A Unified Framework of Epidemic Spreading Prediction by Empirical Mode Decomposition-Based Ensemble Learning Techniques

Yun Feng and Bing-Chuan Wang

***Abstract*— In this paper, a unified susceptible-exposed-infected-susceptible-aware (SEIS-A) framework is proposed to combine the epidemic spreading process with individuals' self-query behaviors on the Internet. An epidemic spreading prediction model that contains two phases is established based on the SEIS-A framework. To deal with the nonstationary complex characteristic of the time series data of disease density, it is decomposed through the empirical mode decomposition (EMD) method to obtain the intrinsic mode functions (IMFs) in phase I. To enhance the prediction performance, the ensemble learning techniques that use the self-query data as an external input are applied to these IMFs in phase II. Finally, an empirical study on the prediction of weekly consultation rates of hand-foot-and-mouth disease (HFMD) in Hong Kong is conducted to validate the effectiveness of the proposed method. The main advantage of this method is that it outperforms other learning methods on fluctuating complex epidemic spreading data.**

***Index Terms*— Empirical mode decomposition (EMD), ensemble learning, epidemic spreading prediction.**

I. Introduction

SINCE the pioneering work of Bernoulli [1], mathematical modeling of epidemic spreading has received continuous attention for more than 200 years. Classical epidemic models, such as the susceptible-infected-susceptible (SIS) [2]–[11] model and the susceptible-infected-recovered (SIR) [12]–[17] model, have been well studied. Based on these models, more realistic models have been proposed to better illustrate real disease spreading behavior. For instance, the susceptible-infected-recovered-susceptible (SIRS) [18] model describes that the immunity of a recovered individual may be lost, causing him or her susceptible again, while the susceptible-exposed-infected-recovered (SEIR) [19] model contains exposed individuals who have been infected by the disease but cannot transmit it to others yet.

The interaction between epidemic spreading and the awareness of individuals transmission has been considered in some studies [20]–[22]. For example, in [21], the analysis of the interrelation between two processes accounting for the spreading of an epidemic, and the information awareness to prevent its infection, on top of multiplex networks was presented. All of these studies were focused on the impact of this interaction on the disease spreading behavior. However, the awareness of individuals may also lead to self-query behaviors on the Internet. For instance, if an individual has some pre-symptoms such as a runny nose and cough, when he tries to find out whether he has been infected by influenza, the fastest way is to search these symptoms on search engines such as Google on the internet. These self-query behaviors can provide more insights into the upcoming outbreak of a certain disease. In addition, these behaviors often occur days or even weeks before this individual is completely infected and seeks medical help in hospitals or clinics. Therefore, the search data on the Internet can be utilized to predict the spreading scale of a certain disease weeks before its outbreak.

Recently, some researchers have investigated the disease spreading prediction problem with search data on the Internet [23]–[27]. In [27], an influenza tracking model, AutoRegression with Google search data (ARGO) that uses publicly available search data on the Internet, is proposed. Despite these innovative methods proposed in these studies, a unified framework to combine disease spreading with individuals' self-query behaviors is missing. Moreover, all these studies utilize influenza data with significant seasonal characteristic and little fluctuation. For some other disease data that contain more randomness and fluctuations, the performance of these methods may not be satisfied. Furthermore, for the complex disease spreading data, there exist many features and one learning method can only learn part of them. The major difficulty is how to design learning methods to capture these features effectively to accomplish the prediction task.

As a nonlinear and nonstationary time-domain decomposition method, empirical mode decomposition (EMD) [28]–[30] decomposes a time series into multiple empirical modes, known as the intrinsic mode functions (IMFs). One intrinsic mode can comprise fluctuations with a variety of wavelengths at different time steps for signals with intermittent fluctuations. Hence, EMD has natural advantages in decomposing nonstationary complex data. Motivated by the advantages of EMD in decomposing nonstationary time series, it is intuitively used to decompose these original disease spreading time series into multiple IMFs with different frequencies. In this manner, these features can be scattered to different IMFs. The problem that comes with it is how to design learning methods to

Manuscript received May 10, 2018; revised January 28, 2019; accepted April 30, 2019. *(Corresponding author: Bing-Chuan Wang.)*

The authors are with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong (e-mail: yun.feng@my.cityu.edu.hk; bingcwang3-c@my.cityu.edu.hk).

Digital Object Identifier 10.1109/TCSS.2019.2915615

capture these scattered features. Since the ensemble learning techniques have been designed to combine multiple learning methods together to improve the learning performance in the community of machine learning, they are used to capture the features in different IMFs as a matter of course.

The proposed EMD-based method is inherently superior over traditional ensemble learning methods due to the fact that every single method in traditional ensemble learning is trained with the same original time series. In this manner, these features are still mixed up and difficult to learn.

Motivated by the above-mentioned considerations, a unified framework named susceptible-exposed-infected-susceptible-aware (SEIS-A) is proposed. This framework combines the traditional SEIS [31]–[34] epidemic model with the self-query behaviors of individuals on the Internet. To better accommodate the fluctuating complex data, EMD is employed to decompose the complex time series data into the IMFs. For each IMF, one learning method is utilized. The ensemble learning techniques that use the self-query data as an external input are applied to these IMFs.

The main contributions of this paper can be summarized as follows.

1) A novel unified SEIS-A framework that combines epidemic spreading and individuals' self-query behaviors on the Internet is proposed.
2) A prediction model that combines the advantages of EMD and ensemble learning techniques is presented based on the proposed SEIS-A framework.
3) The proposed method outperforms other learning methods on the empirical study of the prediction of hand-foot-and-mouth disease (HFMD) spreading in Hong Kong.

The rest of this paper is organized as follows. The SEIS-A framework and problem description are given in Section II. EMD and ensemble learning techniques are described in Section III. In Section IV, the prediction results on HFMD in Hong Kong are presented. Finally, we conclude this paper in Section V.

II. SEIS-A Framework and Problem Description

Considering the fact that individuals in the exposed (E) state may search for suspicious symptoms and other keywords related to a certain disease through search engines on the Internet, an SEIS-A framework is proposed in Fig. 1. In this framework, S denotes the susceptible individuals that are vulnerable to the considered disease, E denotes the exposed individuals that are in the latency period and have presymptoms but cannot transmit the disease to others, and I denotes the infected individuals; Another state named Aware (A) is introduced to denote individuals that are aware of the disease. Individuals in the A state may search for some keywords related to the disease on the Internet, such as the symptoms, treatment approaches, and so on. These searching behaviors reveal information on potentially infected individuals. Therefore, the self-query data related to a certain disease can be utilized to predict the epidemic spreading scale.

