Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Feature Selection for Dengue using Principal Component Analysis

Julia Figueredo[1]
Christian E. Schaerer
Santiago Gómez Guerrero[2]
Gustavo Sosa-Cabrera
Polytechnic School, National University of Asuncion, San Lorenzo, Paraguay
Alejandra Rojas
Cynthia Bernal
Fátima Cardozo
Institute for Health Science, National University of Asuncion, San Lorenzo, Paraguay
Teresita Báez
ASESTPY, Paraguayan Association of Statisticians, Asunción, Paraguay

Dengue fever is one of the top ten global health threats, is endemic in more than 100 countries [1]. In Paraguay, dengue has been endemic since 2009. When an outbreak occurs, the number of non-recorded and under-registered cases increases due to the fact that confirmatory analyses are, specially for developing countries, complex and expensive. For that reason, these laboratory tests are not fully available. In these situations, detection of dengue cases becomes an important issue. The problem consists of establishing some variables (associated with traditional laboratory results) to determine a positive and a severe case of dengue [2].

For this type of prediction, this work proposes to analyze the association between laboratory variables (from ambulatory and hospitalized patients) under suspicion of dengue in order to identify which variables (or combination thereof) best predict positive and severe dengue cases.

The data set was extracted from medical records and consists of 72 laboratory variables that also include, for experimental reasons, positive dengue tests, and its severity (OMS, 2009). The variables were grouped by day, and standardized so that they had a 0 mean and a standard deviation of 1.

Principal Component Analysis (PCA) was run on three subsets of the data, aiming at identifying the proper variables for feature selection [3–5]. Records with missing data on the variables under consideration were previously eliminated. In each of the three experiments or runs, the best 3 components were taken.

The table shows the results obtained from the selection of 16 blood counts and clinical variables recorded on days 1, 2 and 3 of the disease. The first experiment takes data from 93 patients (50 positive, 43 negative dengue on day 1 of the symptoms); for these the PCA run explains 65.14% of the accumulated variability and the variables are Absolute Neutrophils (PC1), Relative Neutrophils (PC2) and Hemoglobin (PC3).

For the second experiment, 267 patients (125 positive, 142 negative dengue on day 2 of symptoms). In this run PCA explains 56.17% of the accumulated variance and the most representative variables of each component are Absolute Neutrophils (PC1), Relative Lymphocytes (PC2) and Erythrocytes (PC3).

[1] naty-1804@fpuna.edu.py
[2] sgomezpy@gmail.com

Variables highlighted by principal component

Experiment	Variables	Patients	Positive	Negative	Day of symptoms	% Explained variance	PC 1	PC 2	PC 3
1	16	93	50	43	1	65.14%	Absolute Neutrophils	Relative Neutrophils	Hemoglobin
2	16	267	125	142	2	56.17%	Absolute Neutrophils	Relative Lymphocytes	Erythrocytes
3	16	243	125	118	3	55.78%	Absolute Neutrophils	Absolute Lymphocytes	Hemoglobin

For the last experiment, 243 patients (125 positive, 118 negative dengue on day 3 of symptoms). In this execution, PCA explains 55.78% of the accumulated variance and the most representative variables of each component are Absolute Neutrophils (PC1), Absolute Lymphocytes (PC2) and Hemoglobin (PC3).

So far, the variability explained by PCA in the first three days of the disease justifies further variable and day group selections in order to identify attributes that allow a higher percentage of variance to be explained. In additional runs, this can be accomplished by keeping the variables with high contribution and bringing into the group the most promising of the remaining variables, which can be completed within a reasonable time frame given the relatively low number of variables in the dataset.

As the PCA tool requires all data cells to be non-empty, for future studies more data records with the full set of 8-day observations will be needed in order to have more solid results. Notably, the already known platelet counts variable got eliminated in our dataset because of too many missing data in those critical records.

References

[1] OPS(Organización Panamericana de la Salud)/OMS(Organización Mundial de la Salud. **Módulo de Principios de Epidemiologia para el control de enfermedades.** 2ª ed. N.W. Washington, D.C., 2001. https://www.paho.org/col/dmdocuments/MOPECE5.pdf.

[2] A. Rojas, F. Cardozo, C. Cantero, V. Stittleburg, S. López, C.M. Bernal, F. Giménez, L.P. Mendoza, B. Pinsky, Y. Guillén, M. Páez, and J. Jesse. "Characterization of dengue cases among patients with an acute illness, Central Department, Paraguay." In: **PeerJ** (2019). URL: https://doi.org/10.7717/peerj.7852.

[3] S. Gómez-Guerrero, M. García-Torres, G. Sosa-Cabrera, E.G. Sotto-Riveros, and C.E. Schaerer. "Classifying dengue cases using CatPCA in combination with the MSU correlation". In: **Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May, MDPI**. 2021. DOI: 10.3390/Entropy2021-09828.

[4] S. Gómez-Guerrero, I. Ortiz, G. Sosa-Cabrera, M. García-Torres, and C.E. Schaerer. "Measuring Interactions in Categorical Datasets Using Multivariate Symmetrical Uncertainty". In: **Entropy** (2022). URL: https://doi.org/10.3390/e24010064.

[5] R. Arias-Michel, M. García-Torres, C.E. Schaerer, and F. Divina. "Feature Selection Using Approximate Multivariate Markov Blankets". In: **Hybrid Artificial Intelligent Systems**. Ed. by Francisco Martínez-Álvarez, Alicia Troncoso, Héctor Quintián, and Emilio Corchado. Cham: Springer International Publishing, 2016, pp. 114–125. ISBN: 978-3-319-32034-2.