Unsupervised Learning for Panel Data

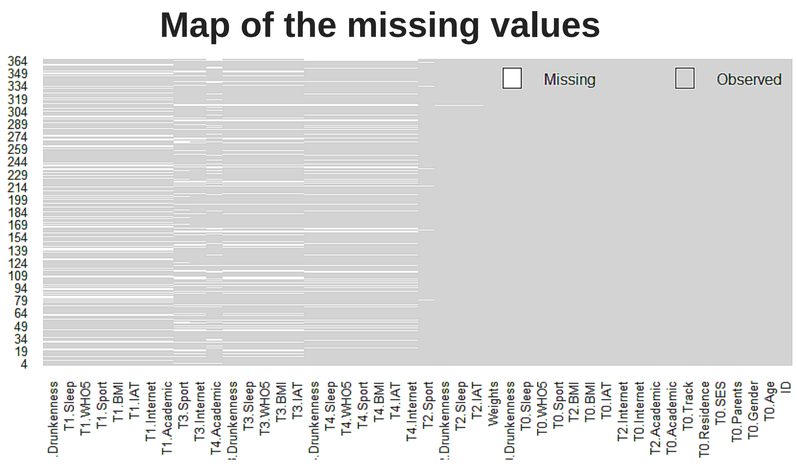
Pierre Alpi, Jülide Yilmaz (master TIDE), Madalina Olteanu (laboratoire SAMM)

Using a stochastic version of self-organizing maps (1) and a classification tree (2), we propose an unsupervised analysis of a longitudinal survey dataset. First, after pre-processing the data through multiple correspondence analysis (MCA), we cluster a first hand of « premium » observations into several groups. Second, the rest of the data are assigned to these groups using the prediction abilities of our model. This allows to replace missing values for several features. Finally, a classification tree (CART) is trained in order to find out which variables are the most relevant for the clustering.

Preliminary analysis

This survey database contains 364 weighted individuals for which we have 46 features:
- Non-temporal qualitative and quantitative features.
- Temporal qualitative and quantitative features over 5 periods (T0-T4)

As often in survey datasets, the data contains quite a lot of missing values. A quick analysis of missing values point out that most of it is generated by unintentional non-response at times 1-4.



- Here, half of the features have missing values, especially from T1 to T3.
- The stripes of white indicates that individuals systematically didn't respond at a given time.
- Such non-response effect should not be taken at any step of the analysis.

Step 1 : prepare the data

- "Premium" individuals i.e. those who answered the survey during the whole period, are separated from the rest. As non-response may be intended in this group, the "unknown" modality is created for qualitative variables in order to capture such effect. The final dataset contains 185 "premium" observations.
- As the SOM Algorithm cannot be used directly for qualitative features, we first perform a Multiple Correspondence Analysis (MCA) in order to obtain a quantitative representation of the information from our dataset.

About self-organizing maps (SOM)

A self-organizing map, also known as Kohonen map, is a machine-learning algorithm performing both clustering and non-linear projection.
SOM may be viewed as a generalization of the k-means algorithm, where clusters are equipped with a topological structure.
Main steps of the online (or stochastic) algorithm, once the prototypes or code-vectors representing the clusters were initialized:
(1) One input is randomly selected and affected to its nearest prototype.
(2) The “winning” prototype and all its neighbours are updated: they are moved towards the current input.
Steps 1 and 2 are repeated until convergence (the cost function does not vary any longer).
The output is, for example, a rectangular grid of clusters, each cluster being resumed by its corresponding prototype

Step 2: Clustering with SOM

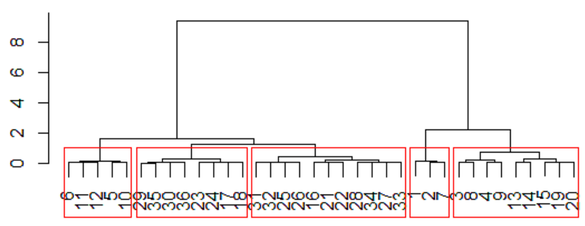
Results for a square grid of clusters with dimension 6x6 shows :

- ***Number of observations by node :*** proportionnaly to the size of the pink square, we can see the number of teenagers classified in a node. The clusters on the left side contain the most homogeneous observations. We can imagine that the observations classified on the left side are very particular.

Number of observations by node

29		15			53
	0				
				4	
		1			

Dendrogram for superclusters



- ***Hierarchical clustering of nodes*** : due to a big number of clusters we made, we applied a hierarchical clustering on the nodes and kept four superclusters.

Supercluster grid

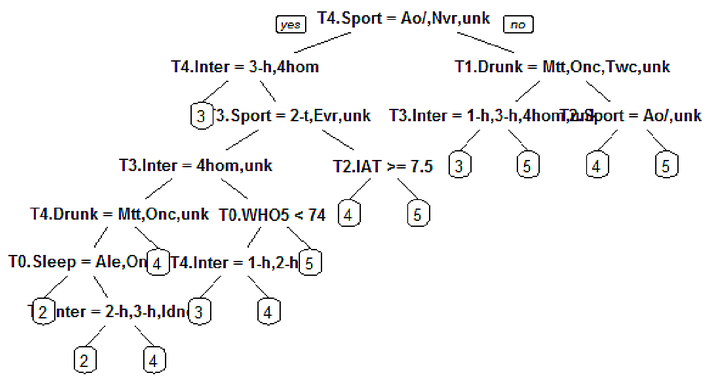
6	12	18	24	30	36
5	11	17	23	29	35
4	10	16	22	28	34
3	9	15	21	27	33
2	8	14	20	26	32
1	7	13	19	25	31

- ***Supercluster grid*** : colors represent each superclass which includes homogeneous clusters. The grid shows no odd union of clusters in a superclass which validates results of the SOM algorithm.