Based on the assumption of homogeneous mixing approximation [2], the mathematical model of the SEIS-A framework can be described by

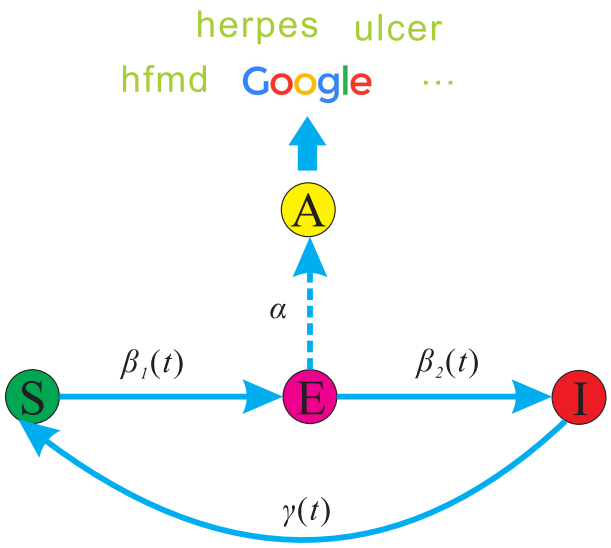

Fig. 1. SEIS-A framework.

$$\begin{cases} \dfrac{ds(t)}{dt} = -\beta_1(t)s(t)i(t) + \gamma(t)i(t) \\ \dfrac{de(t)}{dt} = -\beta_2(t)e(t) + \beta_1(t)s(t)i(t) \\ \dfrac{di(t)}{dt} = -\gamma(t)i(t) + \beta_2(t)e(t) \end{cases} \tag{1}$$

and

$$\begin{cases} a(t) = \alpha e(t) \\ \boldsymbol{Q}(t+1) = \boldsymbol{f}(a(t)) \end{cases} \tag{2}$$

where $s(t)$, $e(t)$, $i(t)$, and $a(t)$ denote the density of individuals in the susceptible state, exposed state, infected state, and aware state, respectively. To be noticed, (2) is based on the assumption that a certain number of individuals in the exposed state become aware of the disease and search for some keywords related to the disease on the Internet. The parameters $\beta_1(t)$ and $\beta_2(t)$ denote the time-varying transmission rate of individuals from the susceptible state to the exposed state and the exposed state to the infected state, respectively. $\gamma(t)$ denotes the time-varying curing rate of infected individuals. In addition, α is the proportion of exposed individuals that is aware of the disease. $\boldsymbol{Q}(t)$ denotes the self-query data that are related to the disease; $\boldsymbol{f}(\cdot)$ denotes the search functions for aware individuals. Equation (1) can be rewritten in the following discrete form:

$$\begin{cases} s(t+1) = s(t) + (-\beta_1(t)s(t)i(t) + \gamma(t)i(t))\Delta t \\ e(t+1) = e(t) + (-\beta_2(t)e(t) + \beta_1(t)s(t)i(t))\Delta t \\ i(t+1) = i(t) + (-\gamma(t)i(t) + \beta_2(t)e(t))\Delta t \end{cases} \tag{3}$$

where Δt denotes the sampling interval for discretization.

Combining (2) with (3), it is easy to obtain the following equation:

$$i(t+1) = i(t) + (-\gamma(t)i(t) + \frac{\beta_2(t)}{\alpha}\boldsymbol{f}^{-1}(\boldsymbol{Q}(t+1)))\Delta t \tag{4}$$

Algorithm 1 EMD

1: For any given data $\{x(t)\}$, identify all the local maxima and minima;
2: Connect all the local maxima and minima by natural cubic spline lines separately to form the upper and lower envelops, denoted by $\{u(t)\}$ and $\{l(t)\}$, respectively;
3: Calculate the mean of the envelops as $m(t) = [u(t) + l(t)]/2$;
4: Define the difference between the data and the mean as the proto-IMF, $h(t) = x(t) - m(t)$;
5: Check the proto-IMF according to the IMF definition and the terminal criterion to determine whether $\{h(t)\}$ is an IMF.
6: If $\{h(t)\}$ is not an IMF, then replace $\{x(t)\}$ with $\{h(t)\}$ and repeat step 1-5 until it satisfies the IMF definition.
7: If $\{h(t)\}$ is an IMF, then assign the proto-IMF $\{h(t)\}$ as an IMF component $\{c(t)\}$.
8: Repeat step 1-7 by replacing $\{x(t)\}$ with the residue $r(t) = x(t) - c(t)$.
9: End the procedure when the residue contains no more than one extremum.

where $\boldsymbol{f}^{-1}(\cdot)$ is the inverse function of the search function $\boldsymbol{f}(\cdot)$. Equation (4) indicates that the disease density at step $t+1$ can be inferred by the disease density at step t and the self-query data collected at step $t+1$. In addition, the time-varying characteristic of parameters $\beta_2(t)$ and $\gamma(t)$ that are due to the randomness characteristic of the disease makes the prediction of disease density more difficult.

Therefore, the disease spreading prediction problem is defined as follows: how to predict the disease density based on the previously collected disease density data and the currently collected self-query data?

To compensate for the unknown time-varying parameters $\beta_2(t)$ and $\gamma(t)$ in (4), it is assumed that these parameters can be inferred by the previous disease density as follows:

$$\begin{cases} \gamma(t) = g_1(i(t), \quad i(t-1), \ldots, i(t-l)) \\ \beta_2(t) = g_2(i(t), \quad i(t-1), \ldots, i(t-l)). \end{cases}$$

Hence, the following disease density prediction model is obtained:

$$i(t+1) = i(t) + (-g_1(\boldsymbol{I}(t)) \cdot i(t) + \frac{g_2(\boldsymbol{I}(t))}{\alpha} \cdot \boldsymbol{f}^{-1}(Q(t+1)))\Delta t \tag{5}$$

where $\boldsymbol{I}(t) = [i(t), i(t-1), \ldots, i(t-l)]^T$ and l denotes the autoregressive (AR) order of the time series.

III. Methodology

In this section, a methodology framework based on EMD and ensemble learning techniques is proposed.

A. Methodology Framework

In the epidemic spreading prediction problem in (5), the disease density $i(t+1)$ at step $t+1$ is associated with the previous disease density $i(t)$, the time-varying model parameters $\beta_2(t)$ and $\gamma(t)$, and the self-query data $\boldsymbol{Q}(t+1)$. The self-query data $\boldsymbol{Q}(t+1)$ refer to the search activities of keywords related to the considered disease from search engines on the Internet such as Google. Due to the timely access of these data, the self-query data at step $t+1$ are used as the external input of the learning model. Since the time-varying parameters $\beta_2(t)$ and $\gamma(t)$ are unknown, in order to establish the epidemic spreading prediction model, the previous disease density data $\boldsymbol{I}(t) = [i(t), i(t-1), \ldots, i(t-l)]^T$ are used to compensate for these unknown parameters.

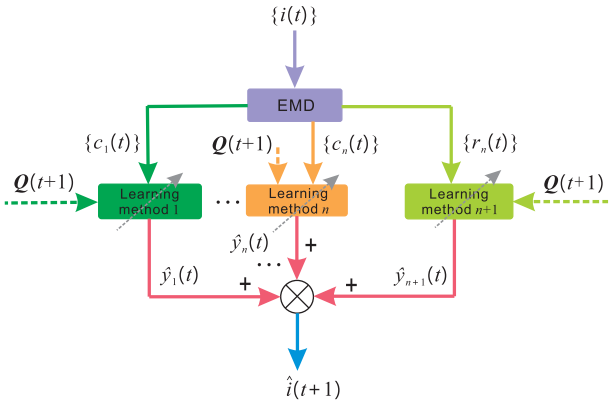

Fig. 2. Methodology framework.

