DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information (supplementary materials)

Introduction

This document is the supplementary material for the original paper DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information, submitted to IAAI 2019. In the following sections, we will elaborate (1) why AI for ILI forecasting; (2) the distribution fitting method introduced in the original paper; (3) datasets for the comparison methods; (4) synthetic training data from simulations; (5) settings of all methods in Experiments of the original paper. (6) related work to epidemic forecasting;

Why AI for ILI forecasting?

We apply artificial neural network method to ILI forecasting for three major considerations: (1) It does not pre-specify the linear relationship between inputs and outputs like statistical time-series models, but let the model to learn the hidden relationship automatically. (2) It relieves the heavy dependency on searching an optimal estimate of the underlying disease model like casual models, as well as the computational intensity in real-time forecasting. It allows a neural network model to learn the hidden information automatically from a region-specific training dataset. Once the model is trained before the season start, it can make real-time forecasting with the minimum computation. (3) The large volumn of synthetic training data generated by causal models in our method allow us to train a more complex ANN model.

Fitting distributions for disease parameters

In $\mathcal{P}(p_E, p_I, \tau, N_I, I_V)$, (τ, N_I) are fitted distributions learned from collected samples through the following way:

Neighbors of VA: ['Kentucky', 'Maryland', 'North Carolina', 'Tennessee', 'West Virginia']. For all states and seasons, there are totally $7 * 6 = 42$ (ar, N_I) samples collected from 2010 to 2016.

Neighbors of NJ: ['Delaware', 'New York', 'Pennsylvania']. For all states and seasons, there are totally $4 * 6 = 42$ (ar, N_I) samples collected from 2010 to 2016.

τ is calibrated using EpiFast by Nelder-Mead (Nelder and Mead 1965) algorithm based on each pair of (ar, N_I).

We fit the samples to several reference distributions, including exponential, normal, uniform, gamma, and lognorm distributions. Then we run the KS-test (the null hypothesis being that the sample is drawn from the reference distribution) to choose a distribution with the highest significance. The best fitted distributions with the significance value for each parameter of each state are shown in Table 1.

Datasets for comparison methods

Our DEFSI method is compared with 5 state-of-the-art methods from ANN methods, statistical methods, and causal methods, respectively. They are *LSTM* (single layer LSTM) and *AdapLSTM* (CDC + Weather data) (Venna et al. 2017) from artificial neural network methods, *ARIMA* (classic ARIMA without exogenous variables) and *ARGO* (CDC + Google data) (Yang, Santillana, and Kou 2015) from statistical methods, and *EpiFast* (Beckman et al. 2014) from agent-based causal models. AdapLSTM, LSTM, ARGO and ARIMA are used for state level forecasting. EpiFast is applied for both state level and county level forecasting.

Google data: The Google correlate terms (available at https://www.google.com/trends/correlate) of each state are queried. Then the Google Health Trends (available at https://trends.google.com/trends/) of each correlated terms for each state is collected and aggregated weekly from $ew40$, 2010 to $ew18$, 2018. Google data is used as surrogate information in ARGO. **Weather data**: We download daily weather data (including max temperature, min temperature, precipitation. Available at https://www.ncdc.noaa.gov/cdo-web/datasets) from Climate Data Online (CDO) for each state and aggregate the daily data to average weekly data from $ew40$, 2010 to $ew18$, 2018. Weather data is used as surrogate information in AdapLSTM.

Synthetic training data from simulations

Large volumes of high-resolution synthetic data are generated by repeating the sampling and simulating process. Let us denote $\Omega = \{(\mathbf{y}_{(i)}, \mathbf{y}^{\mathcal{D}}_{(i)}) \in \mathbb{R}^{\ell \times (K+1)} | i = 1, 2, \cdots, r\}$ as all simulated epicurves, ℓ is the length of the epicurve, K is the number of counties in the state, and r is the total number of simulation runs. These curves are randomly ordered to form a long time sequence of ILI incidence which will be used for training DEFSI model.

