STAT-221: Project

Kevin Kuate Fodouop
Harvard University

Abstract

The Ebola epidemics which has affected West Africa since March 2014 is the most dramatic since the discovery of the virus in 1976, and has generated deep attention and fear among the international community. On several aspects, the response of the WHO has been categorized as too slow or lacking the needed means [2]. This paper attempts to examine subnational data in county touched by Ebola to categorize the intensity of a potential virus outbreak in the region (higher spread or lower spread). We adapt the Grade of Membership (GoM) model studied in Caldas de Castro et al. [5], and assess the relevance on our model based on a previous study of regional Ebola spread [3].

1 Literature Review

This project uses Grade of Membership (GoM) models, and is based on former a study of malaria epidemics using Mixed Membership Models (MMM) [5]. The objective in [5] is to categorize regions as high risk or low risk for malaria, and use this information to optimize allocation of resources to prevent the disease's spread.

MMM are a flexible clustering model, meaning that every point partially belongs to all clusters to a certain degree (in a similar manner as in fuzzy c-means clustering). If the number of cluster is K, each subject has a membership vector $g^{(i)} = (g_1^{(i)}, ..., g_K^{(i)})$ (components adding to 1). The probability of observing a response variable $x_j^{(i)}$ for this subject, given θ_k parameters of the response distribution in each cluster, is $Pr(x_j^{(i)}|g^{(i)}) = \sum_k g_k^{(i)} f(x^{(i)}{}_j|\theta_{kj})$ [6].

The different variables in the response X are considered independent given the GoM score g. [5] also assumes all variables are categorical, so that θ_{kj} will be parameters of a multinomial distribution with $n = 1$. Having $K = 2$, corresponding to low or high profiles, the likelihood for the model reduces to

$$\mathcal{L}(X, G, \Theta) = \prod_{i=1}^{N} \prod_{j=1}^{P} \left[g_{L,i} \prod_{m=1}^{d_j} (\theta_{L,j})_m^{1_{x_j^{(i)}=m}} + g_{H,i} \prod_{m=1}^{d_j} (\theta_{H,j})_m^{1_{x_j^{(i)}=m}} \right] \tag{1}$$

with d_j the number of levels for feature j. The optimal GoM g and distribution parameter θ is obtained by stepwise MLE, detailed in algorithm 1.

Data: Observed categorical data X
Result: MLE of parameter Θ and GoM g.
initialization: $g = g_0$, $\Theta = \Theta_0$ initial guesses (can be derived from former results).
while $|\mathcal{L}(X, G^{(k)}, \Theta^{(k)}) - \mathcal{L}(X, G^{(k-1)}, \Theta^{(k-1)})| > \epsilon$ **do**
- Update step: k = k + 1
1. $G^{(k)} \leftarrow MLE(G)$ with other parameters fixed.
2. $\Theta^{(k)_L} \leftarrow MLE(\Theta_L)$ with other parameters fixed.
3. $\Theta^{(k)_H} \leftarrow MLE(\Theta_H)$ with other parameters fixed.

end

Algorithm 1: 3 step MLE for GoM fitting.

One main challenge of GoM model may be to determine the proper number of clusters K [7], but in [5] it is inherently fixed at 2 by the objectives of the problem (predicting low or high profile of malaria risk). It is however less explicit to determine from which domains select variables for X, and how many level to defined for each of these variables. The categorization is indeed here more driven by "subject matter interpretations of meaningfully different levels of risk" more than on statistical cateogorization, such as quantiles. On that side, a novelty introduced by the paper is to distinguish two domains of features which will determine risk profiles, environmental features and behavioral/economic features.

Most importantly, the categorization is not based on the expected result of the clustering (malaria risk for [5], but on other variables that are likely to influence the result metric. To assess the quality of the resulting high-risk low-risk clustering, the authors of [5] plot g_{high} against the exposure-weighted malaria illness rate (EWR), which is a metric representing risk of malaria in the region.

2 Application to Ebola Propagation

The Ebola epidemics that broke out in March 2014 is still progressing at a steady rate in West Africa (more than 1,000 new cases a week). The 3 mainly affected countries are Guinea, Sirrea Leone and Liberia, which suffer from poor health infrastructures. After a few months of virus propagation, data on the epidemics is available, and organization such as Statistics without Borders have made them publicly available and standardized. This project explores the application of Grade of Membership (GoM) model fitted to geographic regional data to predict the speed of spread (`high-speed` or `low-speed`) after an Ebola outbreak in a region.

