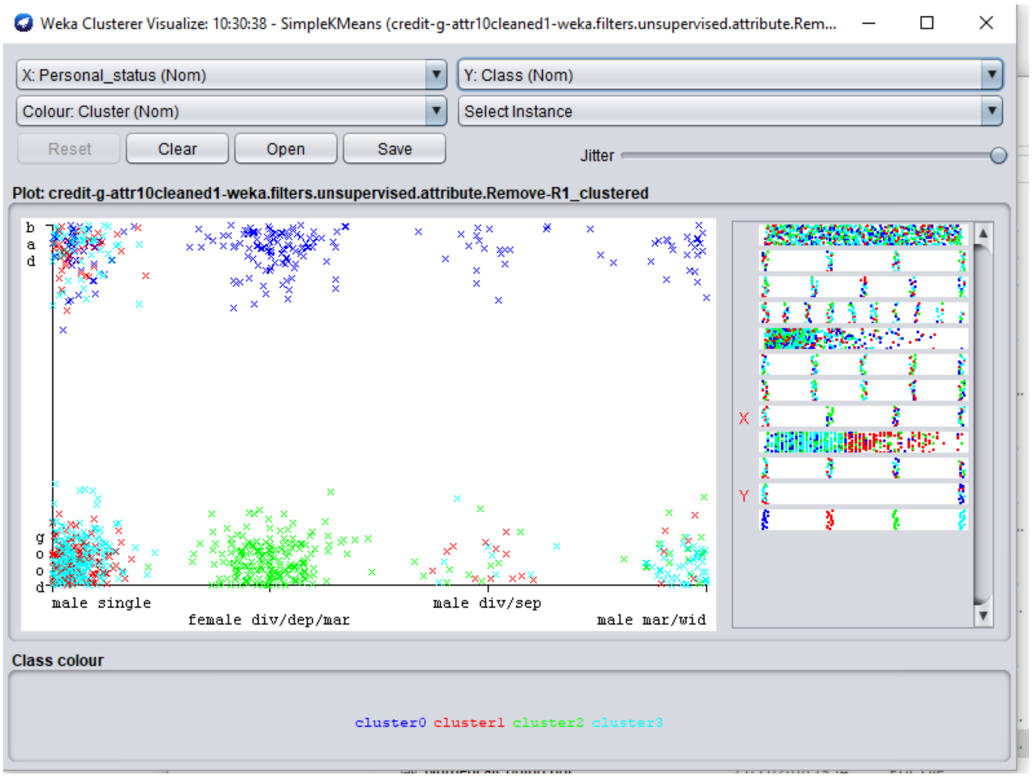
22:05
28/11/2018

K means experiment 3 4 clusters

10:30:38 - SimpleKMeans

```
Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1000
Attributes:   10
              Credit_amount
              Personal_status
              Age
              Class
Ignored:
              Checking_status
              Credit_history
              Purpose
              Saving_status
              Employment
              Job
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 21
Within cluster sum of squared errors: 333.6361967513323

Initial starting points (random):

Cluster 0: 7855,'female div/dep/mar',25,bad
Cluster 1: 6615,'male single',75,good
Cluster 2: 1155,'male mar/wid',40,good
Cluster 3: 754,'male single',38,good

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute                Full Data           0           1                  2           3
                          (1000.0)     (197.0)     (187.0)            (228.0)     (388.0)
==========================================================================================
Credit_amount             3277.656   4311.5939   3571.6738          2498.8026   3068.6649
Personal_status        male single female div/dep/mar male single female div/dep/mar male single
Age                          35.56     30.8579     51.1979            34.5395     31.0103
Class                         good         bad        good               good        good
```