Step 3: CART tree

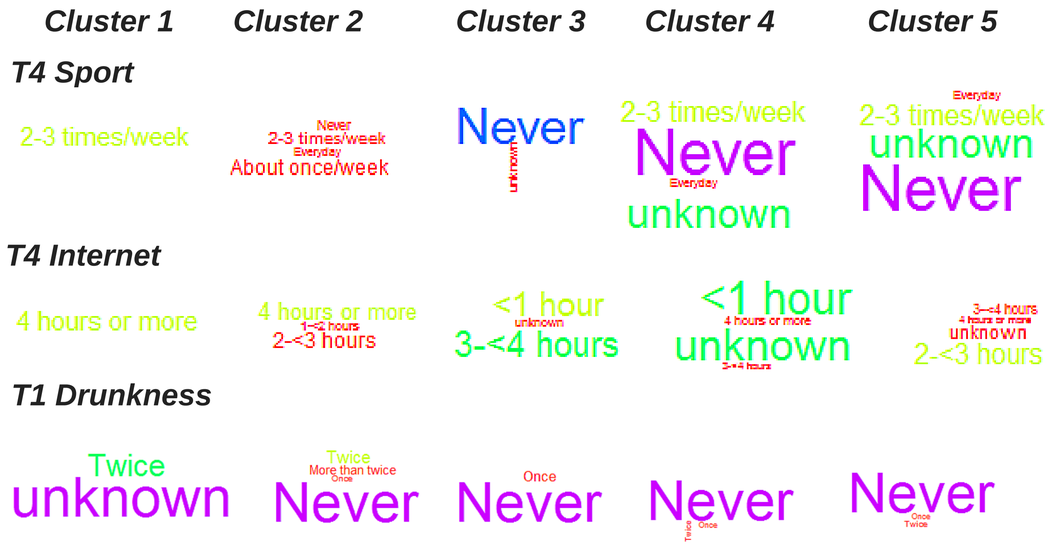
- In order to detect in our dataset the most interesting features, a classification tree is performed on our premium dataset using as predictor the supercluster number.
- Each intersect sorts the observations according to their answer to a binary question.
- Our predictive classification tree shows us the best features for classifying our dataset.

Classification Tree



Most important features :
- T4 Sport: About once a week, Never, unknown
- T4 Internet: 3-4 hours, 4 hours or more
- T1 Drunkness: More than twice, One, Twice, unknown
- T3 Sport:Two times a week, Everyday, unknown

Step 4: Wordclouding on features

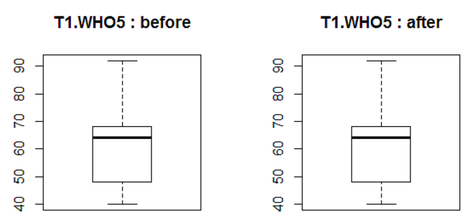


Using wordcloud, for a set of features the most frequent modalities for each supercluster is shown. The first supercluster regroups individuals with an important use of internet in T4, as well as those who practice sports in T4 (2-3 times/week). On the other hand, the third supercluster regroups individuals who do not practice sports in T4 or do not get drunk in T1.

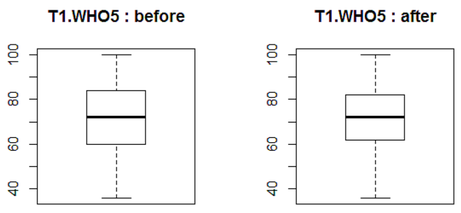
Step 5: Missing value imputation

- After separating our "premium" dataset in clusters with SOM, we assigned each observation of our non-premium dataset to an existing cluster using our prediction model.
- Let's compare the 6th cluster and the 36th clusters, before and after imputation, which respectively contains 29 and 53 observations.

Boxplot for T1 WHO5 for the cluster 6



Box plot for T1.WHO5 for the cluster 36



- We chose T1.WHO5 because half of its values where missing (i.e. missingness map). Before and after imputation for the 6th the distribution of variables does not changes a lot as we can see. For the 36th cluster, which contains more observations, the distribution has changed a bit, in gathering the values around the median.

Conclusion

Using MCA and Self Organising Maps, we proposed a method for handling missing values for quantitatives features. Moreover, a decision tree is included in order to show which features have strong classifying power. In other terms, thses features (T4 Sport, T4 Internet, T1 Drunkness) have a strong influence on behavior of individuals in our dataset. Our results show that imputation was effective enough to leave unchanged boxplots of the dataset.

References

(1) T. Kohonen, Self-Organizing Maps, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 2001
(2) Breiman, Leo, Jerome Friedman, R. Olshen and C. Stone (1984). Classification and Regression Trees. Belmont, California: Wadsworth.
(3) Villa-Vialaneix N., Bendhaiba L., Olteanu M. (2015) SOMbrero: SOM Bound to Realize Euclidean and Relational Output, R package version 1.1