Statistical determination of numerical codes in mixed databases

Ángel Kuri-Morales [1], Raúl Galindo-Hernández [2]

[1] Instituto Tecnológico Autónomo de México, Ciudad de México. akuri@itam.mx
[2] Posgrado en Ciencia e Ingeniería de la UNAM, Ciudad de México. raul.galindo.isc@gmail.com

Abstract

Machine Learning (ML) algorithms try to obtain information from the databases in order to take advantage of this information and generate a model that is able to solve future problems. A smaller part of the existing ML algorithms focus on dealing with non-numerical data (categorical data). This is an area that has received less attention from researchers, mainly because it is harder to map categorical data in metric spaces. By having data in a metric space, richer information can be obtained inherent in the data. In this work the problem of transforming categorical data into numerical data is addressed. Specifically is applied to mixed databases (MDB), which contain both types of data. It is required to transform the categorical variables into numerical variables, preserving the embedded patterns of the MDB, in order to subsequently access the wide range of algorithms that treat only numerical data. A statistical approach is taken to transform the categorical attributes of a MDB into numerical attributes. The central limit theorem [1] and the multivariate approximation [2] are the bases from which the solution of the problem starts. After its transformation to a fully numerical database, it is ready for the application of computational intelligence algorithms based on metrics. Once the MDB is transformed into a fully numerical database, a neural network is trained and the results are compared with 11 different classic machine learning algorithms. A higher accuracy is achieved in several databases which are structurally different.

Keywords *— Machine learning, multivariate approximation, central limit theorem, neural networks.*

References

[1] López-Peña I. Kuri-Morales A.F. Normality from monte carlo simulation for statistical validation of computer intensive algorithms. *Advances in Soft Computing (MICAI)*, 10062, 2017.
[2] Kuri-Morales A. F. Lopez-Peña I. Multivariate approximation methods using polynomial models: A comparative study. *Artificial Intelligence (MICAI)*, 2015.
[3] E. W. Cheney. *Introduction to approximation theory*. McGraw-Hill, 1966.
[4] Gurland J. Ram D. C. Pearson chi-squared test of fit with random intervals. *Biometrika*, 1972.
[5] Cybenko G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303 – 314, 1989.
[6] A. F. Kuri-Morales. Closed determination of the number of neurons in the hidden layer of a multilayered perceptron network. *Soft Computing*, 2017.