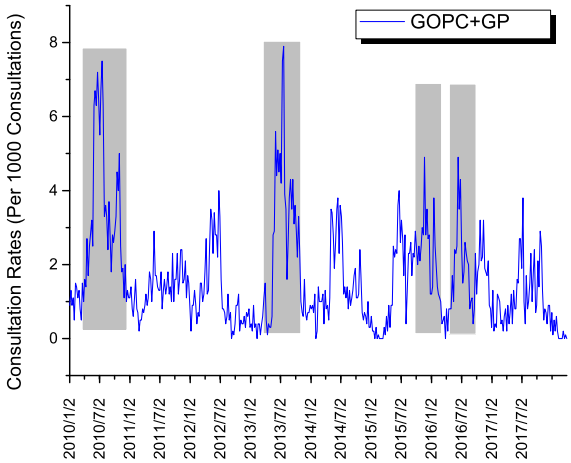

Fig. 3. GOPC + GP from week 1, 2010 to week 13, 2018.

As it is presented in Fig. 2, the proposed methodology contains two phases. In phase I, the collected time series data of disease density $\{i(t)\}$ are decomposed by EMD to obtain the IMFs. In phase II, these IMFs combined with the self-query data $\boldsymbol{Q}(t+1)$ are utilized as the input of various learning methods for ensemble learning techniques to predict the disease density at step $t+1$.

B. Empirical Mode Decomposition

EMD is a data-driven method that decomposes a given time series into the IMFs. The intuitive idea of EMD is to decompose data through a shifting process. For any given data $\{x(t)\}$, it can be decomposed into the IMFs $\{c_j(t)\}$ and the

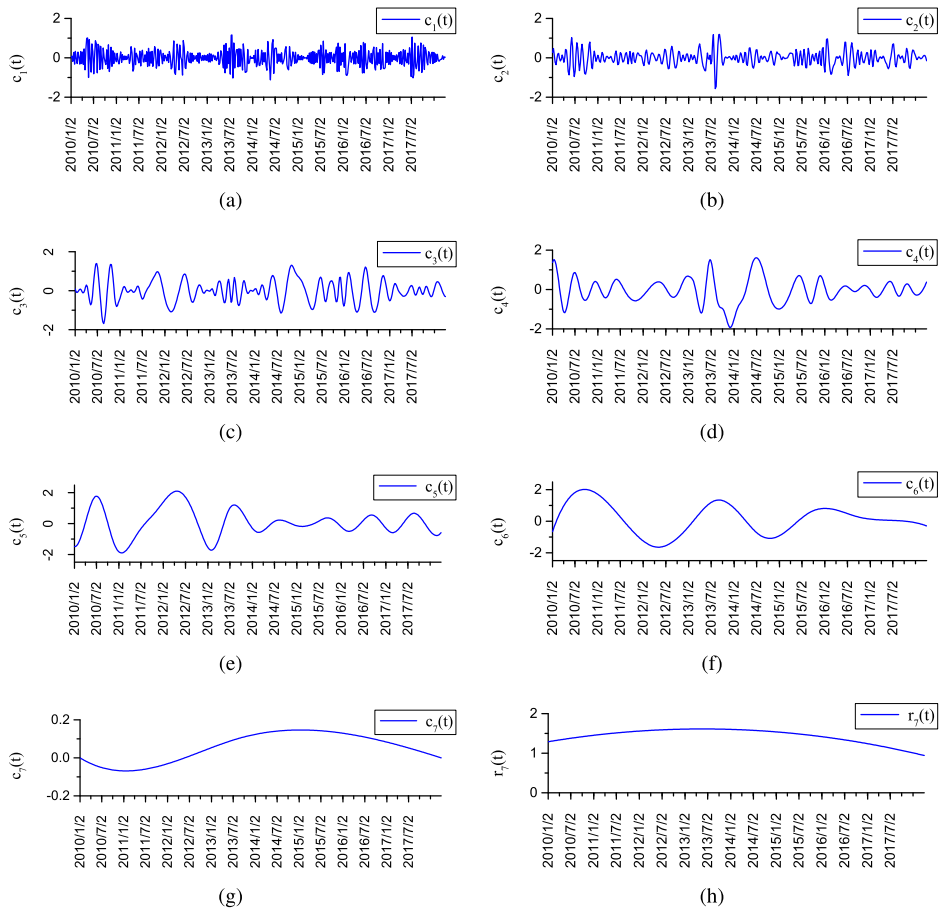

Fig. 4. IMFs and residue obtained through EMD. (a) $c_1(t)$. (b) $c_2(t)$. (c) $c_3(t)$. (d) $c_4(t)$. (e) $c_5(t)$. (f) $c_6(t)$. (g) $c_7(t)$. (h) $r_7(t)$.

residual $\{r_n(t)\}$ as

$$x(t) = \sum_{j=1}^{n} c_j(t) + r_n(t). \tag{6}$$

The detailed steps of EMD are summarized in Algorithm 1.

Once the previous disease density data $\boldsymbol{I}(t) = [i(t), i(t-1), \ldots, i(t-l)]^T$ is collected, first, EMD is utilized to obtain the IMFs and the residue

$$i(t) = \sum_{j=1}^{n} c_j(t) + r_n(t). \tag{7}$$

Then, the IMFs and residue are further processed by the ensemble learning techniques.

C. LASSO-Based Ensemble Learning

LASSO is short for least absolute shrinkage and selection operator [35]. The intuitive idea of LASSO is to minimize the residual sum of squares, while the sum of the absolute value of the coefficients is less than a constant. The details of LASSO can be referred to the Appendix.

As shown in Fig. 2, in phase I, the disease density time series data $\{i(t)\}$ are processed by EMD to obtain the IMFs $\{c_j(t)\}$ and the residue $\{r_n(t)\}$. After decomposition, LASSO is employed on these IMFs and the residue for ensemble learning. Since there are n IMFs and 1 residue, $n+1$ learning methods are utilized for each component separately, and then, ensemble learning techniques that use the self-query data as an external input are applied to obtain the prediction of disease density $\hat{i}(t+1)$ at the next step.

In the training procedure, the inputs for learning method j between 1 and n are the IMF component $\{c_j(t)\}$ and the self-query data $\boldsymbol{Q}(t+1)$. The inputs for learning method $n+1$ are the IMF residue $\{r_n(t)\}$ and the self-query data $\boldsymbol{Q}(t+1)$. The prediction output for each learning method j is denoted by $\hat{y}_j(t+1)$ here. For simplicity, LASSO is assigned for all $n+1$ learning methods. Therefore, for LASSO j where $j < n+1$, the input matrix $\boldsymbol{x}_j$ can be written as $\boldsymbol{x}_j = [\boldsymbol{x}_j(t-m+1), \ldots, \boldsymbol{x}_j(t-1), \boldsymbol{x}_j(t)]^T$, where $\boldsymbol{x}_j(t) = [c_j(t-l), \ldots, c_j(t-1), c_j(t), \boldsymbol{Q}^T(t+1)]^T$, while the output vector is $\hat{\boldsymbol{y}}_j = [y_j(t-m+2), \ldots, y_j(t), y_j(t+1)]^T$. For LASSO $n+1$, we have $\boldsymbol{x}_{n+1}(t) = [r_n(t-l), \ldots, r_n(t-1), r_n(t), \boldsymbol{Q}^T(t+1)]^T$. m denotes the window length for training and l denotes the AR order of the time series. The sliding-window-based-on-line learning scheme [27] is used to accommodate the time-varying parameters in the proposed SEIS-A framework.

