Optimizing Global Influenza Surveillance for Locations with Deficient Data (Student Abstract)

Songwei Shan, [1,2] ‡ **Qi Tan,** [1,2] ‡ **Yiu Chung Lau,** [1,2] **Zhanwei Du,** [1,2] **Eric H.Y. Lau,** [1,2] **Peng Wu,** [1,2]
Benjamin J. Cowling [1,2] *

[1]World Health Organization Collaborating Centre for Infectious Disease Epidemiology and Control, School of Public Health,
The University of Hong Kong, Hong Kong SAR, China
[2]Laboratory of Data Discovery for Health, Hong Kong Science and Technology Park, Hong Kong SAR, China
shansongwei@connect.hku.hk, {tanqi, chunglau, zwdu, ehylau, pengwu, bcowling}@hku.hk

Abstract

For better monitoring and controlling influenza, WHO has launched FluNet (recently integrated to FluMART) to provide a unified platform for participating countries to routinely collect influenza-related syndromic, epidemiological and virological data. However, the reported data were incomplete. We propose a novel surveillance system based on data from multiple sources to accurately assess the epidemic status of different countries, especially for those with missing surveillance data in some periods. The proposed method can automatically select a small set of reliable and informative indicators for assessing the underlying epidemic status and proper supporting data to train the predictive model. Our proactive selection method outperforms three other out-of-box methods (linear regression, multilayer perceptron, and long-short term memory) to make accurate predictions.

Introduction

Influenza is one of the most prevalent diseases around the world, resulting in substantial morbidity and mortality every year worldwide. To effectively monitor and control influenza, WHO has launched FluNet in 1997 (which was integrated into FluMART) (Hamid, Bell, and Dueger 2017) to provide a global platform to share influenza epidemiological data and virological data.

However, the reported data are incomplete. On FluMART, 153 out of 196 countries (78%) can only provide both syndromic and virological data in less than 50% weeks between 2010-2020. For example, as illustrated in Figure 1, Austria started reporting from 2010 while Armenia started reporting from 2011, and Austria suspended reporting occasionally since 2014.The incomplete data makes it difficult to assess the epidemic status of influenza in different countries (Pei et al. 2021). In order to address this problem, we propose a novel surveillance system based on multiple data sources to accurately assess the epidemic status of different countries, especially for those without surveillance in some periods.

There are two advanced features in our proposed surveillance system:

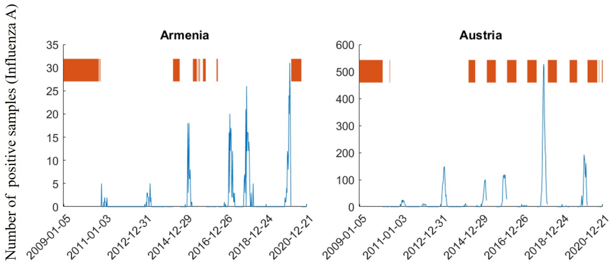

Figure 1: Availability of Influenza A report data(number of influenza A positive samples) in Armenia and Austria. Red indicates the periods when the report data is missing.

1. Robust estimation using reliable indicators. Given the rich amount of the available indicators, the surveillance system is able to automatically select a small set of reliable and informative indicators for assessing the underlying epidemic status;
2. Adaptation to concept drift. The number of reported cases is dependent on both clinical criteria and clinical organization. This surveillance system is able to automatically select data to train the predictive model to reduce thus a bias.

Proposed Method

Multi-Source Predictive Model

Let $X \in \Re^{C\times D\times T}$be the data tensor, where C is the number of countries, D is the number of the attributes and T is the length of time slots. Specifically, $X_{[i,:,:]}$ is the i-th slice of the tensor corresponding to the data of the i-th country of all attributes and all time slots, and similarly $X_{[i,j,t_1:t_2]}$ is data of the j-th attribute and i-th country from t_1 to t_2 time slot. Without loss of generality, we aim to develop a predictive model to estimate the unknown Influenza infections (the $fcol$-th attribute) in the i-th country using all available data:

$$X_{[i,fcol,t]} = F(z_t), z_t = S(X_{[:,:,:t]}) \qquad (1)$$

where S denotes the feature generation function and F denotes predictive function. As the report system is not always reliable, there are many missing values in the data

