ISI2017
MARRAKECH
61st ISI WORLD STATISTICS CONGRESS

Using Knoweldge Discovery approach to understand life conditions in Ecuador

V. Figueroa, MP. Guillaumet, R. Miranda, C. Ramírez, A. Tardío Canito
University Research Institute for Sustainability Science and Technology, Universitat Politècnica de Catalunya-BarcelonaTech (Spain)

Karina Gibert*
Knowledge Engineering and Machine Learning Group (KEMLG)
Department of Statistics and Operations Research
University Research Institute for Sustainability Science and Technology
Universitat Politècnica de Catalunya-BarcelonaTech, Spain – karina.gibert@upc.edu

The goal of this work is to establish a typology of housings in Ecuador on the basis of the information provided by the National Survey on Life Conditions (ENCV) performed by the National Institute of Statistics and Census in Ecuador, which provides information about 30000 households and 132 items. Clustering methods suitable for both numerical and qualitative data are used to this purpose.

Keywords: Clustering; knowledge; pre and post-processing; life conditions

1. Introduction

The National Institute of Statistics and Census in Ecuador performs the National Survey on Life Conditions (ENCV), following almost 30000 households from which 132 variables are taken. The survey provides information about the characteristics of housing in the country on different aspects. In this work, data from ENCV [ENCV, 2015] is exploited under a knowledge discovery approach to understand how people lives in Ecuador.

2. Materials and Methods

In a first analysis, a selection of 24 variables is considered, regarding the materials used to build the different parts of the housings and neighborhoods, structure of housings, access to water, sanitation, waste elimination and type of occupation of the housing. Approximately half of the sample correspond to rural households distributed in the 24 provinces of the country. Data contained a 13.30% of missing values, concentrated in 7 of the variables. The Little [Little, 1970] test indicates that none of them are random. Bivariate analysis permits to understand that most of them are structural missing values that can be substituted accordingly (this is the case, for example of the variable value payed for the water, which is missing for those housings declaring no access to water distribution network).

The goal of this work is to establish a typology of housings. Cluster analysis is performed. Since the ascendant hierarchical clustering is quadratic, the standard implementation in R cannot process the complete sample. For this reason, a subsample, stratifying by the city is extracted. The Ward's method [Ward, 1963] is used. However, since we are interested in building the clusters while considering the interactions between both numerical (like surface of the housings) and qualitative variables (materials of walls, for example), it has been modified to use the Gower [Gower, 1971] distance instead of the classical euclidean distance. The Calinski-Harabaz [Calinski 1974] index was then used to find the better cut of the resulting dendrogram. The two best cuts of the dendrogram are in 2 and 4 clusters. The cut in 4 clusters is analyzed by using post-processing techniques that permit to understand the meaning of the clusters. Profiling graphs are performed and bivariate descriptive analysis is done, by using the class variable as a reference against all variables used from the ENCV survey. Thus, conditional distributions of variables versus the 4 discovered clusters are analyzed. Chi-square tests are used to see which qualitative variables can discriminate the clusters; whereas, for numerical variables either ANOVA or Kruskall-Wallis are used, depending on the variable distribution. Twenty of the variables behave differently in some of the classes. For the significant variables, Traffic Light Panels [Gibert, Conti 2015] are performed to support conceptualization of clusters.

3. Results

From the 4 clusters discovered, two correspond to rural households and two to urban ones, being each group subdivided by the region (Seaside or Mountains).

C1) Rural households in Seaside (includes Amazonia) can be made of wood and can also use wood to cook, being those with more scarcity in access to water and sanitation.

C2) Rural households in the mountains can use adobe to build the housings and often lack of inner pavement is found, they can also cook with wood but access to water through water distribution netwoks (public or not).

C3) Urban households in the seaside have better infrastructures and are mainly build of brik, but it still takes some time to get water due to irregularities on the network.

C4) The urban households in the mountains are stronger and have better access to water and gas, being also the households paying the more for the water and the housing rent as well.

This analysis gives a first idea of where are the more vulnerable households and provide a first reference to design intervention plans to improve quality life in the country.

As a future work, the introduction of Association Rules mining methods [Agrawal, 1994] is intended to be used for an automatic interpretation of the clustering and this first clustering will be used as the kernel for a CURE clustering [Guha 1998] to scale up the clustering to the complete dataset, on the basis of the clusters obtained with the sample. Also, the remaining variables of the survey are going to be included into the analysis, these giving information about life styles of the households in different aspects, like cooking habits or uses of water among others.

References

Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communications in Statistics - Simulation and Computation, 3(1):1—27, Taylor and Francis

ENCV (2015)]http://catalogo.datosabiertos.gob.ec/dataset/encuesta-nacional-de-condiciones-de-vida-inec/resource/c1315f1f-f577-422f-b11a-8191b894c714

Gibert K, D. Conti (2015) aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. Artificial Intelligence Communications 28:113-126, IOSPress

Gower, J.C 1971: A General coefficient of similarity and some of its properties. Biometrics 27, 857—874

S. Guha, R. Rastogi, and K. Shim. Cure: an e_cient clustering algorithm for large databases. In SIGMOD '98 Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pages 73{84. ACM Press, 1998.

J.D. Little. Models and Managers: The Concept of a Decision Calculus. Man. Science 16(8):B466-85, 1970.

Ward, J.H. 1963 Hierarchical grouping to optimize an objective function. J. Am. Statis. Ass. 58: 236--244