For each LASSO training problem, the objective function for parameters $(\boldsymbol{\theta}_j)$ estimation is

$$\hat{\boldsymbol{\theta}}_j = \arg\min \frac{1}{2}\|\boldsymbol{y}_j - \boldsymbol{\phi}^T(\boldsymbol{x}_j)\boldsymbol{\theta}_j\|_2^2 + \lambda\|\boldsymbol{\theta}_j\|_1, \quad j = 1, \ldots, n+1 \tag{8}$$

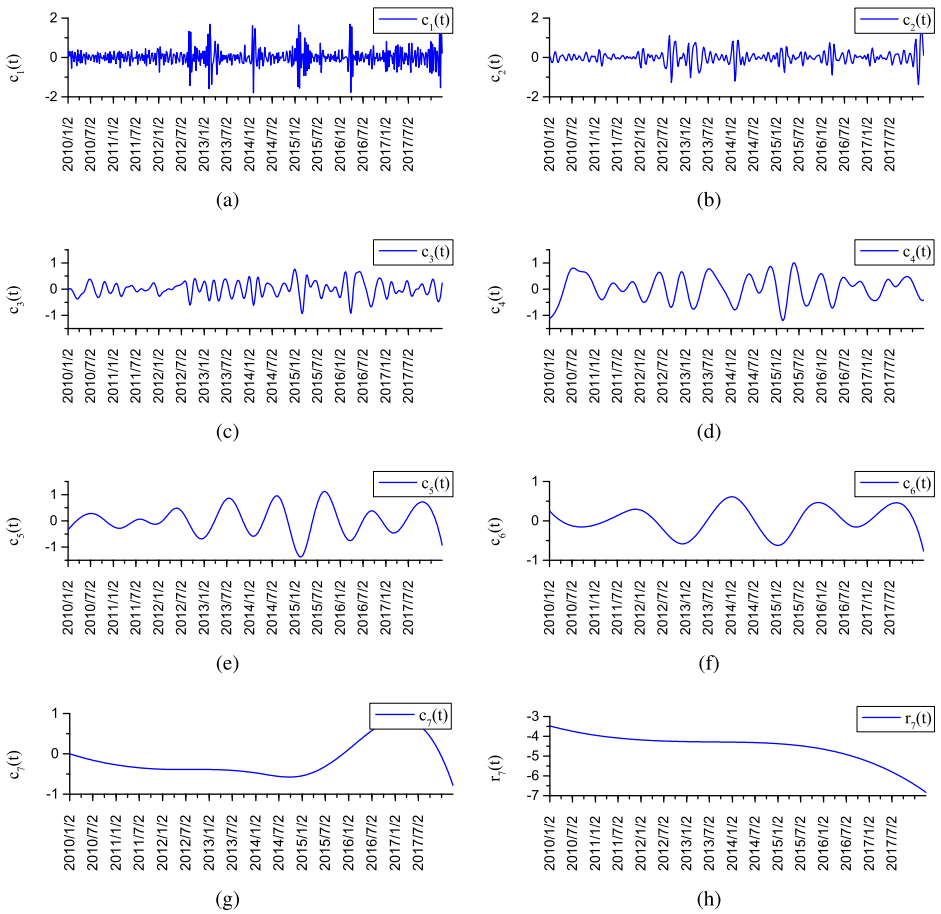

Fig. 5. IMFs and residue obtained through EEMD. (a) $c_1(t)$. (b) $c_2(t)$. (c) $c_3(t)$. (d) $c_4(t)$. (e) $c_5(t)$. (f) $c_6(t)$. (g) $c_7(t)$. (h) $r_7(t)$.

where $\boldsymbol{\phi}$ are the basis functions and λ is the regularization coefficient. $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the L2-norm and L1-norm, respectively.

After obtaining the parameters' estimation $\boldsymbol{\theta}_j$, the prediction for each LASSO model can be obtained as

$$\hat{\boldsymbol{y}}_j = \boldsymbol{\phi}^T(\boldsymbol{x}_j)\hat{\boldsymbol{\theta}}_j \tag{9}$$

where $\hat{\boldsymbol{y}}_j = [\hat{y}_j(t-m+2), \ldots, \hat{y}_j(t), \hat{y}_j(t+1)]^T$. Hence, the disease density prediction $\hat{i}(t+1)$ can be obtained through the following ensemble learning techniques:

$$\hat{i}(t+1) = \sum_{j=1}^{n+1} \rho_j \hat{y}_j(t+1) \tag{10}$$

where ρ_j denotes the aggregate coefficient.

IV. Empirical Study

In this section, a real disease spreading case is considered. HFMD is a common seasonal infectious disease among children. Large outbreaks of HFMD have been occurring in Asia since 1997. HFMD is popular in spring, summer, and fall, and it can even cause death for severe victims. Therefore, the prediction of HFMD spreading is of great significance. In this section, the HFMD spreading data and the Google search data in Hong Kong are utilized to validate the effectiveness of the proposed method. The descriptions of these data are presented in the following.

A. HFMD Data Description

The weekly consultation rates (per 1000 consultations) of HFMD by General Out-Patient Clinics (GOPC) and General Practitioners (GP) from week 1, 2010 to week 13, 2018 are considered. The data are collected from the Hong Kong Centers for Health Protection (CHP) [36]. The weekly consultation rates represent the disease spreading scale in the current week.

As it is presented in Fig. 3, the weekly consultation rate reveals seasonal fluctuations, which indicates the seasonal characteristic of HFMD. From week 1, 2010 to week 13, 2018, there are four major HFMD spreading seasons, and all of them occur in summer months. Apart from the seasonal characteristic, there exist numerous fluctuations, which makes the prediction problem more difficult.

B. Mapping the SEIS-A Framework to Observations

To be noticed, the SEIS-A framework represents the disease incidence on a per capita basis, or incidence rate, and includes asymptomatic and mildly symptomatic infections. Due to the unavailability of the actual disease density data $\boldsymbol{I}(t)$, only the weekly consultation rate can be used for prediction. To address this discordance, a scaling factor μ is used to map the SEIS-A framework to the weekly consultation rate observation [37]. Denote δ as the rate of consultation individuals who are infected and $\eta(t)$ as the weekly consultation rate in week t. By Bayes' rule, the probability of an individual infected by