10:35



K means 4 10 clusters

10:49:50 - SimpleKMeans

```
Number of iterations: 11
Within cluster sum of squared errors: 108.22604158267255

Initial starting points (random):

Cluster 0: 7855,'female div/dep/mar',25,bad
Cluster 1: 6615,'male single',75,good
Cluster 2: 1155,'male mar/wid',40,good
Cluster 3: 754,'male single',38,good
Cluster 4: 3966,'female div/dep/mar',33,bad
Cluster 5: 753,'female div/dep/mar',64,good
Cluster 6: 1413,'male single',55,good
Cluster 7: 362,'female div/dep/mar',52,good
Cluster 8: 2670,'male single',35,good
Cluster 9: 1346,'male single',42,good

Missing values globally replaced with mean/mode

Final cluster centroids:
                                         Cluster#
Attribute               Full Data               0               1               2               3               4               5               6               7               8
                         (1000.0)         (162.0)          (30.0)          (86.0)         (188.0)         (124.0)          (59.0)          (34.0)         (149.0)          (66.0)
==============================================================================================================================================================================
Credit_amount            3277.656       4540.0309      10109.7667       1825.9535        2237.383        3421.629       2131.4746       2562.2647       2731.3691       5893.2576
Personal_status       male single     male single     male single    male mar/wid     male single female div/dep/mar female div/dep/mar     male single female div/dep/mar     male single        mal
Age                         35.56         37.7654         40.8333         29.6512         30.8351         29.8871         51.5763         63.5588         27.7248         32.7879
Class                        good             bad            good            good            good             bad            good            good            good            good




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0      162 ( 16%)
 1       30 (  3%)
 2       86 (  9%)
 3      188 ( 19%)
 4      124 ( 12%)
 5       59 (  6%)
 6       34 (  3%)
 7      149 ( 15%)
 8       66 (  7%)
 9      102 ( 10%)
```

K means 5 10 attribute

10:55:05 - SimpleKMeans

```
Initial starting points (random):

Cluster 0: 'critical/other existing credit',7855,'female div/dep/mar',25,bad
Cluster 1: 'critical/other existing credit',6615,'male single',75,good
Cluster 2: 'existing paid',1155,'male mar/wid',40,good
Cluster 3: 'existing paid',754,'male single',38,good
Cluster 4: 'critical/other existing credit',3966,'female div/dep/mar',33,bad
Cluster 5: 'existing paid',753,'female div/dep/mar',64,good
Cluster 6: 'existing paid',1413,'male single',55,good
Cluster 7: 'critical/other existing credit',362,'female div/dep/mar',52,good
Cluster 8: 'existing paid',2670,'male single',35,good
Cluster 9: 'existing paid',1346,'male single',42,good

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                       Cluster#
Attribute                   Full Data                         0                              1               2               3                  4
                             (1000.0)                    (43.0)                        (183.0)          (77.0)         (157.0)             (95.0)
=================================================================================================================================================
Credit_history          existing paid critical/other existing credit critical/other existing credit   existing paid   existing paid      existing paid
Credit_amount                3277.656                 5804.9302                      3071.8962       1816.2987       2475.4586          2701.6316
Personal_status           male single        female div/dep/mar                    male single    male mar/wid     male single female div/dep/mar
Age                             35.56                   34.2326                        41.5301         28.5714         31.1274            27.9895
Class                            good                       bad                           good            good            good                bad




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0       43 (  4%)
 1      183 ( 18%)
 2       77 (  8%)
 3      157 ( 16%)
 4       95 ( 10%)
 5      136 ( 14%)
 6      140 ( 14%)
 7       76 (  8%)
 8       45 (  5%)
 9       48 (  5%)
```

Kmeans 6 5 attribute seed 10

```
              Age
              Job
              Class
Ignored:
              Credit_history
              Credit_amount
              Saving_status
              Personal_status
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1858.6373302936158

Initial starting points (random):

Cluster 0: 'no checking','new car',1<=X<4,25,skilled,bad
Cluster 1: <0,'used car',unemployed,75,'high qualif/self emp/mgmt',good
Cluster 2: 0<=X<200,radio/tv,>=7,40,'unskilled resident',good
Cluster 3: 0<=X<200,retraining,>=7,38,skilled,good
Cluster 4: <0,'new car',>=7,33,skilled,bad

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                Cluster#
Attribute                   Full Data                  0                         1                  2            3            4
                             (1000.0)            (357.0)                    (93.0)            (142.0)      (210.0)      (198.0)
===============================================================================================================================
Checking_status           no checking        no checking                        <0        no checking     0<=X<200           <0
Purpose                      radio/tv            new car                  used car           radio/tv     radio/tv      new car
Employment                     1<=X<4             1<=X<4                unemployed                >=7          >=7          >=7
Age                             35.56            31.7899                   41.1505             44.993      34.9381      33.6263
Job                           skilled            skilled high qualif/self emp/mgmt unskilled resident      skilled      skilled
Class                            good               good                      good               good         good          bad




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       357 ( 36%)
1        93 (  9%)
2       142 ( 14%)
3       210 ( 21%)
4       198 ( 20%)
```

