entropy 2021

sciforum

Abstract

Classifying Dengue Cases Using CatPCA in Combination with the MSU Correlation †

Santiago Gómez-Guerrero [1], Miguel García-Torres [2], Gustavo Sosa-Cabrera [1], Emilio G. Sotto-Riveros [1] and Christian E. Schaerer [1]

[1] Universidad Nacional de Asunción, San Lorenzo, Paraguay
[2] Universidad Pablo de Olavide, Sevilla, Spain

† Presented at the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May 2021; Available online: https://sciforum.net/conference/Entropy2021/.

Published: 5 May 2021

Dengue is a mosquito-borne viral infection that is a leading cause of serious illness and death among children and adults in many countries across the world. In Paraguay, dengue incidence has been increasing especially in urban areas, becoming endemic and epidemic in the last few years.

This work seeks to understand what factors are responsible for the epidemic and hemorrhagic varieties of dengue. Considering that collected data are of mixed nature (nominal and continuous), Categorical Principal Components Analysis (CatPCA) is adopted as a first tool. However, interpretation of CatPCA output can be challenging, partly because the same variable may appear throughout several of the principal components.

Multivariate Symmetrical Uncertainty (MSU), an entropy-based similarity measure, allows quantifying correlations in a multivariate environment and detecting both linear and nonlinear associations. In this work, the MSU measure is used in combination with CatPCA to obtain greater insight regarding the relevance of each variable.

We apply the two techniques combined in stages, using nation-wide data collected by the country's Sanitary Surveillance Department from nearly 200,000 suspected and confirmed cases throughout 5 years. The first few runs of CatPCA help to discard the less relevant attributes. A subsequent run of CatPCA provides principal components that account for a high percentage of the total variance. Working with the attribute sets identified by CatPCA, MSU finds *n*-way interactions and correlations, and groups those attributes for further selection. Segregation of attributes in disjoint groups can be done at this stage; this allows for an easier interpretation of groupings including those containing the key linear and nonlinear correlations.

The outcomes from this combined approach are better than the CatPCA alone, identifying individual and grouped variables that contribute to the behavior of the class.