SET09120 Data Analytics 2020/21

David Ciocoiu

40323308

1. Introduction

A dataset about credit applications recorded by a German bank has been provided.
We understand that banks profits due to the interests applied on their loans and that the business's main objective is to increase revenue.
A big issue lies between loan repayments and banks must decide whether to take the risk of awarding a loan. The purpose of the analysis is to provide interesting patterns through data mining techniques in order to allow the bank to make informed decision based on applicant's profiles.

2. Data Preparation

Data quality is paramount for cost reduction, increased efficiency and informed decisions (The cost of bad data, D.David, 2014); before undertaking any analysis, our dataset must be cleaned using the provided metadata and prepared for future model applications.
OpenRefine has been used to clean and transform the data.

2.1 Data Cleaning

The first step undertaken is to add missing headers to each column according to the supplied metadata:

case_no	checking_status	credit_history	purpose	credit_amount

saving_status	employment	persona_status	age	job	class

Then we detect any errors and outliers in the data; notice that the quotes in the various attributes have been removed for consistency in the data

Attribute Name	Error Value	Corrected Value	Info
checking_status			
	'<0'	<0	Consistency: Extra quotes
	'>=200'	>=200	Consistency: Extra quotes
	'0<=X<200'	0<=X<200	Consistency: Extra quotes
	'no checking'	no checking	Consistency: Extra quotes
credit_history			
	'all paid'	all paid	Consistency: Extra quotes
	'critical/other existing credit'	critical/other existing credit	Consistency: Extra quotes
	'delayed previously'	delayed previously	Consistency: Extra quotes
	'existing paid'	existing paid	Consistency: Extra quotes
	'no credits/all paid'	no credits/all paid	Consistency: Extra quotes
purpose			
	'domestic appliance'	domestic appliance	Consistency: Extra quotes
	'new car'	new car	Consistency: Extra quotes
	'used car'	used car	Consistency: Extra quotes

	ather	other	Misspelling
	busness	business	Misspelling
	busines	business	Misspelling
	Eduction	education	Misspelling, Capital letters
	Radio/Tv	radio/tv	Capital letters
credit_amount			
case_no: 432	111328000	13280	Outlier: compared against age > 27 and purpose 'other', new value seems to fit in.
case_no: 444	7190000	7190	Outlier: compared against age > 40, employment >= 7 and purpose 'education', new value seems to fit in.
case_no: 452	5180000	5189	Outlier: compared against: Age between 20 and 29; employment 1<=X<4 and purpose 'radio/tv', new value seems to fit in.
case_no: 514	5850000	5850	Outlier: compared against: Age between 20 and 20; employment 1<=X<4 and purpose 'radio/tv', new value seems to fit in.
case_no: 560	19280000	1928	Outlier: compared against: credit_history 'critical/other existing credit'; checking_status 0<=X<200 and purpose 'furniture/equipment, new value seems to fit in.
case_no: 595	13580000	13580	Outlier: compared against age > 27 and purpose 'other', new value seems to fit in.
case_no: 648	13860000	1386	Outlier: compared against: saving_status 500<=X1000, new value seems to fit in.
case_no: 660	63610000	6361	Outlier: compared against: credit_history 'critical/other existing credit'; checking_status 0<=X<200 and purpose 'furniture/equipment, new value seems to fit in.
saving_status			
	'<100'	<100	Consistency: Extra quotes
	'>=1000'	>=1000	Consistency: Extra quotes
	'100<=X<500'	100<=X<500	Consistency: Extra quotes
	'500<=X<1000'	500<=X<1000	Consistency: Extra quotes
	'no known savings'	no known savings	Consistency: Extra quotes
employment			
	'<1'	<1	Consistency: Extra quotes
	'>=7'	>=7	Consistency: Extra quotes