Compared with CDC surveillance data, the training dataset Ω is prominent in two aspects: (i) it includes high-

Table 1: Marginal distributions of each parameters

Parameter	State	Name	Distribution	P-value
p_E	VA	-	(1:0.3, 2:0.5, 3:0.2) (Marathe et al. 2011)	-
	NJ	-	(1:0.3, 2:0.5, 3:0.2) (Marathe et al. 2011)	-
p_I	VA	-	(3:0.3, 4:0.4, 5:0.2, 6:0.1) (Marathe et al. 2011)	-
	NJ	-	(3:0.3, 4:0.4, 5:0.2, 6:0.1) (Marathe et al. 2011)	-
τ	VA	Normal	$\mathcal{N}(\mu = 4.88\text{e}{-5}, \delta = 9.33\text{e}{-7})$	0.74
	NJ	Normal	$\mathcal{N}(\mu = 4.63\text{e}{-5}, \delta = 1.05\text{e}{-6})$	0.85
N_I	VA	Uniform	$\mathcal{U}(7224, 14798)$	0.99
	NJ	Exponential	$exp(\mu = 708, \beta = 2013)$	0.93
I_V	VA	Discrete Uniform	6 vaccination schedules (CDC 2018)	0.99
	NJ	Discrete Uniform	6 vaccination schedules (CDC 2018)	0.93

resolution information; (ii) it includes realistic scenarios which has not happened in the past. The model trained on the simulated training dataset allows a tolerance of flu season variance.

Settings of comparison methods

We elaborate the details of settings for each method in the following.

(1) *DEFSI*: For each state, we generated $r = 5000$ simulated curves with the length $\ell = 52$ of weekly ILI incidence on both state level and county level. The simulations are initializing by samples from parameter space $\mathcal{P}$ (shown in Table 1). In DEFSI model, we set look back window size of within-season observations as 10, and between-season observations as 5. (2) *Single layer LSTM model (LSTM)*: This is a baseline from artificial neural network methods. No surrogate indicator used in this model. We set the look back window size as 10. (3) *AdapLSTM* (Venna et al. 2017): This method makes predictions using a simple LSTM model, then adjusts the predictions by applying impacts of weather factors and spatio-temporal factors. LSTM model has the same setting with (2) in our experiment. In (Venna et al. 2017), the weather data features include maximum temperature, minimum temperature, humidity, and precipitation, while humidity is not included in our experiment. The confidences of symbolic pairs in our experiment are less than 0.3, which will lead to arbitrary adjustment for predictions. The neighbors of each state used for spatio-temporal adjustment factor are the same with neighbors described in Fitting distributions for disease parameters. (4) *Simple ARIMA model (ARIMA)*: This is a baseline from statistical methods. No exogenous variable used in this model. The order for ARIMA is $(10, 1, 0)$. (5) *AutoRegression with GOogle search data (ARGO)* (Yang, Santillana, and Kou 2015): The method proposes an autoregression model utilizing Google search data. We use the public available tool from (Yang, Santillana, and Kou 2015). In our experiment, We set look back window size as 52 and training window as 104. All of the top 100 Google correlate terms of VA are flu related, while only less than 1% of the top 100 Google correlated terms of NJ are flu related. (6) *EpiFast* (Beckman et al. 2014): This model follows the definition in SEIR-based Epidemic Simulation, the parameters tuned in EpiFast are p_E, p_I, N_I, τ. They are optimized by minimizing the error of the predicted and the actual ILI incidence via Nelder-Mead method (Nelder and Mead 1965).

Related work

Large amounts of research work have made outstanding contributions to the field of infectious disease forecasting. In this section, we review related work that is most relevant to ours. Interested readers can refer to (Alessa and Faezipour 2018; Chretien et al. 2014; Nsoesie et al. 2014) for more details. We discuss the existing ILI forecasting methods categorized based on their underlying methodologies: causal methods, statistical methods, and artificial neural network methods.

