Preliminary Draft - VA Final Project

Maria Ludovica Costagliola - 1657716
Emanuele De Santis - 1664777

A.Y. 2017/2018

1 Dataset

In this final project we decided to use *UCI Default Credit Card Dataset*[1], that is composed by about 30000 instances, each of them with an ID and 24 attributes:

- Amount of given credit
- Gender
- Education
- Marital status
- Age
- History of last 6 months payments
- History of last 6 months bill statements
- Binary classification

This dataset is composed of both categorical (e.g Gender, Education,...) and numerical (e.g. Hisotry of last 6 months payments) attributes. Moreover this dataset is near to the topic we presented in the discussion during the lectures.

We have already filtered the dataset from entries that have some missing attribute, so now the number of tuples is about 15000. We have also already developed some python scripts to extract the information needed from the dataset (for example the 2-component PCA values and the frequency of the classification among all the categorical attributes).

2 Visualization and Analytics

In this final project we would make 6 views:

[1] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

- A scatterplot where we represent the dataset using PCA. Using this view is possible to spot immediately the spatial distribution of the default credit card owners. In this way is possible to do the classification of new instances even without a machine learning algorithm
- A linechart where we represent the history of payments and bill statements of a given user (selecting it by clicking on a point in the scatterplot)
- A stack barchart for each categorical attribute to represent frequency distribution of default and correct owners among these attributes.

Each view will be coordinated: clicking on a dot in the scatterplot will make the linechart represent the history of payments and bill statements of the selected user, while the stack barcharts highlight the correspoding value of the categorical attributes for that user. Moreover clicking on a bar of a stack barchart will select all the users that are represented by that bar, so all the other barcharts highligh the bars corresponding to these users, the scatterplot will highlight the points corresponding to these users and the linechart will represent all the histories of each of these users.

For the analytics part we would create a brushing on the scatterplot, that manually overrides the classification given in the dataset (we may think that this tool can be used in case we don't have a good training set, so, exploiting patterns in the scatterplot, we can define an area of "default owners" points).

Moreover we would let the user insert new tuples, so the system has to classify these new instances, updating at the end all the views according to the new dataset.