	'1<=X<4'	1<=X<4	Consistency: Extra quotes
	'1<=X<7'	1<=X<7	Consistency: Extra quotes
personal_status			
	'female/div/dep/mar'	female/div/sep/mar	Consistency: Extra quotes; Misspelling.
	'male div/sep'	male div/sep	Consistency: Extra quotes
	'male mar/wid'	male mar/wid	Consistency: Extra quotes
	'male single'	male single	Consistency: Extra quotes
age			
	222	22	Outlier: Removed extra digit on every "222" instance.
	333	33	Outlier: Removed extra digit on every "333" instance.
	6	33	Outlier: Underage: transformed to median 33.
	1	33	Outlier: Underage: transformed to median 33.
	-34	34	Invalid entry: negative age not possible.
	-35	35	Invalid entry: negative age not possible.
	-29	29	Invalid entry: negative age not possible.
	0.44	44	Invalid entry: fractional age not possible.
	0.24	24	Invalid entry: fractional age not possible.
	0.35	35	Invalid entry: fractional age not possible.
job			
	'high qualif/self emp/mgmt'	high qualif/self emp/mgmt.	Consistency: Extra quotes
	'unemp/unskilled non res'	unemp/unskilled non res	Consistency: Extra quotes
	'unskilled resident'	unskilled resident	Consistency: Extra quotes
	'skilled'	skilled	Consistency: Extra quotes
	yes	skilled	Assumed mistyping.

Lastly no duplicates have been found when comparing against the "case_no", therefore we directly proceed removing this column as it will not be relevant for the purpose of the analysis.
Before going further, it's important to notice that in the data set many instances are un-employed with more or less skilled job abilities, these have been interpreted as temporarily unemployed, hence left untouched.

2.2 Data Conversion

From the cleaned data set, two additional data sets have been produced, although for the analysis only the nominal and the mixed data set (obtained from the above cleaning) have been used.
In order to transform the nominal data set below, many attempts have been done, splitting the data in different size bins until a good overall accuracy has been found.

Nominal Data Set	Original Attribute	Transformed Attribute
credit_amount		
	Numerical series of integers between 392 and 18424	[392-1765] (1765-3279] (3279-4793] (4793-6308] (6308-7822] (7822-9337] (9337-10851] (10851-12366] (12366-13880] (13880-15395] (15395-16909] (16909-18424]
age		
	Numeric series of integers between 19 and 75	[19-32] (32-41] (41-53] (53-64] (64-75]

Numeric Data Set	Original Attribute	Transformed Attribute
checking_status		
	no checking	0
	<0	1
	0<=X<200	2
	>=200	3
credit_history		
	critical/other existing credit	0
	delayed previously	1
	existing paid	2
	all paid	3
	no credits/all paid	4
purpose		
	other	0
	repairs	1
	domestic appliance	2
	radio/tv	3
	furniture/equipment	4
	retraining	5
	education	6
	business	7
	used car	8
	new car	9
saving_status		
	no known savings	0
	<100	1
	100<=X<500	2

	500<=X<1000	3
	>=1000	4
employment		
	unemployed	0
	<1	1
	1<=X<4	2
	4<=X<7	3
	>=7	4
personal_status		
	male single	0
	male mar/wid	1
	male div/sep	2
	female div/sep/mar	3
job		
	unemp/unskilled non res	0
	unskilled resident	1
	skilled	2
	high qualif/self emp/mgmt	3
class		
	bad	0
	good	1

3 Data Analytics

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. (Data Mining, Witten & Frank, Preface).
This information can be used to make decisions and allows for the so-called "Knowledge Discovery in Databases".
Various methods are available, here we will adopt: Classification, Clustering an Association.

3.1 Classification

The purpose of classification models is to classify examples based on a target attribute, which in our case is deciding whether to give a loan or not. There are many techniques available and each one of them has different benefits (i.e. OneR, ID3), here we will look at J48.

3.1.2 J48 Algorithm

C4.5, available under the name of J48 in Weka generates a pruned decision tree, not only giving us a deeper insight into the dataset (as opposed to using OneR), but also simplifies the tree avoiding overfitting with important benefits on interpretability and generalization (hence why we use J48 and not ID3).
Although in a real life scenario we would use validation techniques as k-fold for better accuracy, for simplicity here we will adopt a training set validation (testing the model on data that it has already seen).
By using the "training set" testing option on the nominal bank data set we get an accuracy of 80.6%, a good value, with the decision tree picking the "checking_status" as the first most useful determiner (the one with the highest information gain).
Rules have been picked as a trade off between highest accuracy and highest coverage, and by looking at them we can deduce that "checking_status", "credit_history" and "purpose" play an important role

in the decision making.
Additionally, the confusion matrix below gives us a deeper insight: 662 values have been correctly classified as good (true positives) and 38 have been misclassified as bad (false negatives), whereas 144 values have been correctly classified as bad (true negatives) and 156 have been misclassified as good (false positives).
Obviously in a real life scenario we would also have to decide an acceptable error percentage between false negatives and false positives, according to what would be more damaging to the bank: either not awarding a loan to a false negative (hence missing on a chance of getting some revenue), or awarding a loan to a false positive (hence risking of loosing money).

