Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 6, n. 2, 2018.

Trabalho apresentado no XXXVIII CNMAC, Campinas - SP, 2018.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Feature Selection with Multivariate Symmetrical Uncertainty to predict Dengue Cases using Deep Learning

Marcos Ortega[1]
College of Science and Technology, Computer Engineering, UNCA, Coronel Oviedo, Paraguay
Santiago Gómez[2]
Center for Research in Mathematics, San Lorenzo, Paraguay
Fredy Ramírez[3]
College of Science and Technology, Computer Engineering, UNCA, Coronel Oviedo, Paraguay
Héctor Estigarribia[4]
College of Science and Technology, Computer Engineering, UNCA, Coronel Oviedo, Paraguay

1 Abstract

Dengue is a viral disease transmitted by the *Aedes aegypti* female mosquito, affecting vast areas of the world. In the last 50 years, its incidence in Paraguay has increased, accompanying the persistent migration into the cities [1]. Approximately 80 million cases appear every year in more than 100 countries, and about 2.5 billion people live in countries with endemic dengue. Paraguay is part of this list of countries, as one of the most affected by the disease [3].

Since the appearance of dengue in Paraguayan territory there has been a scalar increase in policies, strategies and public health services that prevent and combat the outbreaks. Despite all these efforts, large epidemics were recorded in the 1988-1989; 1999-2000; 2006-2007 and 2012-2013 periods [1]; and currently there are many cases of the disease in the country.

In this work we propose a model to forecast the number of probable dengue cases using two techniques in tandem. First, we implement a novel technique for feature selection using **Multivariate Symmetric Uncertainty** (MSU) [2], which we employ to compare feature sets. Secondly, the selected feature sets are used to feed a deep learning neural network.

2 Datasets and Methodology

We applied both techniques on data sets corresponding to the time interval between 2009 and 2014. Dengue notification data were originally compiled by the General Direc-

[1] maortega@fctunca.edu.py

[2] sgomezpy@gmail.com

[3] framirez@fctunca.edu.py

[4] hestigarribia64@fctunca.ed

torate of Health Surveillance in Paraguay, and climate data were downloaded from the *www.wunderground.com* website and complemented with local meteo data.

These two data sets contain the dengue fever cases grouped by district and epidemiological weeks, and the history of weather by date and location respectively; for example temperature (in °C), humidity (in percentage) and precipitation (in millimeters).

The **Multivariate Symmetrical Uncertainty** is an entropy-based measure of correlation [2] defined as the total correlation between n categorical variables, normalized by the sum of all individual entropies of those variables. MSU makes possible to choose the most informative features because it can detect bivariate and multivariate correlations between the variables. The measure depends on the number of features, their informativeness, their cardinalities and the sample size. Values of MSU range from 0 to 1; the more informative the features under consideration, the closer the value of MSU gets to 1.

Deep learning (DL) is a branch of machine learning based on artificial neural networks. It is used as a tool to learn from data about many phenomena and its main applications are speech recognition, computer vision and natural language processing [4]. The main strength of deep learning is its robustness with high-capacity models having many parameters. There are frameworks for DL in various programming languages; in this work we use Keras, a library in Python.

3 Conclusions

The utilization of MSU facilitates feature selection for Deep Learning. Selecting really relevant feature sets imply more accurate predictions with lower error rates. Some of the error rates get as low as 1%, collecting only the relevant features. Given the fast computational times allowed by the MSU plus DL combination, these results can be useful for better and more responsive decision making in the fight against dengue.

References

[1] Dirección General de Vigilancia de la Salud. *Boletín Epidemiológico Semanal.* URL *http://www.vigisalud.gov.py/boletin_epidemiologico.* Paraguay, 2009-2014.

[2] G. Sosa-Cabrera, M. García-Torres, S. Gómez, C. Schaerer and F. Divina, *Understanding a Version of MSU to assist in Feature Selection.* In: Proceedings of the 4th Conference of Computational Interdisciplinary Science (CCIS 2016), Sao José dos Campos, Brazil.

[3] World Health Organization. *A description of the reality of the dengue cases around the world.* OMS Dengue, Organización Mundial de la Salud, 6 July 2017.

[4] I.H. Witten. *Data mining: Machine Learning Tools and Techniques*, Elsevier, Amsterdam, 2017.