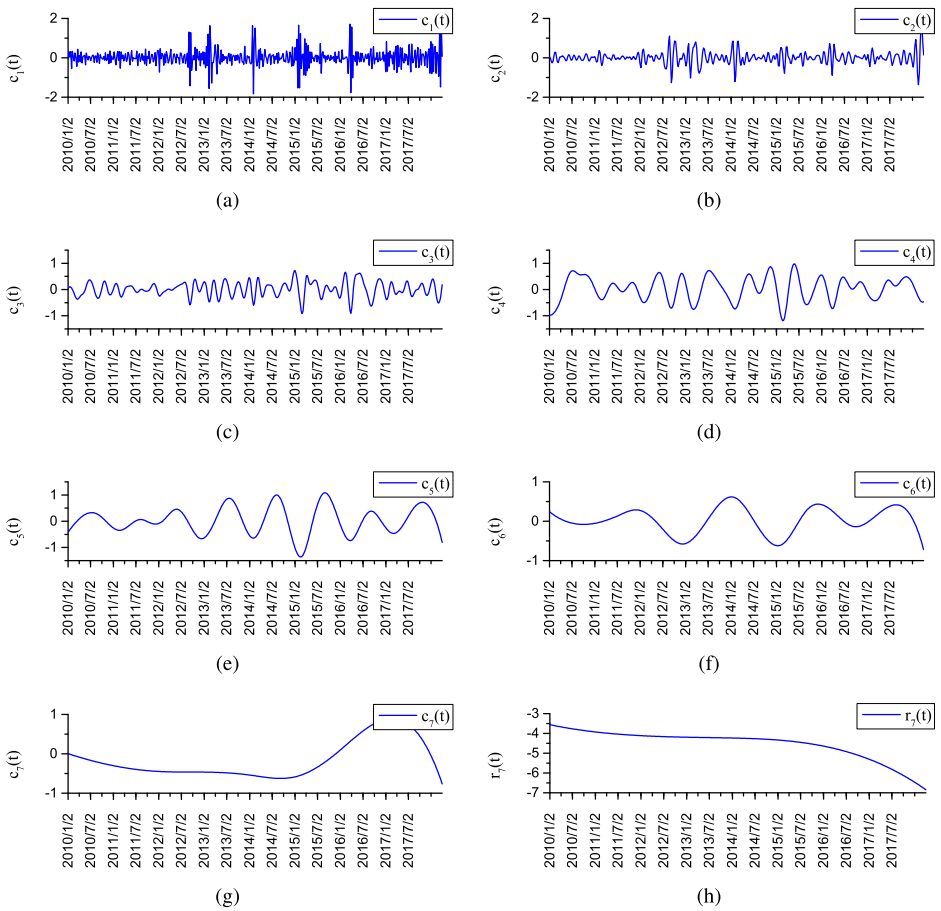

Fig. 6. IMFs and residue obtained through CEEMD. (a) $c_1(t)$. (b) $c_2(t)$. (c) $c_3(t)$. (d) $c_4(t)$. (e) $c_5(t)$. (f) $c_6(t)$. (g) $c_7(t)$. (h) $r_7(t)$.

the disease during a given week, $p(i)$, is

$$p(i) = \frac{p(m)}{p(m|i)} \times p(i|m)$$

where

$$p(i|m) \approx \delta \times \eta(t)$$

$p(i|m)$ denotes the probability that an individual seeking medical attention m is infected. Denoting $\mu = p(m)/p(m|i)$ as the scaling factor, one has

$$i(t) = p(i) \approx \mu\delta\eta(t). \tag{11}$$

On the one hand, as shown in (11), the disease density $i(t)$ is associated with the observed weekly consultation rate, that is, the increase in the weekly consultation rate can cause the outbreak of the disease. On the other hand, the actual disease density data $\boldsymbol{I}(t)$ are unavailable. Therefore, the weekly consultation rate is employed as the objective of prediction. Then, the epidemic spreading prediction problem in (5) can be converted into the weekly consultation rate prediction problem as follows:

$$\eta(t+1) = \eta(t) + (-G_1(\boldsymbol{\eta}(t)) \cdot \eta(t) + G_2(\boldsymbol{\eta}(t)) \cdot \boldsymbol{f}^{-1}(Q(t+1)))\Delta t$$

where $\boldsymbol{\eta}(t) = [\eta(t), \eta(t-1), \ldots, \eta(t-l)]^T$, $G_1(\boldsymbol{\eta}(t)) = g_1(\mu\delta\boldsymbol{\eta}(t))$, and $G_2(\boldsymbol{\eta}(t)) = g_2(\mu\delta\boldsymbol{\eta}(t))/\alpha\mu\delta$.

According to the proposed methodology framework, first, the weekly consultation data are processed through EMD to obtain the IMFs and residue. The obtained IMFs and residue are shown in Fig. 4, and the frequency of the data is decreasing from $c_1(t)$ to $c_7(t)$. For each IMF and residue, an independent LASSO model is assigned.

Moreover, two extensions of EMD, i.e., the ensemble EMD (EEMD) and the complementary EEMD (CEEMD), are also employed as the decomposition tools in the proposed methodology framework. The decomposition results are presented in Figs. 5 and 6, respectively.

C. Google Search Data Description

In the proposed SEIS-A framework, one important feature is that the self-query data $\boldsymbol{Q}(t)$ play a significant role in the disease density prediction. Here, the search activities in Google for keywords related to HFMD are considered. Due to the unavailability of Google Correlate [27] in Hong Kong, the keywords that are related to HFMD have to be chosen carefully. Meanwhile, since Hong Kong is a city with mixed cultures and languages, both traditional Chinese and English are widely used, and keywords related to HFMD in both languages should be considered to improve the prediction precision. The keywords selected for HFMD spreading prediction are based on the previous results [24]–[26].

hand foot and mouth disease	hand foot and mouth
hand foot mouth	hand foot mouth disease
hfmd	herpes
ulcer	anorexia
手足口病(hfmd)	手足口病症状(hfmd symptoms)
手足口病病徵(hfmd symptoms)	手足口病治療(hfmd treatment)
手足口病图片(hfmd pictures)	皰疹(herpes)
潰瘍(ulcer)	厭食(anorexia)

Fig. 7. Keywords selected for HFMD prediction.

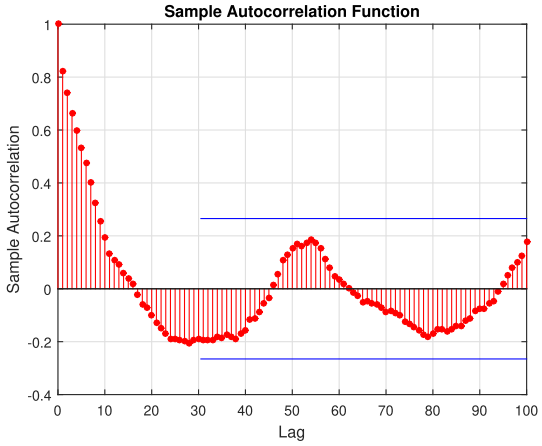

Fig. 8. ACF results of GOPC + GP from week 1, 2010 to week 13, 2018.

The normalized weekly search frequency data for the 16 keywords in Fig. 7 are collected from Google Trends [38]. The search frequency data are collected from week 1, 2010 to week 13, 2018. Meanwhile, since the self-query data on Google are one week earlier than the consultation data announced by CHP, the disease density prediction can be achieved by using the current self-query data on Google and the previous consultation data.

Remark 1: To be noticed, the self-query data $\boldsymbol{Q}(t+1)$ are different from the weekly consultation rate data $\boldsymbol{\eta}(t)$. The self-query data are collected by search engines on the Internet like Google, and it records individuals' search activities for keywords related to a certain disease. However, the weekly consultation rate data $\boldsymbol{\eta}(t)$ are directly related to the disease density data $\boldsymbol{I}(t)$ as it is shown in (11). The weekly consultation rate is employed as the prediction objective due to the unavailability of the actual disease density data $\boldsymbol{I}(t)$, while the self-query data $\boldsymbol{Q}(t+1)$ are used as an external input of the proposed method due to the timely access of it.

D. HFMD Spreading Prediction

First, the autocorrelation function (ACF) [39] is utilized to obtain the model order of the weekly consultation rates. Based on the ACF results in Fig. 8, the value of AR order l is chosen as 30. Meanwhile, the length of the window m is selected as 300 in this empirical study.