Kmeans 7

```
              Age
              Job
              Class
Ignored:
              Checking_status
              Credit_history
              Credit_amount
              Saving_status
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 1777.8123008497948

Initial starting points (random):

Cluster 0: 'new car',1<=X<4,'female div/dep/mar',25,skilled,bad
Cluster 1: 'used car',unemployed,'male single',75,'high qualif/self emp/mgmt',good
Cluster 2: radio/tv,>=7,'male mar/wid',40,'unskilled resident',good
Cluster 3: retraining,>=7,'male single',38,skilled,good
Cluster 4: 'new car',>=7,'female div/dep/mar',33,skilled,bad

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute            Full Data                 0                          1                    2            3            4
                      (1000.0)           (256.0)                     (96.0)              (188.0)      (332.0)      (128.0)
==========================================================================================================================
Purpose               radio/tv           new car                   used car             radio/tv     radio/tv      new car
Employment              1<=X<4            1<=X<4                 unemployed               1<=X<4          >=7          >=7
Personal_status    male single female div/dep/mar               male single          male single  male single  male single
Age                      35.56           28.0508                    42.5833              38.0798      34.5542      44.2188
Job                    skilled           skilled high qualif/self emp/mgmt unskilled resident      skilled      skilled
Class                     good               bad                       good                 good         good          bad




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      256 ( 26%)
1       96 ( 10%)
2      188 ( 19%)
3      332 ( 33%)
4      128 ( 13%)
```

EM experiment 1

10:34:18 - EM

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1000
Attributes:   10
              Credit_amount
              Personal_status
              Age
              Class
Ignored:
              Checking_status
              Credit_history
              Purpose
              Saving_status
              Employment
              Job
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


EM
==

Number of clusters selected by cross validation: 10
Number of iterations performed: 100
```


Attribute	Cluster 0	1	2	3	4	5	6	7	8	9
	(0.07)	(0.12)	(0.1)	(0.11)	(0.02)	(0.1)	(0.13)	(0.17)	(0.04)	(0.13)
Credit_amount										
mean	6386.8008	2717.6895	6505.5951	2575.9303	13825.3359	1111.0686	1469.5755	2905.6607	6535.3643	1174.3218
std. dev.	2382.5878	975.1509	2463.9121	872.7375	1835.7777	380.3868	444.3521	941.9585	1324.4795	481.7894
Personal_status										
male single	32.2444	32.7023	88.4721	76.4359	16.8133	18.8542	58.7633	131.5839	26.8723	76.6582
female div/dep/mar	34.2652	73.9889	11.1386	23.8888	5.5063	55.579	41.2231	22.3622	10.8894	42.0753
male div/sep	2.4231	2.7582	5.0109	9.2487	2.9953	2.4507	8.6248	15.7211	5.1364	5.6388
male mar/wid	10.0524	14.7756	3.1217	4.1689	1.0035	28.1998	25.0617	2.769	1.6136	11.2417
[total]	78.9852	124.225	107.7434	112.8933	26.3185	105.0838	133.673	172.4363	43.6276	135.014
Age										
mean	26.4992	24.5959	36.9846	48.0081	39.5627	24.6682	32.7926	33.3055	51.535	47.8548
std. dev.	2.9266	2.6217	6.1763	9.9887	13.8611	2.4987	4.6392	4.9396	10.9239	10.8699

Class

10:34:18 - EM

