Visual Analytics on Credit Cards Defaults

Visual Analytics project, A.Y. 2017/2018

Maria Ludovica Costagliola Emanuele De Santis

ABSTRACT

The project was developed during the Visual Analytics course. It concerns the visualization of credit cards owners data in order to make the bank director knowing the customers that are supposed to not be able to pay the credit card bill in the next month.
All data are represented using simple and well-known views that immediately highlights similarities among customers and give to the user an overview on all customers.

1. INTRODUCTION

After the paper presentation done during the lectures, we decided to focus our attention on a dataset related to bank transactions. Most of the bank transactions datasets are not public available (or they contains few useful information to protect users' privacy), but we were able to find a dataset related to this field.
We were thinking about the need for a bank director to always know how customers with a credit card from his financial institution behave. Particularly, we pay attention to the last payments and to the corresponding bank account balances of those customers.
From these data and from some other personal information of the customer (for example age, marriage status, ...), it is possible to identify the ones that probably will not be able to pay the credit card bill in the next month.
The prediction is done by a machine learning algorithm, but the result is useless if it is not combined with an efficient visualization of the whole data. In fact, with this visualization a bank director is able to better understand the result of the machine learning algorithm, considering also the similarity between the result and some preexisting patterns or clusters.

2. DATASET

The dataset used in this project is taken from UCI database [1], it contains about 30000 tuples, each with 24 attributes. Some tuples lack of some values and so they were all removed in order to have a completely useful dataset. A few attributes for each remaining tuple were also removed because they were of no interest for us. We get in this way a more manageable dataset thanks to a lower number of tuples (15337) and of attributes (19).
About the used attributes, we can identify four of them that are categorical: age, sex, marriage status, education. These represent characteristics of the owner of a particular credit card and they are used for statistical considerations. We have six numerical attributes named *Amount of bill statement*, one for each month from September 2005 to April 2005 and the corresponding six numerical, named *Amount of previous payment*.
The last attribute of each row is the *target*, that is the prediction about the ability to pay on October.

3. REFERENCES

[1] I.-C. Yeh. UCI machine learning repository, 2016.