HOMEWORK OF VISUAL ANALYTICS

Preliminary approval

Carla Pascual 1840620

Federico Arcos 1879280

1. GENERAL IDEA

We chose the PCA algorithm to preprocess the dataset for choosing the right dimensions, using a scatterplot to spot the spatial distribution of the default credit card owners. We can clasify new instances without the necessity of a machine learning algorithm.

For the attributes to represent frequency distribution of default owners we use a stack barchart and we represent the payments and bill statements of a given user with a linechart.

The stack barcharts shows the value of the attributes for the user.

We select the users represented with a bar clicking on that bar of a stack barchart .

The linechart represent the histories of the users and the scatterplot shows the point of those users.

The linechart represent the history of payments and bill statements of the selected user clicking on a dot in the scatterplot.

ANALYTICS

We change the clasification of the dataset creatingld create a brushing on the scatterplot, that manually overrides the classi cation given in the dataset.

The user can put new tuples. To do that, we need to classify the new instance updating the views at the end by the system.

2. THE DATASET (characteristics and context)

UCI Default Credit Card Dataset is our decisión (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients). As we can see there, there are 24 attributes and 30.000 instances. The data set characteristics are multivariate and the area is business.

ATTRIBUTE INFORMATION:

 This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

There are categorical atributes like Gender or Education and numerical too, like History of payments atributes.