a	b	<-- classified as
662	38	a = good
156	144	b = bad

R.1	IF checking_status = <0 AND credit_history = critical/other existing credit THEN good (67.0/18.0)
	Coverage: 67 **Accuracy:** 73.13% Out of 67 instances, 18 were misclassified. If a client has a checking status inferior to 0 and a critical or other existing credit history then the loan can be awarded.
R.2	IF checking_status = <0 AND credit_history = existing paid AND saving_status = no known savings AND job = skilled THEN bad (15.0/5.0)
	Coverage: 15 **Accuracy:** 66.66% Out of 15 instances, 5 were misclassified. If a client's checking status is less than zero, with existing credits paid, no savings and a skilled job, then the loan should not be awarded.
R.3	IF checking_status = <0 AND credit_history = existing paid AND saving_status = <100 AND purpose = new card THEN bad (31.0/9.0)
	Coverage: 31 **Accuracy:** 70.96% Out of 31 instances, 9 were misclassified. A loan should not be awarded if a client's checking status is less than zero, there is an existing paid credit history, the savings are less than 100 and the purpose for the loan is buying a new car.
R.4	IF checking_status = no checking THEN good (394.0/46.0)
	Coverage: 394 **Accuracy:** 88.32% Out of 394 instances, 46 were misclassified. A loan can be safely awarded if there is no checking status.
R.5	IF checking_status = >=200 THEN good (63.0/14.0)
	Coverage: 63 **Accuracy:** 77.77% Out of 63 instances, 14 were misclassified. A loan can be awarded if the checking status is bigger or equal to 200.
R6	IF checking_status = 0<=X<200 AND credit_amount = [392-1765] AND purpose = radio/tv THEN good (36.0/8.0)
	Coverage: 36 **Accuracy:** 77.77% Out of 36 instances, 8 were misclassified. The bank shall award a loan if the customer's checking status is between 0 and 200 included and the credit amount is between 392 and 1765 included (lowest tier range), with the purpose of buying a radio or tv.

3.2 Association

The aim of association is to find any "correlations" or associations between the attributes (something J48 can not do), hence describing frequent patterns in the data set. This can help us understand what attributes are important when deciding if to award a loan or not.
One of the simplest algorithms for association is Apriori, which given a minimum support uses an

iterative approach applying breadth first search on the dataset in order to find frequent items sets and generate rules according to their confidence (cases in which the rule application is correct).
Rules have been picked as a trade-off between high confidence and amount of attributes involved, and the results seem quite promising, especially when looking at the overall confidence which ranges between 90% and 94%, meaning that the rules are highly accurate.

R.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)
	Very high confidence: 94% 127 **Support**: 127 Instances If a customer has no checking status and the purpose for the loan is to purchase a radio or a tv then it is safe to award a loan.
R.2	checking_status=no checking credit_amount=[392-1765] 145 ==> class=good 136 <conf:(0.94)> lift:(1.34) lev:(0.03) [34] conv:(4.35)
	Very high confidence: 94% **Support**: 145 Instances If a customer has no checking status and the requested credit is between 392 and 1765 (both inclusive) then it is safe to award a loan.
R.3	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)
	Very high confidence: 93% **Support**: 153 Instances If a customer has no checking status with a credit history being either critical or having other existing credit then it is safe to award a loan.
R.4	checking_status=no checking employment=>=7 115 ==> class=good 107 <conf:(0.93)> lift:(1.33) lev:(0.03) [26] conv:(3.83)
	High confidence: 93% **Support**: 115 Instances If a customer has no checking status and their employment status is higher or equal to 7 years, then it is safe to award a loan.
R.5	checking_status=no checking personal_status=male single job=skilled 151 ==> class=good 139 <conf:(0.92)> lift:(1.32) lev:(0.03) [33] conv:(3.48)
	High confidence: 92% **Support**: 151 Instances If a customer has no checking status and they are a single male with a skilled job, then it is safe to award a loan.
R6	checking_status=no checking credit_history=existing paid job=skilled 130 ==> class=good 117 <conf:(0.9)> lift:(1.29) lev:(0.03) [26] conv:(2.79)
	High confidence: 90% **Support**: 130 Instances If a customer with a skilled job has no checking status and their have existing and paid credit history, then it is safe to award them a loan.