```
Number of clusters selected by cross validation: 10
Number of iterations performed: 100
```

Attribute	Cluster 0	1	2	3	4	5	6	7	8	9
	(0.07)	(0.12)	(0.1)	(0.11)	(0.02)	(0.1)	(0.13)	(0.17)	(0.04)	(0.13)
Credit_amount										
mean	6386.8008	2717.6895	6505.5951	2575.9303	13825.3359	1111.0686	1469.5755	2905.6607	6535.3643	1174.3218
std. dev.	2382.5878	975.1509	2463.9121	872.7375	1835.7777	380.3868	444.3521	941.9585	1324.4795	481.7894
Personal_status										
male single	32.2444	32.7023	88.4721	76.4359	16.8133	18.8542	58.7633	131.5839	26.8723	76.6582
female div/dep/mar	34.2652	73.9889	11.1386	23.8888	5.5063	55.579	41.2231	22.3622	10.8894	42.0753
male div/sep	2.4231	2.7582	5.0109	9.2487	2.9953	2.4507	8.6248	15.7211	5.1364	5.6388
male mar/wid	10.0524	14.7756	3.1217	4.1689	1.0035	28.1998	25.0617	2.769	1.6136	11.2417
[total]	78.9852	124.225	107.7434	112.8933	26.3185	105.0838	133.673	172.4363	43.6276	135.014
Age										
mean	26.4992	24.5959	36.9846	48.0081	39.5627	24.6682	32.7926	33.3055	51.535	47.8548
std. dev.	2.9266	2.6217	6.1763	9.9887	13.8611	2.4987	4.6392	4.9396	10.9239	10.8699
Class										
good	36.2585	79.772	84.715	91.3714	6.6648	64.6388	87.8903	131.913	17.0982	109.6893
bad	40.7266	42.453	21.0283	19.5219	17.654	38.4481	43.7827	38.5233	24.5374	23.3247
[total]	76.9852	122.225	105.7434	110.8933	24.3185	103.0838	131.673	170.4363	41.6276	133.014

```
Time taken to build model (full training data) : 25.85 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       73 (  7%)
1      124 ( 12%)
2       83 (  8%)
3       83 (  8%)
4       22 (  2%)
5      127 ( 13%)
6      140 ( 14%)
7      186 ( 19%)
8       32 (  3%)
9      130 ( 13%)


Log likelihood: -14.16701
```

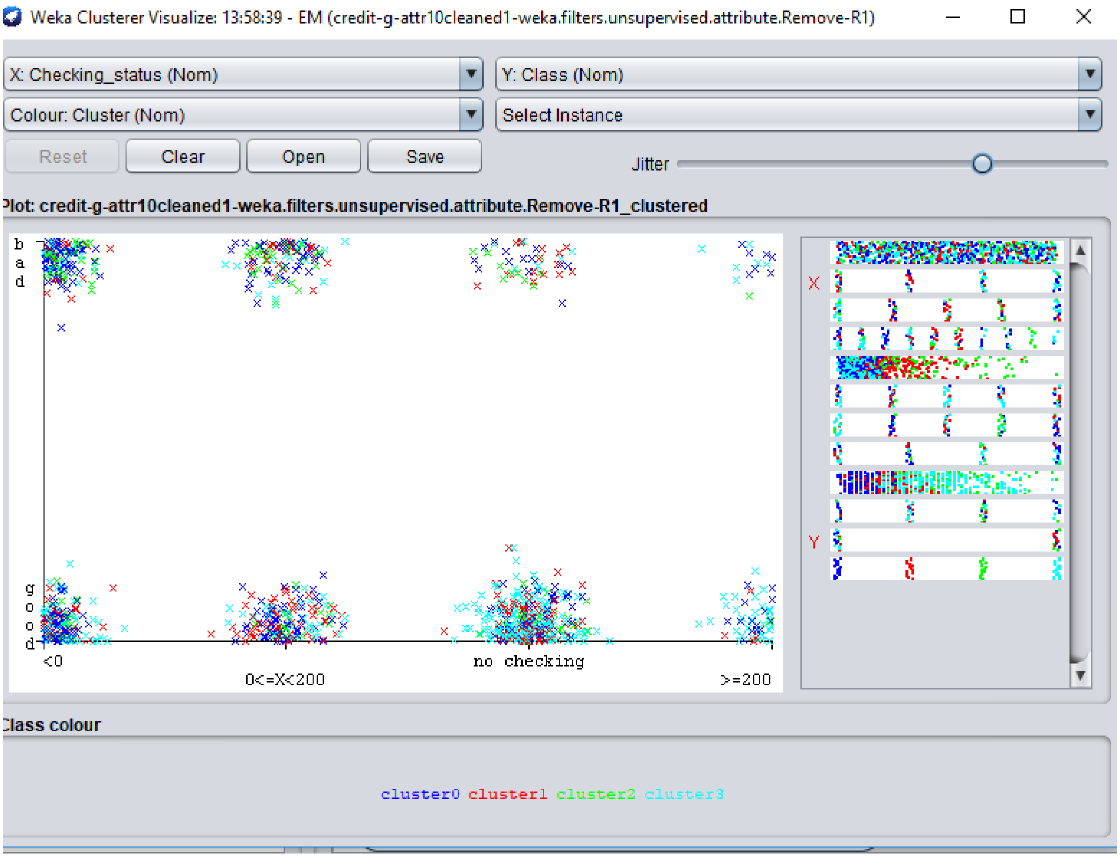
Em experiment 2 4 clusters

13:58:39 - EM

