Data Mining Report

by
Mikkel Stolborg
Hlynur Örn Haraldsson

IT University of Copenhagen
MDMI, S2015
Anders Hartvig Hartzen
Hajira Jabeen
Héctor Pérez Martínez
Sebastian Risi
Noor Shaker
May 14, 2015

Contents

1 Introduction

1.1 Data selection

We have worked on a data set regarding the passengers on the titanic. The data structure is presented in table 1 with a short name and the value type associated with the variable.

Variable name	Short name	Value Type
survival	Survived	binary
pclass	Passenger Class	numeric
name	Name	string
sex	Sex	binary string
age	Age	numeric
sibsp	Number of Siblings/Spouses Aboard	numeric
parch	Number of Parents/Children Aboard	numeric
ticket	Ticket Number	string
fare	Passenger Fare	numeric
cabin	Cabin	string
embarked	Port of Embarkation	attribute table

Table 1: Titanic data set variables with short description and classification.

The type binary and binary string, means there is only two options, the first of the types is based on numeric binary data, whilst the second is based on string data. The value type attribute table covers three options. The type designation in the table are based on how they are treated in the data mining code. This means that whilst some of the other variables might be a three option choice, it will not be treated as such. The reason for this division is mostly due to the nature of string variables and numeric variables.

1.1.1 Data description

The variables in the table are specified as follows.

Survival is simply a 1 for survived and 0 for not.

Passenger Class is a numeric value which can take on either 1, 2, or 3. The lower the value, the higher the class. The class is a proxy for the passengers socio-economic status.

Name is simply the name of the passenger, starting with surname.

Sex is the sex of the passenger.

Age is the age in years of the passenger. If the age is less than one it is fractional, and if it is estimated the age is in the form xx.5, meaning it has a decimal value as well.

Number of Siblings/Spouses Aboard, is the number of relatives, yet some of the relations are ignored. The meaning of sibling and spouse is summarized below.

Number of Parents/Children Aboard, is similar to the Number of Siblings/Spouses Aboard. The meaning is as well summarized below.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic

Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignored)

Parent: Mother or Father of Passenger Aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins, nephews/nieces, aunts/uncles, and in-laws. Some children travelled only with a nanny, therefore parch=0 for them. As well, some travelled with very close friends or neighbours in a village, however, the definitions do not support such relations.

Ticket Number is the number printed on the ticket. It is basically an identification string with characters and numbers. The tickets have no real pattern.

Passenger Fare is how much the passenger paid for its ticket. The price probably is listed in dollars.

Cabin is the cabin number. Again this is an identification of the rooms, yet there is not really any good patterns within the identification string.

Port of Embarkation is the place of embarking for the passenger. The port is abbreviated as follows. C is for Cherbourg, Q is for Queenstown, S is for Southampton.

1.2 Research question

Our primary question which we want answered was:
"Which attributes contributes mostly to the survival rate of a passenger on the Titanic
Here we wished to figure out what set of parameters would ensure the highest rate of survival on the Titanic. We would use classification through a classification tree to figure out which set gives the highest percentage of survival.

The secondary question arose when looking at clusters of the data.
"Which societal data can be found in clusters of the titanic data"
Looking at the data, we decide to try clustering to see if there was an emergent pattern. It would be interesting to see if the were a relation between wealth and number of children and the like.

1.3 Tools for data mining

We choose to use the free tool called Orange[1]. This tool allowed to quickly manipulate the data, such that we could extract the interesting elements. In the tool you manipulate the data by clicking and drawing connections between data elements and processing methods. Before explaining further, we have included the map of the process used for our data set in the orange framework, see figure 1.