Association seems to suggest that even though there is no checking status, it is safe to award loans when the purpose is to buy a tv or a radio, the employment status is >= 7 or the borrowers are single males with skilled job abilities.

3.3 Clustering

Clustering is an approach which attempts to group our data in order to uncover specific patterns.
For this model the mixed bank dataset has been used defining 6 clusters and validating the model using Weka's "training set" option. The algorithm used is the SimpleKMeans with the EuclideanDistance function which measures the distance between values in order to correctly group them together.
The default settings have been kept with a cluster number of 6 and a seed of 20 which performs quite

well compared on the sum of squared errors against other seed values.
By looking at the last row we can also see that the instances are mostly evenly clustered, with cluster 0 having the least amount of instances. Overall, the attributes that vary the most amongst clusters are age, credit amount to be borrowed and employment time, giving us interesting insights into the data.
Although Weka has good evaluation methods as the "classes to cluster evaluation", the latter reaches its efficiency only when the cluster number is set to 2 and the seed to 10 with only 35% of the instances being misclassified; an alternative is to visualize our 6 clusters.
By plotting cluster_number and instance_number against the class (Fig.1) we can see that the majority of the bad class is assigned to c-0 and c-3, giving good reasons to believe these were correctly clustered.

Clustering (1000 instances)	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
checking_status	<0	no checking	no checking	0<=X<200	0<=X<200	<0
credit_history	existing paid	critical/other existing credit	existing paid	existing paid	critical/other existing credit	existing paid
purpose	new car	new car	radio/tv	radio/tv	radio/tv	furniture/equipment
credit_amount	4086	2913	3042	3111	4400	2906
saving_status	<100	<100	no known savings	<100	<100	<100
employment	4<=X<7	>=7	1<=X<4	<1	1<=X<4	1<=X<4
personal_status	male single	male single	male single	female div/sep/mar	male single	male single
age	39	44	36	29	33	31
job	unskilled resident	skilled	skilled	skilled	skilled	skilled
class	bad	good	good	bad	good	good
Clustered Instances	102 (10%)	198 (20%)	194 (19%)	186 (19%)	140 (14%)	180 (18%)

C-0: This groups is mainly composed by single males at the end of their 30's wanting to buy a new car. They have existing paid credit and good employment status, although a warning might come from the fact that they are unskilled residents wanting to borrow a considerable sum (4k), hence no loan should be extended to them.

C-1: Single men in their early 40's wanting to buy a new car get a loan even though they have other existing credit. This sounds much like the group in cluster 0, although the first one doesn't get a loan. Trust in cluster 2 may come from the fact that the sum to be borrowed is not very high (4k) and they are skilled workers which have been employed for 7 or more years (hence employment time is important).

C-2: Cluster 2 is are the people without known savings. They have existing paid credit and they are skilled workers which have been employed between 1 and 4 years, aiming to borrow a reasonable amount of money (3k) to buy either a radio or a tv. It is safe to award them a loan.

C-3: Young females in their late 20's which have been employed for less than a year although skilled and with existing paid credit shall not be awarded a loan when borrowing sums around 3k.

C-4: Single males in their early 30's with a good checking status wanting to borrow a high sum of money (around 4k) are good assets for the bank. Another positive factor could be the employment status between 1 and 4 years combined with the skilled job.

C-5: The final group is similar to the the one on cluster 4, although they want to buy furniture or equipment. Even though these young males in their early 30's have a checking status lower than 0 they have been employed between 1 and 4 years and the credit amount to borrow is not very high (3k), hence they shall be awarded a loan.