We adapt the Mixed Membership Models (MMM) introduced in [6], and used in [5] to predict the high or low risk of malaria propagation in regions of the Amazon. The actual low-speed or high-speed Ebola propagation is based on a former data science study of Ebola cases prediction [3]. This project uses the SIR model, an epidemiology model for understanding complex dynamics of disease propagation systems [4]. Figure 1 shows 2 shapes of ebola cases and deaths time series. The SIR model computes estimates for contamination, recovery and death rates, and implies an exponential evolution of cumulative number of cases. Hence the region `GN-8` of Guinea would be suited for the SIR model, and categorized `high-speed`, whereas `GN-38` would be better fitted by an S-curve model, and is categorized `low-speed`.

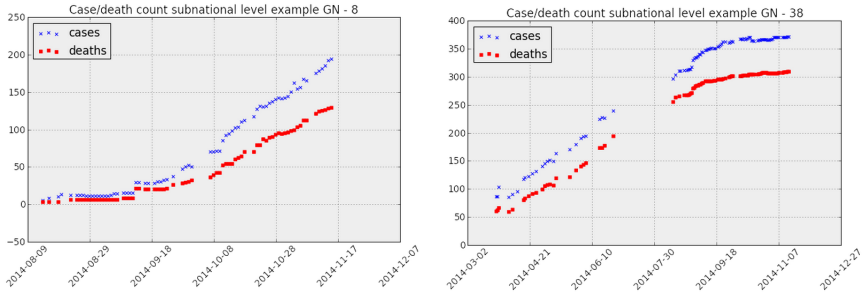

Figure 1: *Regions with cases shape adapted to SIR model (left), and not adapted to SIR model (right). Low model fit corresponds to a S-curve behavior, instead of exponential.*

The SIR model only uses the cases (or deaths) time series data, and do not take into account regional features, such as quality of health infrastructures or access to clean water. Figure 2 displays a SIR fit for cases and deaths time series, for a region adapted to the model framework.

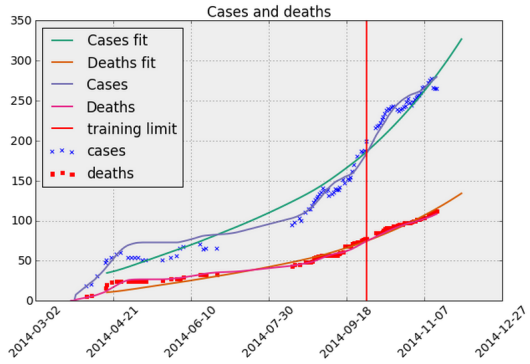

Figure 2: *Fit of the SIR model on a region showing exponential behavior in cases and deaths numbers.*

The Ebola virus is transmitted to humans from wild animals, but then spreads in populations through human-to-human transmission. Reducing that human-to-human transmission is one the main objective of the WHO action [1]. The WHO action focuses on quick response to Ebola outbreaks, and acute monitoring of those outbreaks. However, few measures seem to be taken to preventively strengthen health infrastructures or sanitation procedures in region not yet touched by Ebola, but which would suffer fast propagation of the virus [2]. The goal of this project is to use GoM models to predict the intensity of potential future outbreaks for that purpose.

As in [5], we use categorical features, and obtain a likelihood function of the form (1). The main challenge was to assess which variables to include in the model, and how to define the corresponding categories (starting with the number of such categories). Lacking the empirical research and subject matter expertise on feature importance for disease propagation, we adopted a statistically driven approach for the definition of such categories. For a defined number of categories, levels are computed comparing the observed value to the quantiles on the observations. The definition of such features to be included and their number of level is called a *spread profile schema*.

To assess the quality of a *spread profile schema*, we compute a commonly used goodness of fit metric for the resulting clustering, defined as

$$R_{clust} = \frac{sep_{inter}}{var_{intra}} \tag{2}$$

sep_{inter} being the inter-centroidal separation and var_{intra} the intra-cluster variance. An adapted metric for fuzzy clustering is also computed. Results for some *spread profile strucure* are documented in Table 1.

