A Machine-learning Mobile App to support prognosis of Ebola Virus Diseases in an evolving environment

Supervisers: Anne-Marie Hartley, Martin Jaggi

Sina Fakheri

May 2017

1 Introduction

Ebola Virus Disease (EVD) is a strong deadly virus. It works by attacking directly the immune system. It disables the cells detective mechanism and reprograms it so the cells become a virus production machine [1]. Despite it's discovery near Ebola river in DRC in 1976, its impact was quite local and limited. However this changed in 2013 where it became a global epidemic until 2016. It infected 28700 persons among which 11325 died (mostly from West Africa) [2]. Very recently, Dr Chan from WHO stated that another Ebola epidemic in future is probable [3]. Therefore, we need to create conditions where we are better prepared for such an event. Here is where technology can be of great help.
The lack of high standard medical structures in West Africa , the fast rate at which the virus spreads and the absence of EVD specific symptoms implies to have cheap and easy methods to deal with Ebola-suspect cases. More specifically, we need an efficient and robust method for triage and quarantines. The current triage protocol misclassifies more than half of the cases, putting at risk of death healthy people while depriving Ebola-affected patients for treatement.

Current state-of-art solutions are both rare and limited. Hartley et al.[4] proposed a scoring system for triage which is based in statistical modeling. The major limitation is due to the fact that this score is static and is not easy to adapt to new variables like a new behaviour of the virus, a new treatment center in another location, etc. A Machine-learning based solution intrinsically take these changes into account and it's performance generally improves as more data is available.

We propose a solution via a smartphone app, based on machine learning techniques, to predict the risk of a person, based on his current symptoms and health state, being EVD+ (diagnosis) and also, in the case, his survival chance

(prognosis). This will improve both resource management and EVD+ detection precision which in turn slows down the propagation rate of EVD. This solution is cheap, highly portable and easy-to-use. Health care assistants just have to install our app on their Android smartphones. Our solution is highly dynamic and self adapts to eventual virus mutations in time and location specificities.

2 Methodology

2.1 Clinical dataset

We used the GOAL dataset from the study conducted by Hartley et al.[4]. "This retrospective cohort study uses anonymized patient data collected between December 14, 2014 and November 15, 2015 at the GOAL ETC in Port Loko, Sierra Leone. Data comprised patient demographics, geographic location, clinical signs and symptoms, and laboratory results (for malaria infection and semi-quantitative Ebola viremia), as well as the final patient out- come of death or survival." It contains 575 patiens. The evd_ct (inversely proportional to viral load in patient's body) feature contains some missing values. This concernes the EVD- cases. For prognosis, all these cases were removed from dataset, bringing down our sample size to 144 patients.

We also had access to another dataset, from data collected in a district hospital in Sierra Leone, Kenama Government Hospital (KGH). The initial idea was to use this dataset for our external validation. Finally, because of the great number of missing values (more than 80%) and the low number of samples in this dataset, we decided not to use it as such few number of samples can not be of real help to externally validate our model. Additionally, Doing imputation on very few available data gives generally poor results. (idea to be confirmed by Annie)

2.2 Missing data and imputation

The three following features contain missing values : 'malaria,'referral_time' and 'quarantine'.
The total number of rows with missing values was relatively small (26 rows). Nevertheless, We once ran our models with the missing values removed and another time imputed. We found that imputing results gives a model that performs slightly better on the validation set. Moreover, imputation by mean and KNN imputation were tested and the results were very similar. This can be explained by the low proportion of missing values and the fact that the concerned features were not among the most prevalent for the classification. Finally, for simplicity, we retained the imputation by mean option.

2.3 Preprocessing and feature selection

The dataset has two set of features: the first one concerns information and measurements gathered at arrival, in the Ebola Treatment Center (ETC), and the

second one is the same set of information gathered after some days (determined by 'days_admitted') spent in ETC. This approach allows us to better understand how the evolution of certain measures influence the death and also help us to see if the same subset of features are the most determinant for each prognostic model.

2.4 Feature engineering

A case of interest for health experts is the difference between prediction models among different locations and different periods on time, i.e. population selection. This allows to see the evolution of the virus over time (observe eventual virus mutations). Additionally, it helps to adjust the treatment with the specifities of each region. To integrate this ability to our model, we artificially added a location categorical feature to the dataset: First using Numpy's random.choice function, we attributed a location (from an array of 5 locations) to each patient. Then, to be able to use this information in our prediction models, we used DictVectorizer package from sklearn.feature_extraction library which allows to add as many binary features as we have distinct location. For each patient, only one of the newly created binary features has a value of 1. It's then straightforward to select the population corresponding to a specific location and create the prediction model based upon it.

2.5 Algorithms used

For our prediction model on prognosis, We have tried four common Machine learning algorithms: Logistic Regression, Random Forest, SVM and Neural Networks. For each of them, we have used grid-search using 10-fold cross-validations to fine tune the hyperparameters and obtain the best model (based on roc-auc score).

For Logistic regression we optimized :

- C (inverse of regularization strength)
- Class-weight : to balance the two classes
- Penality : Lasso (l1) or Ridge (l2)

For the random forest we used n_estimators, min_samples_leaf and max_features.

Finally for NN, we used only number of hidden_layer_sizes which includes both number of hidden layers and number of nodes in each layer

We have used Python as programming language and Scikit-learn library for the machine learning algorithms.

3 Results

4 Discussion

References

[1] The Ebola Virus Explained, 2014
http://kurzgesagt.org/work/the-ebola-virus-explained/

[2] Outbreaks Chronology: Ebola Virus Disease
https://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html

[3] Opening remarks at the Ebola vaccines for Guinea and the world event
http://www.who.int/dg/speeches/2017/ebola-vaccines-guinea/en/

[4] Hartley M-A, Young A, Tran A, Okoni-Williams H-H, Suma M, Mancuso B, et al. *Predicting Ebola Infection: A Malaria-Sensitive Triage Score for Ebola Virus Disease* PLoS Neglected Tropical Disease. 2017.