Predicting Influenza Dynamics using a Deep Learning Approach

Shiran Zhong and Ling Bian

University at Buffalo, the State University of New York, 105 Wilkson Quad, Buffalo, NY 14261
Email: shiranzh@buffalo.edu, lbian@buffalo.edu

Abstract

Disease transmission is a complex spatio-temporal process. A great number of approaches have been developed to predict influenza epidemics. Few of them have focused on the temporal dynamics of individual infected locations. Location networks, where locations are nodes and disease flows between them are links, provide a promising approach for such dynamic analyses, but also present challenges. In this study, we employ a deep learning approach to capture the dynamics of disease flows in location networks. We also analyze how the attributes of locations have an impact on the prediction accuracy via a sensitivity analysis.

1. Introduction

Disease transmission is a complex spatio-temporal process (Ferguson et al., 2005; Charaudeau et al. 2014). Among the prevailing approaches to predicting the dynamics of influenza epidemics, few have focused on the transmission at a location-specific scale. Location networks, where locations are nodes and disease flows between them are links, provide a promising basis for such dynamic analyses (Zhong and Bian 2016), but also present challenges as each location might have peaks and troughs of different magnitudes throughout the epidemic (Bian et al. 2012). Conventional approaches are not adequate to capture such complex, dynamic patterns.

The objectives of this study are two-fold. We explore the use of Deep Convolutional Networks (DCN), a deep learning approach, to capture and predict disease flow dynamics represented by the presence of links between locations. We also analyze how the attributes of locations impact the prediction accuracy via a sensitivity analysis.

2. Data and Study Area

The dataset consists of 73 daily disease flow networks of an urban area. In these networks, all 1,026 location nodes remain the same across 73 days. Links can be present or absent depending on the occurrence of disease flow between locations on any particular day. The dataset was obtained from the China Information System for Diseases Control and Prevention.

3. Methodology

To achieve the objectives stated above, the methodology is divided into three parts: the first part describes the principle of DCN; the second part introduces the training and testing processes involved in the DCN; and the last part focuses on the sensitivity analysis.

3.1. Principle of DCN

In contrast to classic neural networks which require good features for supervised training, DCN is a training process where good features could be automatically learned from input data (LeCun et al. 2015). The workflow of DCN in this study (Figure 1) contains three processes: 1) the convolution process, where a convolution filter is applied on the input data (in a format of matrix) in order to amplify the feature signal and suppress the noise; 2) the pooling process, which extracts features that represent location characteristics in the past few days; and 3) the training and testing process, where the learned features are used as input to predict the presence

and absence of links between locations as output (Busseti et al. 2012; LeCun et al. 2015). Predictions are made from past observations and validated with current observations, The process is repeated as the time window moves forward. Of the three processes, training and testing process is further explained below.

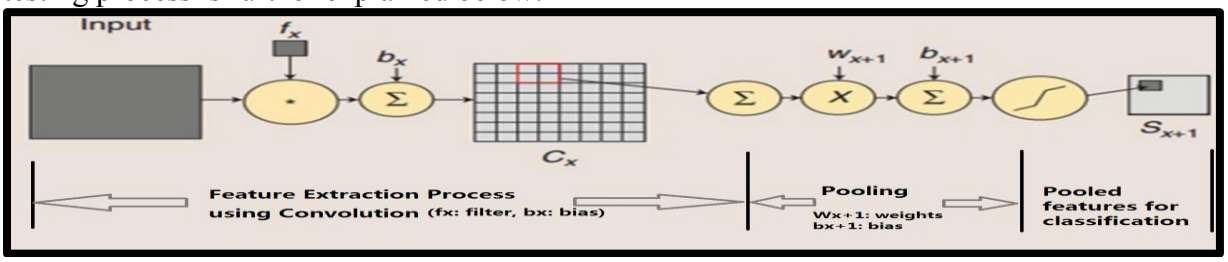

Figure 1. Schematic view of deep convolutional network workflow.

3.2. Training and testing process

The features are represented by 13 attributes of locations in three categories. The first category of attributes describes the epidemiological behavior of locations, such as number of cases at locations, and the previous presence of disease flow between locations. The second category describes the characteristics of location nodes in the networks. These include path length between location pairs, degree, betweenness, closeness, clustering coefficient, eccentricity, bridge, radiality, stress and topological coefficient. The third category takes spatial information into consideration by counting the number of cases at locations' nearest neighbors.

