

Data Mining course.

Master of Information Technologies. Enginyeria Informàtica

FIB 2009-2010

Tomàs Aluja, Lluís Belanche

1st. Practical work: Characterization of the buyers of product Alpha and its competitors.

A certain brand of high consumption is intending to launch a promotion spot on television of one of its products. Before the launch, the assess the behavior of the consumers in regard of its spot and the spots of the competitors; the company decided to conduct a survey (copy test) among a sample of 252 potential customers, whom were interrogated on the last bought brand and their preferences after viewing ads (the trademark in question and those of competitors), as well as their socio-demographic information. With the information gathered it was formed the file ALPHA.XLS.

The purpose if this practical work is to characterize the buyers of different brands ALPHA, BETA, GAMMA and DELTA, moreover it is intended that students get used with the R software.

To this end, the student will use four different approaches: PROFILING, VISUALISATION, CLUSTERING and ASSOCIATION RULES.

Work to be undertaken:

The work to be done is divided into four phases:

- A. Feature selection and profiling of consumers of the different brands.
- B. Multidimensional positioning of brands.
- C. Clustering of consumers.
- D. Association rules for each brand.

A. Profiling of consumers of brands ALPHA, BETA, GAMMA and DELTA.

Let's consider the file ALPHA.XLS. The aim is to obtain an automatic characterization (profile) of the ALPHA brand and all its competitors according the socio-demographic variables.

Import Excel file ALPHA.XLS containing the data and dictionary. Take the socio-economic variables as explanatory, whereas the response variable will be the "preferred brand after viewing the advertising spot" (in the second part of practice, the latest bought brand will also be a response variable, in this first part it is optional to take it as explanatory as well).

1. Perform a first summary of the data file.
2. Impute the “don’t know”s of variables “last bought brand” and “preferred brand”. Treat them as it were missing values.
3. Recode the continuous variables in intervals (4 or 5). Name the modalities with a string which can be directly understandable (not use values "1", "2", but... "age 20", "age 25,"... for example, for the variable age). Define the variable “preferred brand” and “last bought brand” with their respective brand literal (i.e. "alfa1", "beta1"... and "alpha2", "beta2"...) for example, to differentiate the last purchased brand from the preferred brand after viewing the spot. Repeat the same procedure for the variables of profession, but treat now the “don’t know” of profession as a new category for this variable.
4. Represent graphically the dependence of the “preferred brand” with all explanatory variables. For the continuous explanatory variables graph the mean value for each brand whereas for the categorical explanatory variables, create the corresponding tables crossing the “preferred brand” with each explanatory variable and graph the profile of purchase on each brand respect each explanatory variable.
5. Feature selection: Identify the variables more related with the response variable “preferred brand”, first among the continuous variables and later among the categorical. Could we envisage a non linear dependence of some variables? (this part is done as a mere practice of feature selection, since here the number of explanatory variables is very small).
6. Find the profile of each brand. For each brand detect which are their differential characteristics. First detect which continuous explanatory variables are significantly associated with each brand and later which are the modalities of the categorical explanatory variables more associated which each brand.

B. Multidimensional positioning of brands and change induced by the television spot.

We want to obtain a visual representation of the association between each brand and the socio-economic variables and the change induced by the television spot. To do so we take the “last bought brand” the “preferred brand” as response variables in this part of the practice.

7. Carry out a multivariate analysis for a joint display of all socio-economic variables. Take all socio-economic variables in the active data frame and the two responses in the illustrative data frame. Using acm function to perform a Multiple Correspondence Analysis. Interpret the factorial plan formed by the 1st factor and

the 2nd, or by the 1st factor or the 3rd (depending whether of which do you think will be more understandable for the final user).

8. Project the illustrative variables, the “last bought brand” and the “preferred brand”. Develop a possible explanation for the market position of each brand and the potential impact of spot on the change of position of each of them.

C. Socio-economic clustering of consumers.

We will perform the clustering using the factorial coordinates of the MCA already obtained.

9. Decide the number of factors to take into account for the clustering phase.
10. Perform a Hierarchical Clustering with consolidation. Obtain the histogram of height values. Decide the number of classes to cut the tree.
11. Interpret the partition. Use the same procedure used in part 5.
12. Identify the classes most disposed to each brand.

D. Association rules of each brand

Which are the events most associated to each brand?. We are interested in the rules having as consequent the categories of the “preferred brand” variable and as antecedent any combination of the categories of the socio-economic variables (also you can include the categories of the “last bought brand variable”).

13. Se the program “data loader” of R in order to cast the categorical variables into a transaction R object or matrix suitable for mining with “apriori” (see the document ‘arules.pdf’ (p.16) in the web page to get more information on how can this be done keeping the original labels).
14. Run “apriori” to find a set of interesting rules for each brand. Prepare a table for each brand, giving the antecedents, the support, the confidence and the lift of each rule.

Conclusions:

Taking into account all obtained results, write your own conclusions about the differences that exist between brands.