Causal Methods In epidemiology, infectious disease models for within-host progression include: susceptible-infectious-recovered (SIR), susceptible-exposed-infectious-recovered (SEIR), susceptible-infectious-recovered-susceptible (SIRS), and their extensions (Bailey 1975; Kuznetsov and Piccardi 1994). Forecasting methods employing these models are called causal methods because they describe the causal mechanisms of infectious diseases. Flu forecasting by causal methods works by identifying a good estimate of the underlying disease model that generates the incidence data that best fit the observations. In these methods the underlying epidemic model can be either a compartmental model (CM) (Flahault et al. 2006; Lee et al. 2012; Lunelli, Pugliese, and Rizzo 2009) or an agent-based model (ABM) (Parker and Epstein 2011; Chao et al. 2010). In a compartmental model, a population is divided into compartments (e.g. S, E, I, R) and differential equations characterize change of sizes of the compartments due to disease propagation and progression. In an agent-based model, disease spreads among heterogeneous agents through an unstructured network. The individual level details in an agent-based model can be easily aggregated to obtain epidemic data of any resolution, e.g. number of newly infected people in a county in a specific week. To get county level epidemics in a compartmental model, however, one needs to create compartments in each county, where county population size and between county travel data become crucial. Researchers of many scientific community developed innovative methods to predict influenza activity. Shaman et al. (Shaman and Karspeck 2012) developed a framework for initializing real-time forecasts of seasonal influenza outbreaks, using a data assimilation technique commonly applied in numerical weather prediction. Tuite et al. (Tuite et al. 2010) used an SIR CM to estimate parameters and morbidity in pandemic H1N1. Yang et al. (Yang, Karspeck, and Shaman 2014) applied various filter methods to model and forecast influenza activity using an SIRS CM. In (Nsoesie et al. 2013), authors proposed an simulation optimization approach based on SEIR ABM for epidemic forecasting. Zhao et al. (Zhao et al. 2015) infer the parameters of the SEIR ABM from social media data for ILI forecasting. *Limitations: There are plenty of parameters within a causal model and each parameter often has extremely high dimensionality. Thus, the optimal estimate is quite difficult in terms of searching and computing. Especially, methods employing agent-based models are computationally intensive in real-time forecasting.*

Statistical methods Statistical methods employ statistical and time-series based methodologies to learn patterns in historical epidemic outbreaks and leverage those patterns for forecasting. Popular statistical methods for ILI forecasting include e.g. generalized linear models (GLM), autoregressive integrated moving average (ARIMA), and generalized autoregressive moving average (GARMA) (Bardak and Tan 2015; Benjamin, Rigby, and Stasinopoulos 2003; Dugas et al. 2013). Wang et al. (Wang et al. 2015) proposed a dynamic Poisson autoregressive model with exogenous input variables (DPARX) for flu forecasting. Yang et al. (Yang, Santillana, and Kou 2015) proposed ARGO, an autoregressive-based influenza tracking model for nowcasting that incorporated CDC ILI data and Google search data. The extensive work based on ARGO are discussed in (Yang et al. 2017). *Limitations: Although existing work has made significant achievements in short-term forecasting, none of them could make high-resolution forecasting due to the lack of high-resolution observations for most of regions around the world. In addition, in statistical methods, we often assume a linear relationship between the inputs and outputs, which may not true for real cases.*

Artificial neural network methods Artificial neural networks (ANN) have gained increased prominence recently in epidemic forecasting due to their self-learning ability without prior knowledge. Xu et al. (Xu et al. 2017) first introduced feed-forward neural networks (FNN) into surveillance of infectious diseases and investigated its predictive utility using CDC ILI data, Google search data, and meteorological data. Recurrent Neural Network (RNN) has been demonstrated to be able to capture dynamic temporal behavior of a time sequence. In (Volkova et al. 2017) Volkova et al. built an LSTM model for short-term ILI forecasting using CDC ILI and twitter data. Venna et al. (Venna et al. 2017) proposed LSTM based method that integrates the impacts of climatic factors and geographical proximity to achieve better forecasting performance. Wu et al. (Wu et al. 2018) constructed a deep learning structure combining RNN and convolutional neural network to fuse information from different sources. *Limitations: Similar to statistical methods, they could not make high-resolution forecasting without corresponding observations. Moreover, overfitting often happens due to the small training data points available in epidemic history.*

References

Alessa, A., and Faezipour, M. 2018. A review of influenza detection and prediction through social networking sites. *Theoretical Biology & Medical Modelling* 15:2.

Bailey, N. T. J. 1975. *The Mathematical Theory of Infectious Diseases and Its Applications*. Griffin, 2 edition.

Bardak, B., and Tan, M. 2015. Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data. In *IEEE 15th BIBE*, 1–6.

Beckman, R. J.; Bisset, K. R.; Chen, J.; Lewis, B. L.; Marathe, M. V.; and Stretz, P. E. 2014. ISIS: a networked-epidemiology based pervasive web app for infectious disease pandemic planning and response. In *KDD*.

Benjamin, M. A.; Rigby, R. A.; and Stasinopoulos, D. M. 2003. Generalized autoregressive moving average models. *Journal of the American Statistical Association* 98(461):214–223.

CDC. 2018. Historical seasonal influenza vaccine schedule. `https://www.cdc.gov/flu/professionals/vaccination/vaccinesupply.htm`. Accessed November 1, 2017.