In the training process, we use **DCN** to build an optimized weight matrix $\boldsymbol{W}$, which represents how the presence of link on a certain day is associated with its connecting nodes' attributes in the past few days. The weight matrix W is trained according to Equation 1:

$$\text{Output } (Y_{ijm}) \Leftarrow \text{Input } (X_{ijm-1}, X_{ijm-2}, \ldots X_{ijm-n}) * \boldsymbol{W} \qquad \boldsymbol{(1)},$$

where $\boldsymbol{Y_{ijm}}$ represents the presence/absence of the link between Locations $\boldsymbol{i}$ and $\boldsymbol{j}$ observed on Day $\boldsymbol{m}$; $\boldsymbol{X_{ijm}}$ indicates the input attributes on Locations $\boldsymbol{i}$ and $\boldsymbol{j}$ observed on Day $\boldsymbol{m}$. $\boldsymbol{X_{ijm-1}, X_{ijm-2}, \ldots X_{ijm-n}}$ represent attributes of Locations $\boldsymbol{i}$ and $\boldsymbol{j}$ observed one day, two days, up to n days prior to Day m, respectively. $\boldsymbol{n}$ is the temporal lag, which is initially set as five days. The 73 daily networks are divided into two parts, a training set and a testing set. The training set contains the daily network of the first 50 days, and the remaining 23 days are used as the testing set.

3.3. Sensitivity Analysis

We perform a sensitivity analysis to evaluate the impact of the 13 attributes and the training parameters, i.e. the temporal lag and the length of training/testing sets, on the prediction. The impact of the 13 attributes is evaluated by removing them in two approaches: 1) a conventional approach in which one attribute is removed at a time, yielding 13 scenarios, and 2) an alternative approach in which the exhaustive combination of multiple attributes are removed, yielding 2^13=8,192 scenarios. Regarding the impact of training parameters, the temporal lag $\boldsymbol{n}$ is varied from 3-10 days and the training set is lengthened to 50 to 59 days, while the testing set is shortened accordingly.

Four criteria are utilized to evaluate the accuracy of predicted presence and absence of links: 1) Overall Accuracy (OA): the sum of correctly predicted presence and absence divided by the sum of observed presence and absence over the testing period, 2) Precision of Presence (PP): the correctly predicted presence divided by the total number of predicted presence over the testing period, 3) Precision of Absence (PA): the correctly predicted absence divided by the total number of predicted absence over the testing period, and 4) F1 Score: the balance between PP

and PA as shown in Equation 3(Powers 2011). All accuracy measurements are standardized from 0 to 1.

$$F1\ score = \frac{(2 * PP * PA)}{(PP + PA)} \quad (3)$$

For comparison purposes, the classic neural networks analysis (FNN) is also applied to the same 73 daily networks to predict the present/absence of links, using the same set of attributes, temporal lag, training and testing length division, the sensitivity analysis, and the four accuracy criteria.

4. Results and Discussion

Figure 2 illustrates the prediction results using DCN (a) and classic neural networks (b), with respect to the varying temporal lag and training and testing length division. For the DCN results, OA and PA are above 92%. PP and F1 Score are above 80% when the temporal lag is within six days. As the observed peaks and troughs of the epidemic last approximately 12-14 days, the 6-day half cycle corresponds to the rising slope of the epidemic peak before it declines. The same principle is applicable to the troughs. The training and testing length division does not have a noticeable impact on the prediction results. In addition, the DCN produces results with considerably higher accuracy than that generated by classic neural networks in terms of all four criteria (Figure b).