Conclusion

Classification, association and clustering have been chosen over regression as they tempt to be more powerful than a regression model, particularly in our case where we need to interpret our data in a variety of ways. For instance, we might want to show an association between many bank attributes, this would not be possible using regression. (M. Abernethy, Classification and Clustering, 2010). Decision Trees score quite well not only on the training set (80%) but also on K-folds with 10 Ks (73%). Among the 3 chosen models, association stands out for its accuracy (93%-90%), although there aren't rules describing bad cases. For this purpose, clustering could be used in concomitance, though keeping in mind that it is quite a tricky model as we must change different parameters to get a good result (i.e cluster numbers and seed value), and requires human interpretation to reach conclusions.

Overall, we see that older customers which have been employed for a consistent amount of time in a skilled job and wanting to borrow a reasonable amount of money, are a good potential asset for the bank. The bank should watch itself from young customers, especially females with a low employment time wanting to borrow a high credit amount. To support this evidence Fig.2 and Fig.3 show that c-3 (labeled as bad) is the "female" cluster which tempts to have high values for the "credit_amount".
In a social environment this could be reasonably labelled as biased since it doesn't award loans to young females, although we understand the bank wants to maximize its profits adopting a capitalist approach, and it's not concerned with this (socially important) ethical aspect.
As the bank can be safe by not loaning money to these groups, a future analysis may concern finding a strategy for turning these people into potential assets (maybe by loaning less money to them or refining the groups to find the few "safe" individuals among them).

References:

- B. Devis, The cost of bad data, 2014, Available at: https://econsultancy.com/the-cost-of-bad-data-stats/
- DePaul University, Available at: http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/k-means.html
- M. Abernethy, Classification and Clustering, 2010, Available at: https://developer.ibm.com/articles/os-weka2/
- M. Berthold, C. Borgelt, F. Hoppner and F. Klawonn Guide to Intelligent Data Analysis, SpringerVerlag, 2010
- I. Witten, E. Frank, M. Hall and C. Pal (2017) Data Mining: Practical Machine Learning Tools and Techniques

Appendix

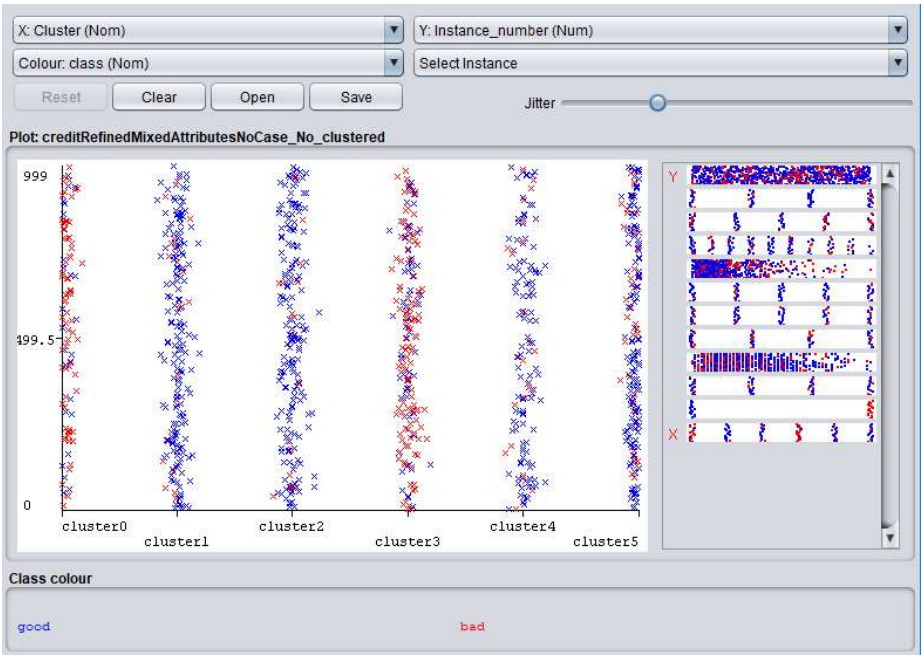

Fig.1 Clusters vs instance_number by class (colour).. Notice how the bad classes are mainly clustered into c-0 and c-3.