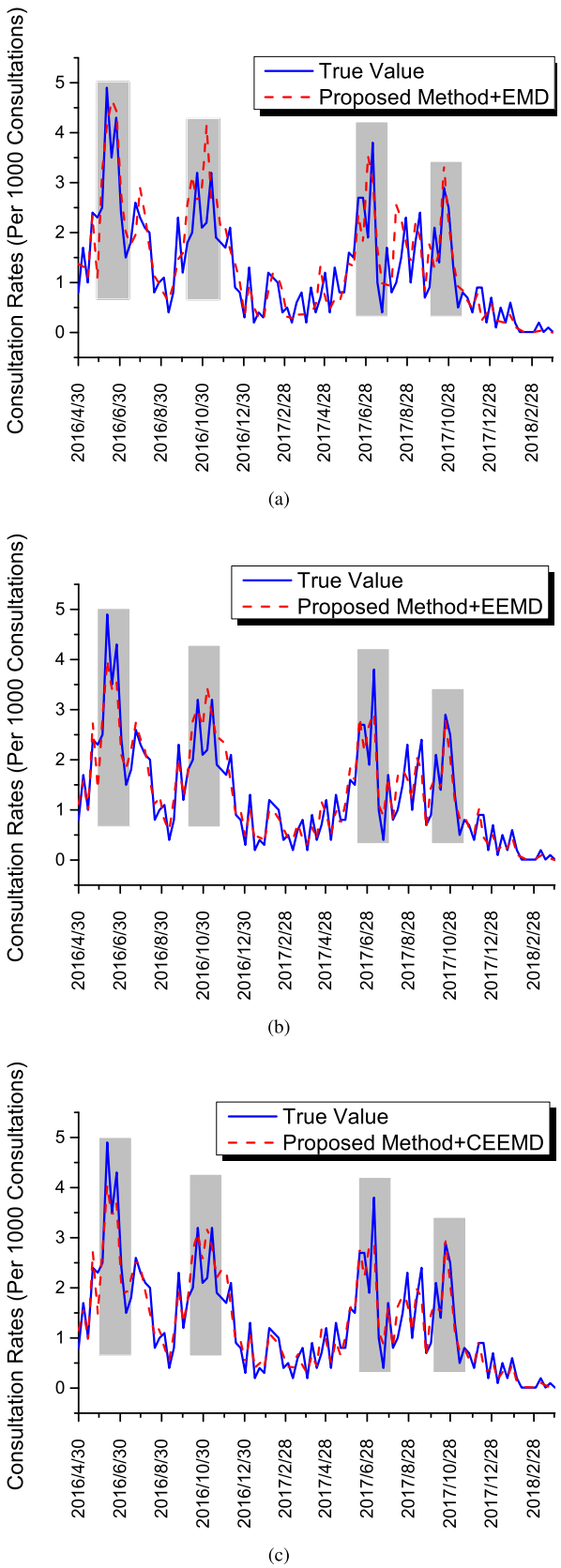

Fig. 9. Prediction results of GOPC + GP from week 18, 2016 to week 13, 2018 based on the proposed framework. (a) EMD. (b) EEMD. (c) CEEMD.

Several common evaluation metrics are adapted to compare the performance of different methods: the root mean square error (RMSE), the mean absolute error (MAE), and the mean

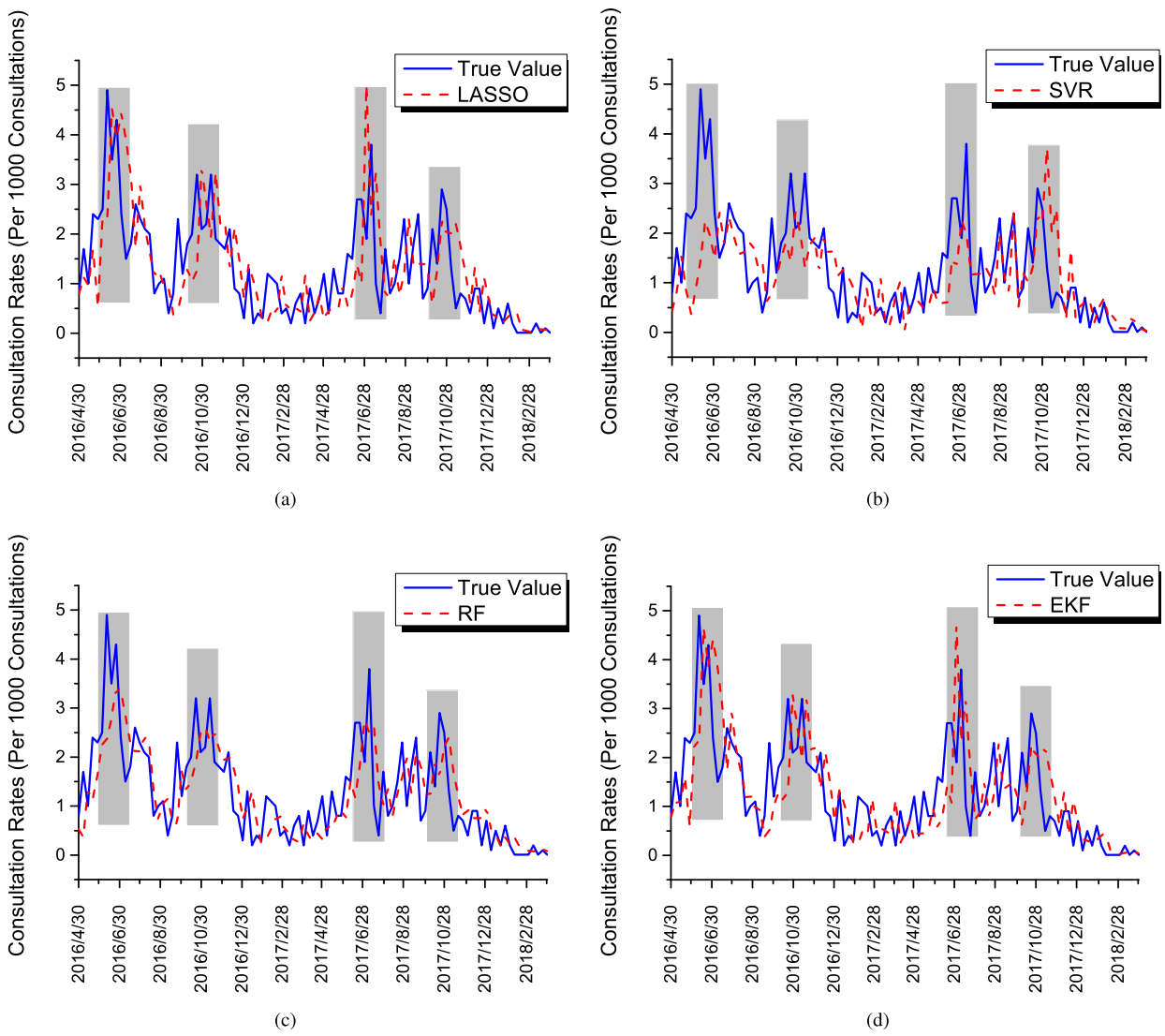

Fig. 10. Prediction results of GOPC+GP from week 18, 2016 to week 13, 2018 by other methods. (a) LASSO. (b) SVR. (c) RF. (d) EKF.

absolute percentage error (MAPE). Denote $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ as a series of the prediction values and $(y_1, y_2, \ldots, y_n)$ as the real values. Then, the above-mentioned metrics have the following forms:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} (\hat{y}_j - y_j)^2}{n}}$$

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^{n} |\hat{y}_j - y_j|$$

$$\text{MAPE} = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{\hat{y}_j - y_j}{y_j} \right|.$$

The prediction results of the proposed method using EMD from week 18, 2016 to week 13, 2018 are presented in Fig. 9(a). During this period, there exist four major outbreaks of HFMD. It is obvious that the proposed method performs well in the disease spreading prediction and all four major outbreaks are predicted with minor errors.

