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 **??** with a short name and the value type associated with the variable. 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

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.

VARIABLE DESCRIPTIONS: survival Survival (0 = No; 1 = Yes) pclass Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd) name Name sex Sex age Age sibsp Number of Siblings/Spouses Aboard parch Number of Parents/Children Aboard ticket Ticket Number fare Passenger Fare cabin Cabin embarked Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES: Pclass is a proxy for socio-economic status (SES) 1st Upper; 2nd Middle; 3rd Lower

Age is in Years; Fractional if Age less than One (1) If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch) some relations were ignored. The following are the definitions used for sibsp and parch.

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 neighbors in a village, however, the definitions do not support such relations.

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.

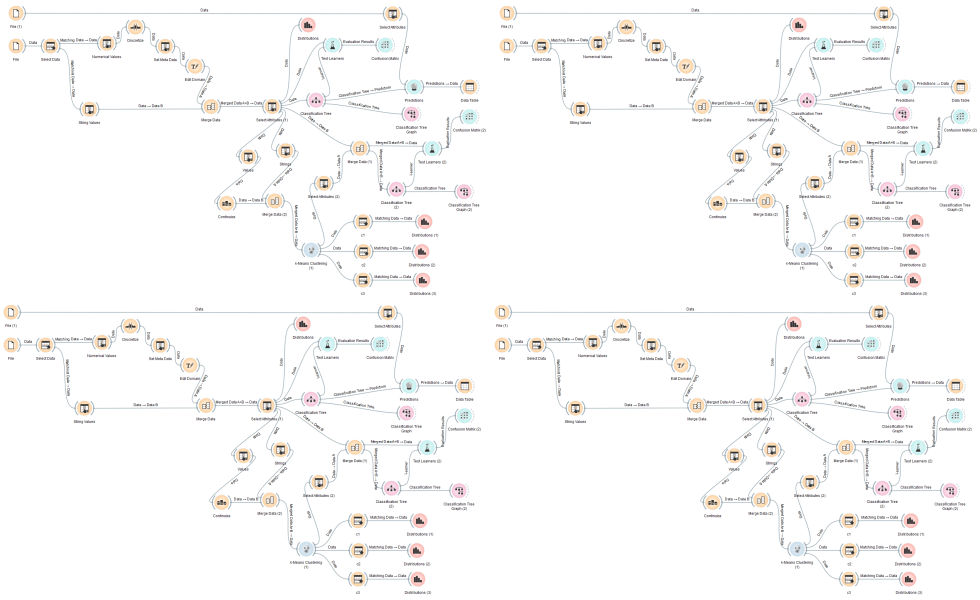

Figure 1: The map of the methods and processes used on the data.

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

3 Conclusion

3.1 Societal impact

References

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