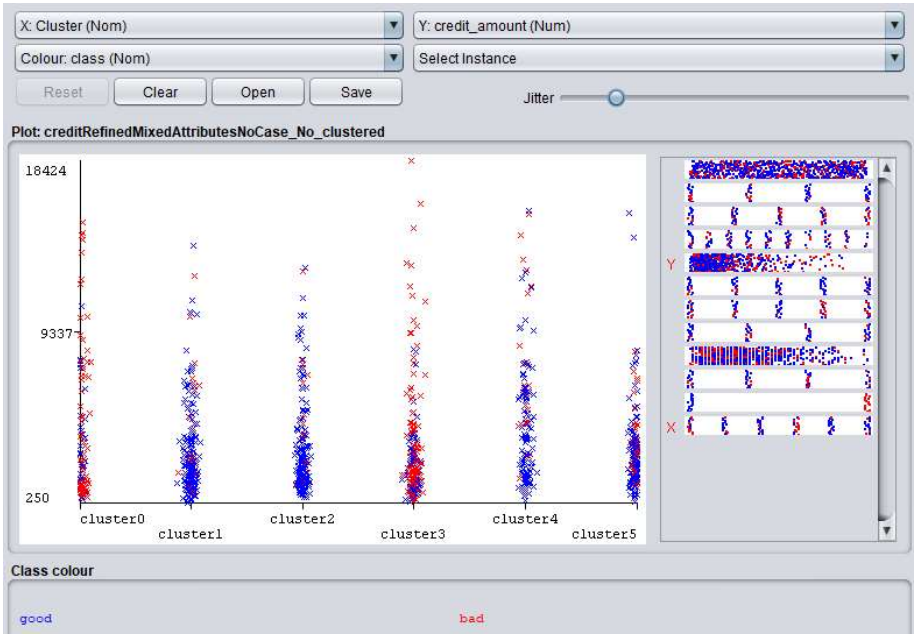

Fig.2 Clusters vs credit_amount by class (colour).. Notice how c-3 has high values labelled as bad.

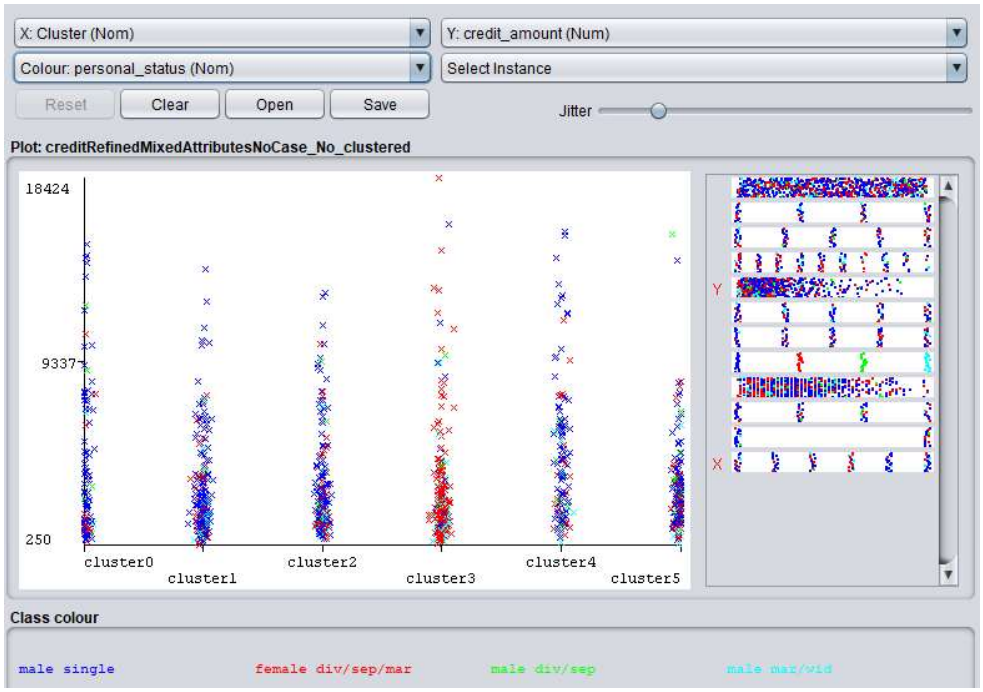

Fig.3 Clusters vs credit_amount by personal_status (colour). Notice how females are mainly clustered in c-3 associated with the bad class.