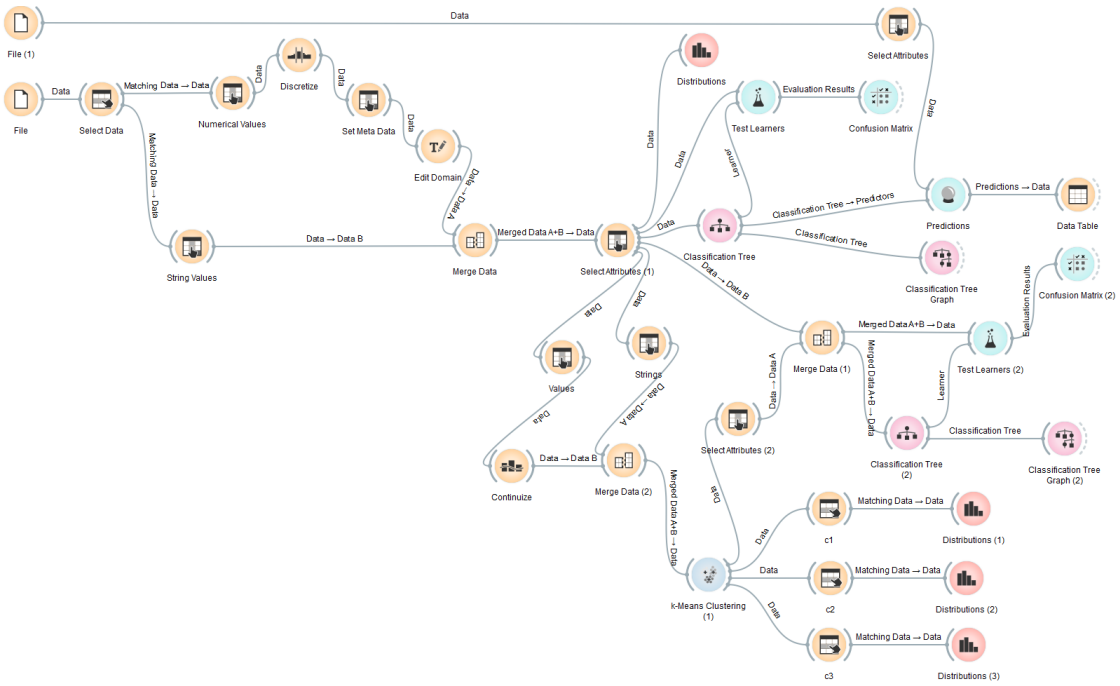

Figure 1: The map of the methods and processes used on the data.

