entropy 2021

sciforum

Abstract

Multivariate Symmetrical Uncertainty as a Measure for Interaction in Categorical Patterned Datasets †

Santiago Gómez-Guerrero [1], Gustavo Sosa-Cabrera [1], Miguel García-Torres [2], Inocencio E. Ortiz-Samudio [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

Interaction between three or more variables is often found in statistical models where the response variable is numeric. Techniques like regression or analysis of variance can show interaction as a composite-variable term in the model, and their algorithms include calculations to determine the size of the interaction. However, there is a lack of methods to appropriately detect and measure interactions when the variables are a mix of numerical and categorical.

In this work, we present a way of measuring interactions between n categorical variables for the case of samples with patterned records. In these datasets, of all the possible attribute value combinations only some of them are present. We explore various datasets using the Multivariate Symmetrical Uncertainty, which is a recently developed entropy-based correlation measure. MSU is unbiased for representative samples, and it detects linear and non-linear associations between any mix of categorical and discretized numerical variables.

More precisely, we explore the behavior of a number of known 3-variable record structures such as XOR, AND, OR, NAND and others, plus their extensions to more variables. Simulations using different sampling scenarios on each record structure show that every n-variable pattern possesses a characteristic minimum value M_L and a characteristic maximum value M_U for the MSU correlation.

It is observed that the M_L value, attained when the pattern occurs in a certain combination of frequencies, hints that interaction is intrinsically expressed by this minimum value. Other sampling scenarios resulting in higher MSU values carry this intrinsic interaction due to the pattern itself, plus additional correlation due to extra occurrences of some configurations.

This method of quantifying n-way categorical interactions opens up new questions on the behavior of datasets that exhibit multivariate correlation, as for example in semi-patterned and non-patterned datasets.