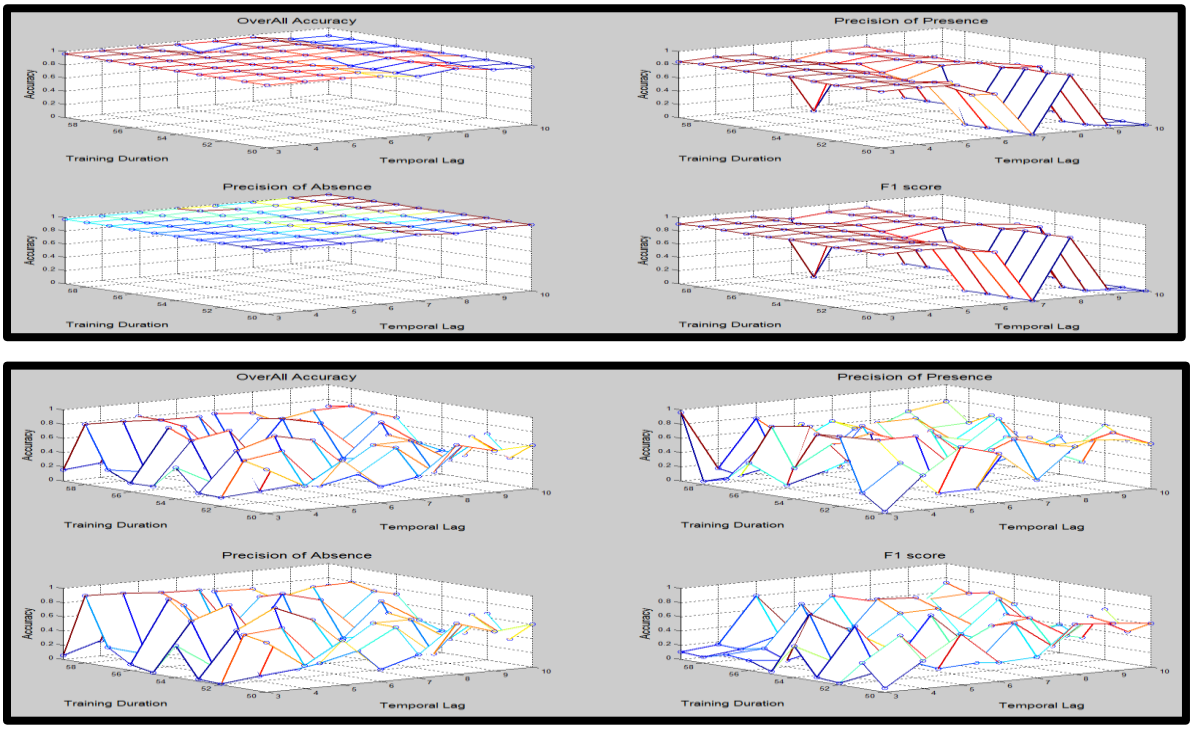

Figure 2. Prediction results using DCN (a) and classic neural networks analysis (b). Each sub figure corresponds to one of the four evaluation criteria: OA, upper left; PP, upper right; PA, lower left, and F1 score, lower right (x-axis: temporal lag; y-axis: length of training period; vertical axis: prediction accuracy).

Figure 3 illustrates the results from the sensitivity analysis regarding the impact of attributes on prediction results, with the F1 score reported as a balanced evaluation. Among the

13 scenarios yielded by removing one attribute at a time, The prediction is most sensitive to the removal of network path length. This attribute measures the length of disease flow pathways.

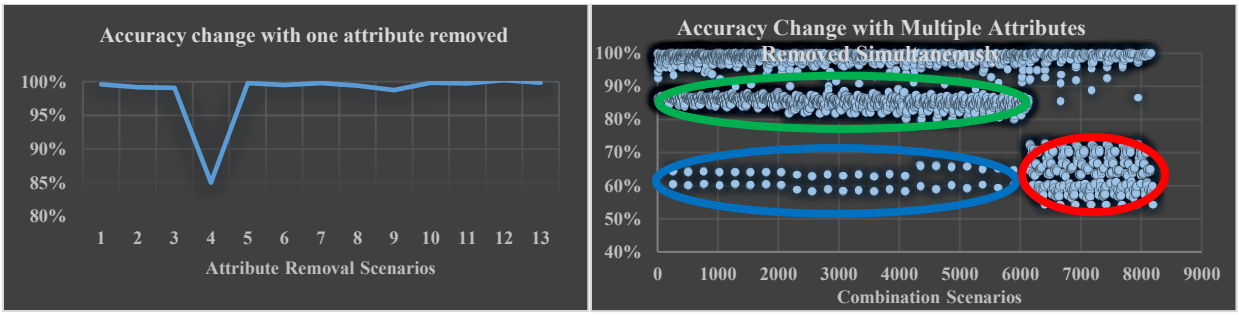

Figure 3. Sensitivity analysis results with respect to the 13 attributes: a) 13 scenarios with one attribute removed at a time and b) 8,192 scenarios with multiple attributes removed at a time (x-axis: scenario ID; y-axis: prediction accuracy).

Prediction results from the 8,192 scenarios fall into three groups (Figure 3b). The green circle highlights the scenarios whose accuracy is reduced by 17.75% on average, when two attributes, the previous presence of disease flow between locations and the path length, are removed. The blue circle highlights scenarios whose accuracy has decreased by 39.5% on average, when two attributes, degree and closeness, are removed. The red circle highlights scenarios whose accuracy has decreased by 40.87% on average, with two attributes, neighbors' cases and eccentricity, are removed. These six attributes represent the source and pathways of disease flow. The prediction of links is sensitive to these attributes. These findings help develop location-oriented intervention strategies to mitigate the spread of disease, e.g. quarantine at location pairs of short path length.

Acknowledgements

Research reported in this publication was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM108731. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reference

Bian, L., Huang, Y., Mao, L., Lim, E., Lee, G., Yang, Y., ... & Wilson, D. (2012). Modeling individual vulnerability to communicable diseases: A framework and design. *Annals of the Association of American Geographers*,*102*(5), 1016-1025.

Busseti, E., Osband, I., & Wong, S. (2012). *Deep learning for time series modeling*. Technical report, Stanford University.

Charaudeau, S., Pakdaman, K., & Boëlle, P. Y. (2014). Commuter Mobility and the Spread of Infectious Diseases: Application to Influenza in France. *PloS one*, *9*(1), e83002.

Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*,*442*(7101), 448-452.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*,*521*(7553), 436-444.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Zhong, S., & Bian, L. (2016). A Location-Centric Network Approach to Analyzing Epidemic Dynamics. *Annals of the American Association of Geographers*, 1-9.