2 Data mining

2.1 Preprocessing

2.2 Classification tree

2.2.1 Cross Validation

2.3 K-means Clustering

2.3.1 Data validation

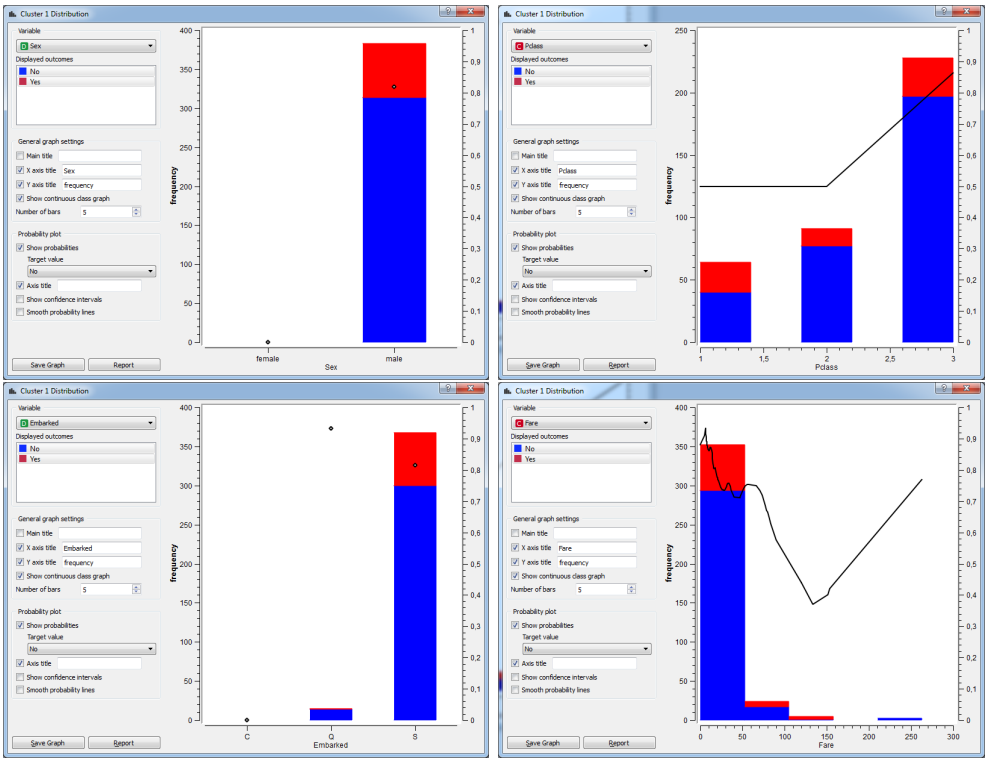

Figure 2: Cluster 1

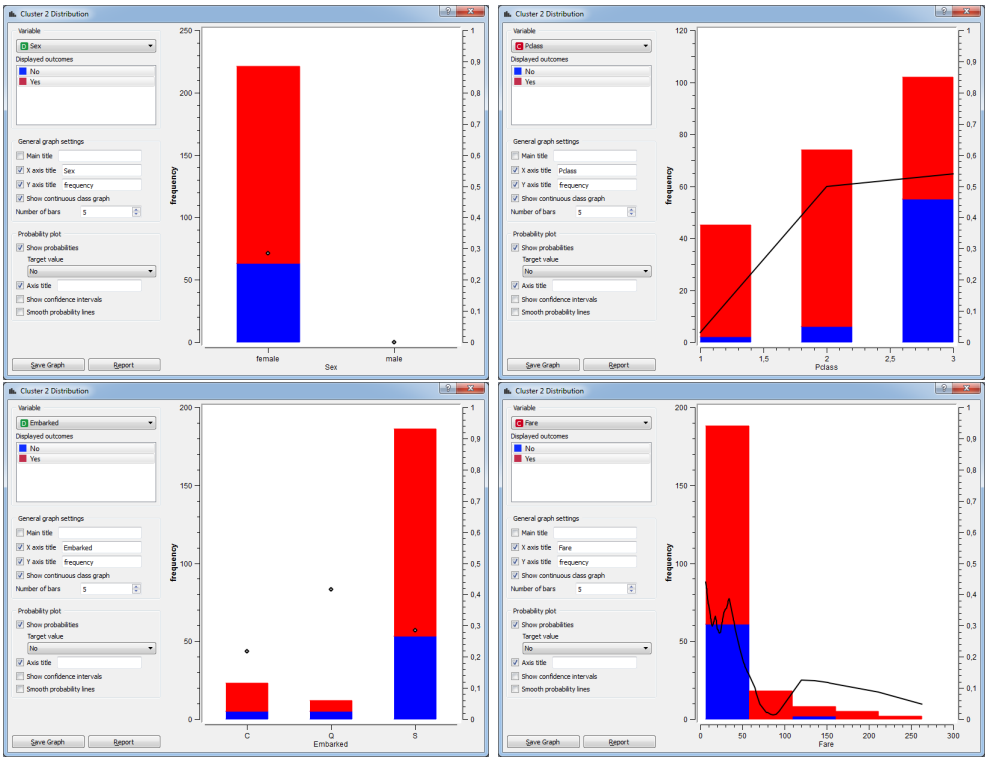

Figure 3: Cluster 2

3 Conclusion

3.1 Societal impact

References

[1] *Orange Data Mining*, http://orange.biolab.si/, 13-05-15