To better illustrate the advantage of the proposed method, three other learning methods, support vector regression (SVR), random forest (RF), and extended Kalman filter (EKF), are compared with the proposed method on the same time series, and the results are presented in Fig. 10(b)–(d). Comparing with Fig. 9(a), it is obvious that the proposed method has an overall advantage over all these methods. In addition, the three evaluation metrics, RMSE, MAE, and MAPE, are calculated for better comparison in Table I. The LASSO without EMD is also investigated, and it is found to be less effective than the EMD-based LASSO.

Moreover, the prediction results of the proposed method with EEMD and CEEMD are presented in Fig. 9(b) and (c), respectively. Based on the value of the metrics in Table I, it is found that the proposed method using EEMD or CEEMD as the decomposition tool can obtain a better performance than that using EMD. This is due to the inherent advantages of EEMD and CEEMD over EMD in mode decomposition. However, since the main contribution claimed in this paper is to introduce an EMD-based ensemble learning

TABLE I
COMPARISON OF DIFFERENT METHODS

	RMSE	MAE	MAPE
Proposed Method+EMD	0.5408	0.3970	0.4043
Proposed Method+EEMD	0.3648	0.2679	0.3020
Proposed Method+CEEMD	0.3415	0.2490	0.2812
LASSO	0.9405	0.6945	1.2815
SVR	0.8785	0.6212	1.3821
RF	0.7108	0.5453	1.4736
EKF	0.9186	0.6812	1.2667

framework for the epidemic spreading prediction problem rather than solving the mode decomposition problem, EMD is used as the decomposition tool in the proposed methodology framework.

Remark 2: The reasons for using EMD as the decomposition tool are twofold. First, EMD is the most classical method for such a mode decomposition problem. Second, it has not been used on the epidemic spreading prediction problem to the best of our knowledge. On the one hand, although using the EEMD or CEEMD method as the decomposition tool can further improve the prediction performance, it may still be improved by the state-of-the-art EMD extensions. On the other hand, the proposed method using EMD as the decomposition tool already performs better over all the methods without mode decomposition. Based on the above-mentioned considerations, EMD is used as the decomposition tool in the proposed method.

V. CONCLUSION

In this paper, a unified SEIS-A framework is proposed to combine individuals' self-query behaviors on the Internet with the epidemic spreading process. An epidemic spreading prediction method that combines EMD with ensemble learning techniques is established based on the proposed SEIS-A framework. An empirical study on the prediction of weekly consultation rates of HFMD in Hong Kong based on the self-query data on Google is conducted. The results indicate that the proposed method outperforms other learning methods. The future direction may be the combination of epidemic spreading over complex networks with evolutionary computation [40]–[43].

APPENDIX

A. Introduction to LASSO

The objective function for parameters ($\boldsymbol{\theta}$) estimation of LASSO is

$$\hat{\boldsymbol{\theta}} = \arg\min \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \tag{12}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ denote the predictor variables and the responses, respectively. Here, $\boldsymbol{\phi}$ are the basis functions and λ is the regularization coefficient. $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the L2-norm and L1-norm, respectively. In order to optimize the objective function, the parameters $\boldsymbol{\theta}$ can be formulated as a difference between two vectors with positive entries

$$\boldsymbol{\theta} = \boldsymbol{\theta}^+ - \boldsymbol{\theta}^-$$
$$\text{s.t. } \boldsymbol{\theta}^+ \geqslant \mathbf{0}, \boldsymbol{\theta}^- \geqslant \mathbf{0}.$$

Hence, the objective function can be regenerated as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}^+,\boldsymbol{\theta}^-} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\phi}^T(\boldsymbol{x})(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)\|_2^2 + \lambda\|(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)\|_1$$
$$\text{s.t. } \boldsymbol{\theta}^+ \geqslant \mathbf{0}, \boldsymbol{\theta}^- \geqslant \mathbf{0}. \tag{13}$$

Define $\boldsymbol{X} = [\begin{smallmatrix}\theta^+ \\ \theta^-\end{smallmatrix}]$, and then, the objective function can be further simplified to the compact form of quadratic programming

$$\min \frac{1}{2}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{X} + \boldsymbol{c}^T\boldsymbol{X} \tag{14}$$

where

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^T(\boldsymbol{x}) & -\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^T(\boldsymbol{x}) \\ -\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^T(\boldsymbol{x}) & \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^T(\boldsymbol{x}) \end{bmatrix}, \quad \boldsymbol{c} = \lambda\mathbf{1} - \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{y} \\ -\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{y} \end{bmatrix}.$$

Here, $\mathbf{1}$ is the column vector of one. The parameters $\boldsymbol{\theta}$ can be obtained by solving this optimization problem. The main advantage of LASSO is that some parameters of trivial features shrink to zero by the L1-norm, and hence, the model complexity can be reduced.

REFERENCES

[1] D. Bernoulli, "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir," *Histoire de l'Acad., Roy. Sci. (Paris) Avec Mem*, pp. 1–45, 1760.

[2] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Rev. Mod. Phys.*, vol. 87, no. 3, p. 925, 2015.

[3] Q. Liu and P. Van Mieghem, "Burst of virus infection and a possibly largest epidemic threshold of non-Markovian susceptible-infected-susceptible processes on networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 97, no. 2, Feb. 2018, Art. no. 022309.

[4] K. Devriendt and P. Van Mieghem, "Unified mean-field framework for susceptible-infected-susceptible epidemics on networks, based on graph partitioning and the isoperimetric inequality," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 96, no. 5, Nov. 2017, Art. no. 052314.

[5] Y. Feng, Q. Fan, L. Ma, and L. Ding, "Epidemic spreading on uniform networks with two interacting diseases," *Phys. A, Stat. Mech. Appl.*, vol. 393, pp. 277–285, Jan. 2014.

[6] Y. Feng, L. Ding, Y.-H. Huang, and L. Zhang, "Epidemic spreading on weighted networks with adaptive topology based on infective information," *Phys. A, Stat. Mech. Appl.*, vol. 463, pp. 493–502, Dec. 2016.

[7] Y. Feng, L. Ding, and P. Hu, "Epidemic spreading on random surfer networks with optimal interaction radius," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 56, pp. 344–353, Mar. 2018.

[8] P. Hu, L. Ding, and T. Hadzibeganovic, "Individual-based optimal weight adaptation for heterogeneous epidemic spreading networks," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 63, pp. 339–355, Oct. 2018.