```
                                      Cluster
Attribute                                   0         1         2         3
                                       (0.29)    (0.23)    (0.14)    (0.34)
============================================================================
Checking_status
  <0                                 106.4862   43.9326   55.4737   72.1076
  0<=X<200                            88.4681    74.631   53.6298    56.271
  no checking                         75.2574  107.7598   35.1863  179.7965
  >=200                               19.9824    8.2117    3.0322   35.7736
  [total]                             290.194  234.5352  147.3221  343.9487
Credit_history
  critical/other existing credit      39.0341   64.0057   47.2551   146.705
  existing paid                      212.3341  101.0781   58.7726  161.8152
  delayed previously                  10.4116   45.8659   16.5444    19.178
  no credits/all paid                  8.5756   18.8737   12.6251    3.9255
  all paid                            20.8385    5.7116   13.1249   13.3249
  [total]                             291.194  235.5352  148.3221  344.9487
Purpose
  radio/tv                            93.3337    54.758    2.9611  132.9472
  education                            9.4729   12.7784     8.181   23.5677
  furniture/equipment                 78.9776   36.2342   23.6837   46.1045
  new car                             65.4539   36.8204    41.031   94.6947
  used car                             7.4624   49.4057   34.9763   15.1556
  business                            18.7413   43.6538   22.4426   16.1623
  domestic appliance                   7.6296    1.1197    1.2527     5.998
  repairs                              9.9306    2.6019    7.2062    6.2612
  other                                1.5481    1.7952   10.3679    2.2887
  retraining                            3.644    1.3678    1.2195    6.7687
  [total]                             296.194  240.5352  153.3221  349.9487
Credit_amount
  mean                              1926.6309 5077.3178 6563.2371 1809.4153
  std. dev.                          976.5201 2574.3184 4004.2072  866.6382

Saving_status
  no known savings                    32.2361   51.9657   32.5098   70.2883
  <100                               200.3169  121.7581  102.6813  182.2437
  500<=X<1000                          16.829   10.3192    1.6972   38.1546
  >=1000                              10.6063   11.2446    2.7616   27.3875
  100<=X<500                          31.2057   40.2475    8.6722   26.8746
  [total]                             291.194  235.5352  148.3221  344.9487
Employment
  >=7                                  9.6447   33.5343   61.5209  152.3001
  1<=X<4                             123.5639    87.951   28.3164  103.1687
  4<=X<7                              44.0359   70.5031   11.3765   52.0845
  unemployed                          12.3206    3.5064   40.0643   10.1087
  <1                                 101.6289   40.0405     7.044   27.2866
  [total]                             291.194  235.5352  148.3221  344.9487
Personal_status
```

13:58:39 - EM

```
  <100                               200.3169  121.7581  102.6813  182.2437
  500<=X<1000                          16.829   10.3192    1.6972   38.1546
  >=1000                              10.6063   11.2446    2.7616   27.3875
  100<=X<500                          31.2057   40.2475    8.6722   26.8746
  [total]                             291.194  235.5352  148.3221  344.9487
Employment
  >=7                                  9.6447   33.5343   61.5209  152.3001
  1<=X<4                             123.5639    87.951   28.3164  103.1687
  4<=X<7                              44.0359   70.5031   11.3765   52.0845
  unemployed                          12.3206    3.5064   40.0643   10.1087
  <1                                 101.6289   40.0405     7.044   27.2866
  [total]                             291.194  235.5352  148.3221  344.9487
Personal_status
  male single                         78.7787  154.2397   101.284  217.6977
  female div/dep/mar                 148.8291   56.0038   28.8287   80.3384
  male div/sep                        11.6166   10.9518   12.7811   18.6505
  male mar/wid                        50.9696   13.3399    4.4284    27.262
  [total]                             290.194  234.5352  147.3221  343.9487
Age
  mean                                26.2256   31.7894   43.8366    42.486
  std. dev.                            3.7662    6.1506    12.779   10.7842

