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

Marcos Ortega, Santiago Gómez – CIMA.

maortega@fctunca.edu.py, sgomezpy@gmail.com

Introduction

According to the OMS 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.

According to DGVS 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.

Objective

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 Symmetrical Uncertainty (MSU) [2], which we employ to compare feature sets. Secondly, the selected feature sets are used to feed a Deep Learning neural network.

MSU as an Association Criterion

MSU measures multiple correlation among a set of categorical variables. Taking a group of variables including de class, we use MSU to evaluate how strongly correlated the variables are. Correlations closer to 1 indicate good candidate groups. Here is the definition of MSU:

$$MSU(X_{1:n}) \coloneqq \frac{n}{n-1}\left[\frac{C(X_{1:n})}{\sum_{i=1}^{n} H(X_i)}\right]$$

Where,

$$C(X_{1:n}) \coloneqq \sum_{i=1}^{n} H(X_i) - H(X_{1:n})$$

$$H(X_{1:n}) \coloneqq H(X_1, \dots, X_n) \coloneqq -\sum_{x_1} \dots \sum_{x_n} P(x_1, \dots, x_n) log_2[P(x_1, \dots, x_n)]$$

Deep Learning

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 [5].

Datasets And Experiments

The techniques were applied on data sets corresponding to the time interval between 2009 and 2013. Dengue notification data were originally compiled by the General Directorate of Health Surveillance in Paraguay, and climate data came from Paraguayan Weather Service.

['dia', 'mes', 'anio', 'semana', 'institucion_notificacion', 'edad_discret', 'sexo', 'barrio', 'departamento', 'viajo', 'lugar_viajo', 'tuvo_cuadro_similar', 'fecha_cuadro_similar', 'caso_entorno', 'clasificacion_clinica', 'clasificacion_final', 'criterio', 'fallecido', 'sindrome_hemorragico', 'gingivorragia', 'derrame_pleural', 'hemorragias_diversas', 'otros', 'estado_final', 'NS1_AG', 'IgG', 'IgM', 'PCR', 'serotipo', 'alguna_prueba_lab', 'tmax_discret', 'tmin_discret', 'tmed_discret', td_discret', 'pres_est_discret', 'pres_nm.discret', 'prcp_dicret', 'hr_discret', 'helio_discret', 'nub_discret', 'vmax_d_discret', 'vmax_f_discret', 'vmed_discret', 'distrito_notif', 'id_estacion']

Results

After applying MSU on all_data, 2009, 2010, 2011, 2012 and 2013 data sets, the most relevant features found were:

['semana', 'tuvo_cuadro_similar', 'caso_entorno', 'clasificacion_clinica', 'clasificacion_final', 'criterio', 'sindrome_hemorragico', 'gingivorragia', 'derrame_pleural', 'hemorragias_diversas', 'otros', 'NS1_AG', 'IgG', 'IgM', 'PCR', 'serotipo', 'alguna_prueba_lab', 'tmax_discret', 'tmin_discret', 'tmed_discret', 'td_discret', 'pres_est_discret', 'pres_nm_discret', 'prcp_dicret', 'hr_discret', 'helio_discret', 'nub_discret', 'vmax_d_discret', 'vmax_f_discret', 'vmed_discret', 'distrito_notif', 'id_estacion']"

This is the MSU behavior over the five years:

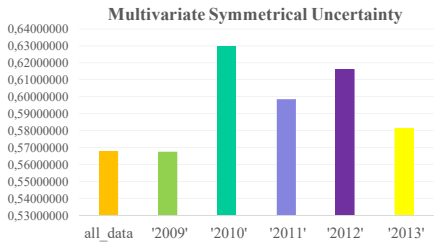

Figure 1: MSU values each year.

Applying Keras on some datasets of these features, we obtained a Deep Learning model with a limit of three epochs without improving that predicts dengue cases per epidemiological week with an average accuracy of 90.10%, and less time of processing in comparison with the original dataset.

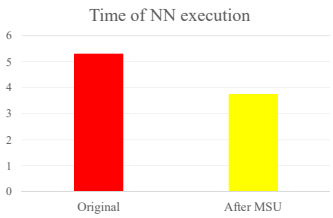

Figure 2: time of execution in minutes of a NN.

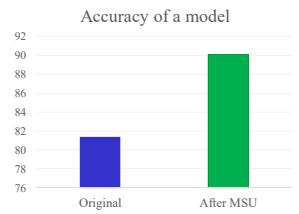

Figure 3: accuracy in percentage of a NN model

Conclusion

The utilization of MSU facilitates feature selection for Deep Learning. Selecting really relevant feature sets imply more accurate predictions with lower error rates when a good is used. Given the fast computational times allowed by the MSU plus Deep Learning combination, these results can be useful for better and more responsive decision making in the fight against dengue.

References

[1] Direccion General de Vigilancia de la Salud. Boletn Epidemiologico Semanal. URL http://www.vigisalud.gov.py/boletin epidemiologico. Paraguay, 2009-2014.
[2] G. Sosa-Cabrera, M. Garca-Torres, S. Gomez, C. Schaerer and F. Divina, Under-standing a Version of MSU to assist in Feature Selection. In: Proceedings of the 4th Conference of Computational Interdisciplinary Science (CCIS 2016), Sao Jose dos Campos, Brazil.
[3] World Health Organization. A description of the reality of the dengue cases around the world. OMS Dengue, rganizacion Mundial de la Salud, 6 July 2017.
[4] I.H. Witten. Data mining: Machine Learning Tools and Techniques, Elsevier, Ams-terdam, 2017.
[5] «Keras Documentation». [On Line]. Available in : https://keras.io/. [Acceded: 14-aug-2018].

Thanks

The authors thank PROCIENCIA-CONACyT-Paraguay. This work is supported by CIMA through research grant PINV15-706 from CONACyT.