[9] Y. Huang, L. Ding, and Y. Feng, "A novel epidemic spreading model with decreasing infection rate based on infection times," *Phys. A, Stat. Mech. Appl.*, vol. 444, pp. 1041–1048, Feb. 2016.

[10] Y. Huang, L. Ding, Y. Feng, and J. Pan, "Epidemic spreading in random walkers with heterogeneous interaction radius," *J. Stat. Mech., Theory Exp.*, vol. 2016, no. 10, Oct. 2016, Art. no. 103501.

[11] Y. Feng, L. Ding, Y.-H. Huang, and Z.-H. Guan, "Epidemic spreading on random surfer networks with infected avoidance strategy," *Chin. Phys. B*, vol. 25, no. 12, 2016, Art. no. 128903.

[12] B. Steinegger, A. Cardillo, P. De Los Rios, J. Gómez-Gardeñes, and A. Arenas, "Interplay between cost and benefits triggers nontrivial vaccination uptake," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 97, no. 3, Mar. 2018, Art. no. 032308.

[13] D. Soriano-Paños, L. Lotero, J. Gómez-Gardeñes, and A. Arenas. (2018). "A framework for epidemic spreading in multiplex networks of metapopulations," [Online]. Available: https://arxiv.org/abs/1802.03969

[14] A. Moinet, R. Pastor-Satorras, and A. Barrat, "Effect of risk perception on epidemic spreading in temporal networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 97, no. 1, Jan. 2018, Art. no. 012313.

[15] K. Omata, "Nonequilibrium statistical mechanics of a susceptible-infected-recovered epidemic model," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 96, no. 2, Aug. 2017, Art. no. 022404.

[16] W. K. Chai and G. Pavlou, "Path-based epidemic spreading in networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 565–578, Feb. 2017.

[17] L. L. Ghezzi and C. Piccardi, "Pid control of a chaotic system: An application to an epidemiological model," *Automatica*, vol. 33, no. 2, pp. 181–191, Feb. 1997.

[18] C.-H. Li, C.-C. Tsai, and S.-Y. Yang, "Analysis of epidemic spreading of an sirs model in complex heterogeneous networks," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 19, no. 4, pp. 1042–1054, Apr. 2014.

[19] R. Almeida, "Analysis of a fractional SEIR model with treatment," *Appl. Math. Lett.*, vol. 84, pp. 56–62, Oct. 2018.

[20] S. Funk, E. Gilad, C. Watkins, and V. A. A. Jansen, "The spread of awareness and its impact on epidemic outbreaks," *Proc. Nat. Acad. Sci.*, vol. 106, no. 16, pp. 6872–6877, 2009.

[21] C. Granell and S. Gómez, and A. Arenas, "Dynamical interplay between awareness and epidemic spreading in multiplex networks," *Phys. Rev. Lett.*, vol. 111, no. 12, Sep. 2013, Art. no. 128701.

[22] Q. Wu, X. Fu, M. Small, and X.-J. Xu, "The impact of awareness on epidemic spreading in networks," *Chaos, An Interdiscipl. J. Nonlinear Sci.*, vol. 22, no. 1, Jan. 2012, Art. no. 013101.

[23] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, p. 1012, Feb. 2009.

[24] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in China with search query from baidu," *PLoS ONE*, vol. 8, no. 5, May 2013, Art. no. e64323.

[25] P. Guo *et al.*, "Monitoring seasonal influenza epidemics by using Internet search data with an ensemble penalized regression model," *Sci. Rep.*, vol. 7, Apr. 2017, Art. no. 46469.

[26] Q. Xu, Y. R. Gel, L. L. R. Ramirez, K. Nezafati, Q. Zhang, and K.-L. Tsui, "Forecasting influenza in Hong Kong with Google search queries and statistical model fusion," *PLoS ONE*, vol. 12, no. 5, May 2017, Art. no. e0176690.

[27] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 47, pp. 14473–14478, Nov. 2015.

[28] Y.-H. Wang, C.-H. Yeh, H.-W. V. Young, K. Hu, and M.-T. Lo, "On the computational complexity of the empirical mode decomposition algorithm," *Phys. A, Stat. Mech. Appl.*, vol. 400, pp. 159–167, Apr. 2014.

[29] G. Wang, X.-Y. Chen, F.-L. Qiao, Z. Wu, and N. E. Huang, "On intrinsic mode function," *Adv. Adapt. Data Anal.*, vol. 2, no. 3, pp. 277–293, Jul. 2010.

[30] N. E. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proc. Roy. Soc. London A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[31] S. Ouaro and D. Ouedraogo, "SEIS model with multiple latent stages and treatment in an exponentially growing population," *Commun. Math. Biol. Neurosci.*, vol. 2017, Nov. 2017.

[32] M. Fan, M. Y. Li, and K. Wang, "Global stability of an SEIS epidemic model with recruitment and a varying total population size," *Math. Biosci.*, vol. 170, no. 2, pp. 199–208, Apr. 2001.

[33] J. Liu and F. Wei, "Dynamics of stochastic SEIS epidemic model with varying population size," *Phys. A, Stat. Mech. Appl.*, vol. 464, pp. 241–250, Dec. 2016.

[34] R. Xu, "Global dynamics of an SEIS epidemiological model with time delay describing a latent period," *Appl. Math. Comput.*, vol. 218, no. 15, pp. 7927–7938, Nov. 2012.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[36] *The Centre for Health Protection (CHP)*. Accessed: Apr. 12, 2018. [Online]. Available: https://www.chp.gov.hk/en/static/24015.html

[37] W. Yang, B. J. Cowling, E. H. Lau, and J. Shaman, "Forecasting influenza epidemics in Hong Kong," *PLoS Comput. Biol.*, vol. 11, no. 7, 2015, Art. no. e1004383.

[38] *Google Trends*. Accessed: Apr. 12, 2018. [Online]. Available: https://trends.google.com/trends/

[39] H. Liu, C. Chen, H.-Q. Tian, and Y.-F. Li, "A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks," *Renew. Energy*, vol. 48, pp. 545–556, Dec. 2012.

[40] Y. Wang, B.-C. Wang, H.-X. Li, and G. G. Yen, "Incorporating objective function information into the feasibility rule for constrained evolutionary optimization," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2938–2952, Dec. 2016.

[41] B.-C. Wang, H.-X. Li, J.-P. Li, and Y. Wang, "Composite differential evolution for constrained evolutionary optimization," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published. doi: 10.1109/TSMC.2018.2807785.

[42] B.-C. Wang, H.-X. Li, and Y. Feng, "An improved teaching-learning-based optimization for constrained evolutionary optimization," *Inf. Sci.*, vol. 456, pp. 131–144, Aug. 2018.

[43] Y. Feng, B.-C. Wang, and L. Ding. (2019). "A constrained cooperative coevolution strategy for weights adaptation optimization of heterogeneous epidemic spreading networks." [Online]. Available: https://arxiv.org/abs/1901.00602

Yun Feng received the B.E. degree in automation and the M.S. degree in control theory and control engineering from the Department of Automation, Wuhan University, Wuhan, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong.

His current research interests include fault diagnosis of distributed parameter systems and complex networks.

Bing-Chuan Wang received the B.E. degree in automation and the M.S. degree in control science and engineering from Central South University, Changsha, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the City University of Hong Kong, Hong Kong.

His current research interests include evolutionary computation and intelligent modeling.