Chao, D. L.; Halloran, M. E.; Obenchain, V. J.; and Longini, Jr, I. M. 2010. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLOS Computational Biology* 6(1):1–8.

Chretien, J. P.; George, D.; Shaman, J.; Chitale, R. A.; and McKenzie, F. E. 2014. Influenza forecasting in human populations: A scoping review. *PLOS ONE* 9(4):1–8.

Dugas, A. F.; Jalalpour, M.; Gel, Y.; Levin, S.; Torcaso, F.; Igusa, T.; and Rothman, R. E. 2013. Influenza Forecasting with Google Flu Trends. *PLoS ONE* 8(2):e56176.

Flahault, A.; Vergu, E.; Coudeville, L.; and Grais, R. F. 2006. Strategies for containing a global influenza pandemic. *Vaccine* 24(44):6751–6755.

Kuznetsov, Y. A., and Piccardi, C. 1994. Bifurcation analysis of periodic SEIR and SIR epidemic models. *Journal of Mathematical Biology* 32(2):109–121.

Lee, J. M.; Choi, D.; Cho, G.; and Kim, Y. 2012. The effect of public health interventions on the spread of influenza among cities. *Journal of Theoretical Biology* 293:131–142.

Lunelli, A.; Pugliese, A.; and Rizzo, C. 2009. Epidemic patch models applied to pandemic influenza: Contact matrix, stochasticity, robustness of predictions. *Mathematical Biosciences* 220(1):24–33.

Marathe, A.; Lewis, B.; Chen, J.; and Eubank, S. 2011. Sensitivity of household transmission to household contact structure and size. *PLoS ONE* 6.

Nelder, J. A., and Mead, R. 1965. A simplex method for function minimization. *The Computer Journal* 7(4):308–313.

Nsoesie, E. O.; Beckman, R. J.; Shashaani, S.; Nagaraj, K. S.; and Marathe, M. V. 2013. A simulation optimization approach to epidemic forecasting. *PLOS ONE* 8(6):1–10.

Nsoesie, E. O.; Brownstein, J. S.; Ramakrishnan, N.; and Marathe, M. V. 2014. A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses* 8(3):309–316.

Parker, J., and Epstein, J. M. 2011. A Distributed Platform for Global-Scale Agent-Based Models of Disease Transmission. *ACM Trans Model Comput Simul* 22(1):2.

Shaman, J., and Karspeck, A. 2012. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*.

Tuite, A. R.; Greer, A. L.; Whelan, M.; Winter, A.-L.; Lee, B.; Yan, P.; Wu, J.; Moghadas, S.; Buckeridge, D.; Pourbohloul, B.; and Fisman, D. N. 2010. Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. *CMAJ* 182(2):131–136.

Venna, S. R.; Tavanaei, A.; Gottumukkala, R. N.; Raghavan, V. V.; Maida, A.; and Nichols, S. 2017. A novel data-driven model for real-time influenza forecasting. *bioRxiv*.

Volkova, S.; Ayton, E.; Porterfield, K.; and Corley, C. D. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLOS ONE* 12(12):1–22.

Wang, Z.; Chakraborty, P.; Mekaru, S.; Brownstein, J.; Ye, J.; and Ramakrishnan, N. 2015. Dynamic poisson autoregression for influenza-like-illness case count prediction. In *KDD*.

Wu, Y.; Yang, Y.; Nishiura, H.; and Saitoh, M. 2018. Deep learning for epidemiological predictions. In *SIGIR*.

Xu, Q.; Gel, Y. R.; Ramirez Ramirez, L. L.; Nezafati, K.; Zhang, Q.; and Tsui, K. L. 2017. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PLOS ONE* 12(5):1–17.

Yang, S.; Santillana, M.; Brownstein, J. S.; Gray, J.; Richardson, S.; and Kou, S. C. 2017. Using electronic health records and internet search information for accurate influenza forecasting. *BMC Infectious Diseases* 17(1):332.

Yang, W.; Karspeck, A.; and Shaman, J. 2014. Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLOS Computational Biology* 10(4):1–15.

Yang, S.; Santillana, M.; and Kou, S. C. 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. *PNAS* 112(47):14473–14478.

Zhao, L.; Chen, J.; Chen, F.; Wang, W.; Lu, C. T.; and Ramakrishnan, N. 2015. SimNest: Social Media Nested Epidemic Simulation via Online Semi-supervised Deep Learning. *Proceedings of IEEE ICDM* 2015:639–648.