‡The first authors contributed equally to this paper.
*Corresponding author: bcowling@hku.hk

tensor X, which poses great challenges to predictive analysis. To generate the reliable features fed into the predictive model, we select several indicators that are available at all time slots. We denote the set of these supporting indicators as $K = [(c_1, a_1), (c_2, a_2), \ldots]$. For predicting the target influenza value at time slot t, the supporting indicators of the previous t_ξ time slots are informative and thus we vectorize them as the features:

$$z_t = S(X_{[:,:,:t]}) = vec(X_{[K,t-t_\xi:t]}) \tag{2}$$

We use a linear model to model the predictive function, because it is efficient for training and capable of modeling the cross-region influence. The mathematical equation of the predictive model is written as:

$$X_{[i,fcol,t]} = F(z_t) = W z_t \tag{3}$$

Model Training

The dimension of the size of the feature would be large, which may cause the prediction to become less robust to the unexpected value missing of the supporting indicators and noise. Therefore, we would expect that the weighting W is sparse so that only some key supporting indicators are selected in the predictive model. Therefore, we add the l1 norm at the weighting as penalization (Tibshirani 1996). The training objective become:

$$\min \sum_{t}^{T} ||X_{[i,fcol,t]} - W z_t||_2^2 + \lambda ||W||_1 \tag{4}$$

Due to the concepts drift, the samples with large time lags are less informative for predicting the current samples. Therefore, we reweight the samples in the training. The training objective is extended as

$$\min \sum_{t}^{T} u_t ||X_{[i,fcol,t]} - W z t||_2^2 + \lambda ||W||_1 \tag{5}$$

where $u_t = 1/\log(\beta(t' - t) + 2)$ and t' is the testing time slot. The parameter β controls the scale of time vanishing.

We use cross-validation to determine two hyper-parameters λ and β. In the training data, we use the first 80% for learning model and the latter 20% as validation. We use grid search to select the value pair of λ and β with the best validation score. The candidate hyper-parameter values for λ is $[10, 1, 0.1, 0.001, 0.0001]$ and the candidate hyper-parameter values for β is $[1, 0.1, 0.01, 0.001, 0.0001]$.

Performance Evaluation

To evaluate the performance of our proposed method, we apply it on the influenza data of 132 countries from 2009 to 2020 which are collected from FluMART(http://apps.who.int/flumart/Default?ReportNo=12). Figure 2 shows the prediction result on the number of influenza A positive samples, where the blue line indicates the ground-truth surveillance data, the red line indicates the fitting result and the yellow line indicates the prediction result.

We compare the RMSE of our method and three others (the out-of-box linear regression, and two out-of-box deep neural network models [e.g., multilayer perceptron and long-short term memory] (Goodfellow, Bengio, and Courville 2016)) on the prediction of numbers of influenza A positive samples in 132 countries through 2009 to 2020, as shown in Table 1. The result shows that our method has the lowest RMSE than the three others.

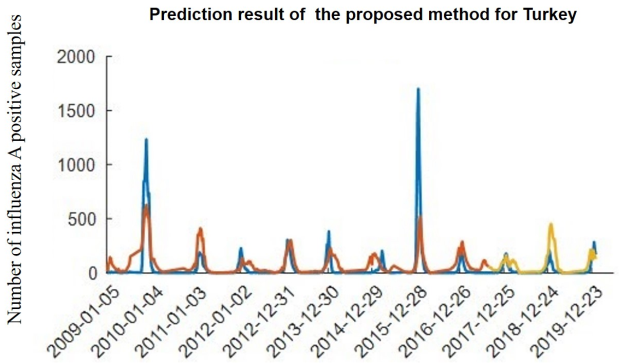

Figure 2: An illustration of prediction result of proposed method on the number of influenza A positive samples. The blue line is the surveillance data, the orange line is fitting result, and the yellow line is the prediction result.

Method	RMSE
Linear without selection	9.3410e+03
Multilayer perceptron	1.1632e+04
Long-short term memory	7.8207e+03
Our method	**6.5383e+03**

Table 1: Prediction error (RMSE) of different methods on Influenza A. The best performance is highlighted in bold.

Conclusion and Future Work

In this paper, we propose a feature selection and temporal weighting model for influenza surveillance and demonstrate its effectiveness. In future we will apply and evaluate this method in the surveillance-related studies.

References

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Hamid, S.; Bell, L.; and Dueger, E. L. 2017. Digital dashboards as tools for regional influenza monitoring. *Western Pacific surveillance and response journal: WPSAR*, 8(3): 1.

Pei, S.; Teng, X.; Lewis, P.; and Shaman, J. 2021. Optimizing respiratory virus surveillance networks using uncertainty propagation. *Nature communications*, 12(1): 1–10.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.