Job
  skilled                              199.02  163.7709   54.5394  214.6696
  unskilled resident                  66.2842   35.4429    6.9787   97.2941
  high qualif/self emp/mgmt           14.9791    34.189   77.2012   25.6307
  unemp/unskilled non res              9.9106    1.1323    8.6028    6.3542
  [total]                             290.194  234.5352  147.3221  343.9487
Class
  good                               169.0599   173.987   65.6584  295.2947
  bad                                119.1341   58.5481   79.6637    46.654
  [total]                             288.194  232.5352  145.3221  341.9487


Time taken to build model (full training data) : 10.99 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       303 ( 30%)
1       218 ( 22%)
2       136 ( 14%)
3       343 ( 34%)


Log likelihood: -21.90924
```



EM experiment 3

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N 4 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     credit-g-attr10cleaned1-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1000
Attributes:   10
              Checking_status
              Credit_history
              Credit_amount
              Saving_status
              Class
Ignored:
              Purpose
              Employment
              Personal_status
              Age
              Job
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


EM
==

Number of clusters: 4
Number of iterations performed: 100


                                  Cluster
Attribute                                0         1         2         3
                                    (0.23)    (0.27)    (0.23)    (0.28)
=========================================================================
Checking_status
  <0                               62.0119   35.9405  135.5444   44.5031
  0<=X<200                         91.1655   49.8442   71.1852   60.8051
  no checking                      77.8999  162.9697    8.6251  148.5053
  >=200                             2.8617   21.2038   15.5019   27.4327
  [total]                          233.939  269.9582  230.8566  281.2462
Credit_history
  critical/other existing credit   61.5054   97.2808   26.3881  111.8257
  existing paid                    94.6628  142.5716  151.6348  145.1309
  delayed previously               41.5333    20.975   14.0124   15.4792
  no credits/all paid              23.4709    3.4314    14.237    2.8606
  all paid                         13.7665    6.6994   25.5843    6.9498
  [total]                          234.939  270.9582  231.8566  282.2462
Credit_amount
  mean                            7130.254 2984.6725 2085.7773 1338.7452
  std. dev.                      3255.4191 1160.3074 1056.6384  443.8413

Saving_status
  no known savings                 56.8426   69.5492   14.9989   45.6093
  <100                            137.9912  127.8693   182.343  158.7964
  500<=X<1000                       7.1072   27.3055    6.6487   25.9386
  >=1000                            5.9132   18.7247    3.5891   23.7731
  100<=X<500                       27.0847   27.5096   24.2769   28.1288
  [total]                          234.939  270.9582  231.8566  282.2462
```


```
EM
==

Number of clusters: 4
Number of iterations performed: 100


                                  Cluster
Attribute                                0         1         2         3
                                    (0.23)    (0.27)    (0.23)    (0.28)
=========================================================================
Checking_status
  <0                               62.0119   35.9405  135.5444   44.5031
  0<=X<200                         91.1655   49.8442   71.1852   60.8051
  no checking                      77.8999  162.9697    8.6251  148.5053
  >=200                             2.8617   21.2038   15.5019   27.4327
  [total]                          233.939  269.9582  230.8566  281.2462
Credit_history
  critical/other existing credit   61.5054   97.2808   26.3881  111.8257
  existing paid                    94.6628  142.5716  151.6348  145.1309
  delayed previously               41.5333    20.975   14.0124   15.4792
  no credits/all paid              23.4709    3.4314    14.237    2.8606
  all paid                         13.7665    6.6994   25.5843    6.9498
  [total]                          234.939  270.9582  231.8566  282.2462
Credit_amount
  mean                            7130.254 2984.6725 2085.7773 1338.7452
  std. dev.                      3255.4191 1160.3074 1056.6384  443.8413

Saving_status
  no known savings                 56.8426   69.5492   14.9989   45.6093
  <100                            137.9912  127.8693   182.343  158.7964
  500<=X<1000                       7.1072   27.3055    6.6487   25.9386
  >=1000                            5.9132   18.7247    3.5891   23.7731
  100<=X<500                       27.0847   27.5096   24.2769   28.1288
  [total]                          234.939  270.9582  231.8566  282.2462
```


```
Class
  good                            123.6041  257.2589   75.1331  248.0039
  bad                             108.3349   10.6993  153.7234   31.2423
  [total]                          231.939  267.9582  228.8566  279.2462


Time taken to build model (full training data) : 0.4 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       205 ( 21%)
1       249 ( 25%)
2       224 ( 22%)
3       322 ( 32%)


Log likelihood: -13.03833
```