As a way to obtain the optimal *spread profile schema* for our problem, we tried implementing a Genetic Algorithm method. Its chromosome structure encodes the number of levels for each features, and its fitness function is the clustering goodness of fit defined in (2) (fuzzy version). However we failed to define a crossover function that would efficiently explore the relevant chromosome space.

Adapting figure 2 of [5], we ploted GoM versus R-square obtained with the SIR model, grouped by country. Plots for bicategory (all category have 2 levels) schema is displayed on figure 3.

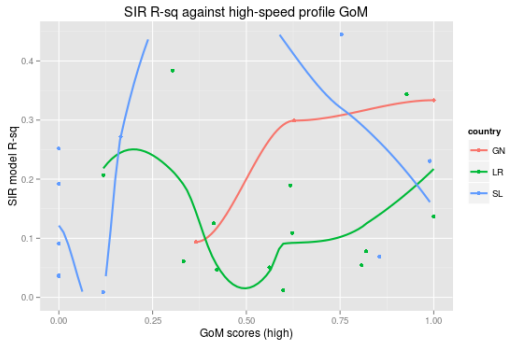

Figure 3: *Subareas SIR model fit (R^2) as a function of GoM scores (degree of belonging to* `high-speed`*).*

3 Implementation

Features were obtained from different sources, mainly the qDatum public sources repository on Ebola [8]. Preprocessing and merging from the several sources was needed before obtaining the final features data in `dat/merged_covariate_df.csv`.

The results of SIR model fits is found in `dat/SIR_fit.csv`. It contains R^2, MSE goodness of fit metrics, and the number of observations. For assessing the relevance of our clustering results, we use the R^2 metric.

The structure of the code is as follows:

- `dat_format.R`: Helper functions to convert data in the relevant categorical format, having specified features to be included and number of levels.
- `MLE.R`: Functions performing the MLE optimization.
- `SIR_fit.R`: Functions to assess the quality of a clustering result and plot metric against GoM g, given the SIR regional fits.

- `local_optim.R`: Test GoM model locally, on first spread profile schemas. Also computes the matrix of suggested schemas to be tried on Odyssey.
- `local_ga.R`: Attempt to use Genetic Algorithm to determine best schema.
- `schemas.R`: Data visualization of some feature distributions, and compute the schemas to be tried on Odyssey.
- `mle_task.R`: Try 100 schemas on 10 machines with Odyssey, triggered by `sbatch -array=1-10 mle.slurm`.

To perform the constraint optimization on the θs, we use the tranformation between the simplex and $\mathcal{R}^{d_p}$:

$$\xi_j = g(u_j) = log(u_j) - \sum_{j=1}^{d_p} log(u_j), \quad u_j = h(\xi_j) = \frac{e^{\xi_j}}{\sum_{j-1}^{d_p} e^{\xi_j}}$$

u being a column of one θ_p, for low or high profiles. We have $h(g(u_j)) = u_i$. After transformation we can use R's `optim` function to perform the log-likelihood maximizations for the different steps described in algorithm 1.

The advantage of this design is that it makes it very easy to change the *spread profile schema* (just change arguments of functions in `dat_format.R`, and then assess the quality of the model. One direct drawback from this implementation, refering to [7], is that the categorization only relies on statistical quantiles, whereas subject specific empirical knowledge would be preferable. Hence even though it did not succeed, the motivation itself of the Genetic Algorithm approach can be challenged (as exogeneously derived categories would be better according to [7]).

4 Conclusion

References

[1] *Ebola Virus Disease*, by WHO Media Centre.

[2] *Ebola Response Roadmap*, by WHO.

[3] *Predict Ebola!*, by Kristen Altenburger, Manuel Andere, Guillaume Sabran, and Shiya Wang.

[4] *Compartmental models in epidemiology*, by Wikipedia.

[5] *Malaria Risk on the Amazon Frontier*, by Marcia Caldas de Castro, Roberto L. Monte-Mor, Diana O. Sawyer, and Burton H. Singer.

[6] *Introduction to Mixed Membership Models and Methods*, by Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva, Stephen E. Fienberg.

[7] *Interpretability Contraints and Trade-offs in Using Mixed Membership Models*, by Burton H. Singer, Marcia C. Castro.

[8] *http://www.qdatum.io/public-sources*, by qDatum.