EM experiment 4

```
                            Cluster
Attribute                         0        1        2        3
                              (0.1)   (0.34)   (0.32)   (0.24)
===============================================================
Purpose
  radio/tv                    7.292  48.6422  142.887  85.1788
  education                  4.4519  27.5044   9.3307  12.7131
  furniture/equipment        15.284  61.5297  43.6597  64.5265
  new car                   32.2239 132.2713  31.5192  41.9856
  used car                  21.4509   6.3827  68.9209  10.2455
  business                    9.287  57.1817  23.4746  11.0567
  domestic appliance         1.2211   3.1051   3.2872   8.3866
  repairs                     3.257  10.3116   5.7128   6.7186
  other                     10.7567   2.0831   2.0265   1.1337
  retraining                 2.0293   3.0819   3.1112   4.7777
  [total]                  107.2539 352.0936 333.9298 246.7226
Employment
  >=7                       28.3644  83.9164 138.3591   6.3601
  1<=X<4                     9.9653 141.8261  93.9568  97.2519
  4<=X<7                     3.4641  66.6372  78.5162  29.3826
  unemployed                51.5206   3.0085   2.4902   8.9807
  <1                         8.9395  51.7055  15.6076  99.7474
  [total]                  102.2539 347.0936 328.9298 241.7226
Personal_status
  male single               56.3764 231.0844 251.1031  13.4361
  female div/dep/mar        32.5121  73.7509  44.2616 163.4755
  male div/sep                 7.21  24.3244   7.1935  15.2721
  male mar/wid               5.1554  16.9339  25.3716   48.539
  [total]                  101.2539 346.0936 327.9298 240.7226
Job
  skilled                    6.5926 209.8712 242.6283 172.9079
  unskilled resident         2.1267 110.9873  40.1935  52.6924
  high qualif/self emp/mgmt 73.4985  23.4457  44.0031  11.0528
  unemp/unskilled non res   19.0361   1.7894    1.105   4.0695
  [total]                  101.2539 346.0936 327.9298 240.7226
Class
  good                      55.6954 207.8381 292.9966 147.4699
  bad                       43.5585 136.2555  32.9332  91.2528
  [total]                   99.2539 344.0936 325.9298 238.7226


Time taken to build model (full training data) : 0.34 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       97 ( 10%)
1      308 ( 31%)
2      323 ( 32%)
3      272 ( 27%)


Log likelihood: -5.80546
```

EM experiment 5

```
=== Clustering model (full training set) ===


EM
==

Number of clusters: 4
Number of iterations performed: 100


                         Cluster
Attribute                      0        1        2        3
                          (0.16)   (0.23)   (0.29)   (0.32)
============================================================
Purpose
  radio/tv               17.7294   57.053  95.8455 113.3721
  education              18.4937  19.9097  10.6665     4.93
  furniture/equipment     35.568   30.023  75.3067  44.1023
  new car                55.5636  62.5337  53.4894  66.4133
  used car                3.1137  27.9751  23.0829  52.8282
  business               28.8011  17.5875  23.2776  31.3338
  domestic appliance      1.4113   4.0996   7.4233   3.0658
  repairs                 5.6849   9.9973   7.8557   2.4621
  other                    5.932   5.1333   1.3294   3.6054
  retraining              1.5597    1.307   2.5339   7.5994
  [total]               173.8574 235.6193 300.8109 329.7124
Personal_status
  male single            93.5144 146.2546  78.2467 233.9843
  female div/dep/mar     42.8656  59.9255 161.9224  49.2865
  male div/sep           27.1113  16.6995   4.8418   5.3473
  male mar/wid             4.366   6.7397  49.7999  35.0944
  [total]               167.8574 229.6193 294.8109 323.7124
Age
  mean                   33.9025  51.4879  25.0015  34.7732
  std. dev.               5.3256   9.7092   2.7364   6.0388

Class
  good                   65.0068 166.4288 175.9059 296.6585
  bad                   100.8506  61.1905  116.905  25.0539
  [total]               165.8574 227.6193 292.8109 321.7124


Time taken to build model (full training data) : 0.36 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      129 ( 13%)
1      206 ( 21%)
2      327 ( 33%)
3      338 ( 34%)


Log likelihood: -7.10093
```