SCIENTIFIC REPORTS

OPEN

Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model

Received: 04 October 2016
Accepted: 20 March 2017
Published: 19 April 2017

Pi Guo, Jianjun Zhang, Li Wang, Shaoyi Yang, Ganfeng Luo, Changyu Deng, Ye Wen & Qingying Zhang

Seasonal influenza epidemics cause serious public health problems in China. Search queries-based surveillance was recently proposed to complement traditional monitoring approaches of influenza epidemics. However, developing robust techniques of search query selection and enhancing predictability for influenza epidemics remains a challenge. This study aimed to develop a novel ensemble framework to improve penalized regression models for detecting influenza epidemics by using Baidu search engine query data from China. The ensemble framework applied a combination of bootstrap aggregating (bagging) and rank aggregation method to optimize penalized regression models. Different algorithms including lasso, ridge, elastic net and the algorithms in the proposed ensemble framework were compared by using Baidu search engine queries. Most of the selected search terms captured the peaks and troughs of the time series curves of influenza cases. The predictability of the conventional penalized regression models were improved by the proposed ensemble framework. The elastic net regression model outperformed the compared models, with the minimum prediction errors. We established a Baidu search engine queries-based surveillance model for monitoring influenza epidemics, and the proposed model provides a useful tool to support the public health response to influenza and other infectious diseases.

Seasonal influenza is a serious public health problem that causes severe illness and death in the world. According to the World Health Organization (WHO), seasonal influenza occurs with an annual attack rate estimated at 5% to 10% in adults and 20% to 30% in children. The epidemics are estimated to result in about 3 to 5 million cases of severe illness and 250,000 to 500,000 deaths worldwide each year[1]. During 2008–2011, an annual average of 92,677 seasonal influenza cases was reported in China[2]. Overall, the influenza pandemics posed a significant burden of excess influenza-associated mortality in the country[3]. To achieve near real-time surveillance of the spread of infectious diseases, several novel approaches based on online surveillance systems and using informal sources such as news reports[4], social media data[5,6], and search query data[7,8] have been proposed.

In 2009, Ginsberg, J. *et al.*[8] first presented a novel method of analyzing large numbers of Google search queries to track influenza-like illness in the United States. The proposed method provided near real-time estimates of seasonal influenza activity each day and overcame the limitation of traditional systems requiring 1–2 weeks to gather and process surveillance data[8]. To estimate the seasonal influenza activity and quickly detect outbreaks in China, several programs were used to predict trends of influenza epidemics[9,10]. However, these techniques used only influenza-like illness or influenza case data. The robust prediction of influenza epidemics could be improved. In 2013, Yuan, Q. *et al.*[11] first explored the use of the combination of influenza case data and internet search query data from the search engine Baidu within a linear regression framework to monitor influenza epidemics in China. This provided a new idea to monitor the spread of influenza in the country. To inform the search behavior of users, Baidu released the search volume daily on the Baidu Index website (http://index.baidu.com). The search volume of different search keywords used can be abstracted to assess changes in the search behavior of users.

Department of Preventive Medicine, Shantou University Medical College, No. 22 Xinling Road, Shantou, Guangdong, 515041, People's Republic of China. Correspondence and requests for materials should be addressed to Q.Z. (email: qyzhang@stu.edu.cn)

According to Yuan, Q. *et al.*[11], the construction of the prediction model involved compositing many search keywords into a single index according to different weights. However, in practice, many search keywords are used to construct the prediction model. The direct compositing of all keywords into a single index is not convenient for assessing the contribution of each keyword to the prediction. Developing robust techniques of search keyword selection and enhancing the ability to predict influenza epidemics remains challenging. Beyond the use of a linear regression model for prediction, we explored an ensemble framework that incorporated different penalized regression algorithms including lasso, ridge and elastic net[12] to avoid the over-fitting problem with various keywords, identify informative predictors from a pool of candidate keywords, and estimate the parameters of the model with low variability.

In our previous study[13], use of a penalized regression model based on random bootstrap samples[14] was able to detect significant variables with better predictive performance. How well a model predicts is practically quantified by performance measures. For example, performance measures such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC)[15] and kappa index of agreement (KIA)[16] are often used to evaluate performance for classification problems. However, in many settings, the assessment of performance by a single measure has inherent problems[17]. For example, in disease surveillance applications, to predict periods of high incidence of infectious disease requires large sensitivity and/or specificity rates in addition to prediction accuracy[18]. Different performance measures reflect different characteristics of the constructed prediction model. Therefore, under many circumstances, several performance measures must be considered simultaneously.

To improve prediction robustness, we sought to develop a Baidu search engine query data-based prediction model whose performance was optimized with respect to a set of measures. A novel ensemble framework was established by combining bootstrap aggregating (bagging) and a multi-objective optimization method in this study. New ensemble penalized regression models using the lasso, ridge and elastic net algorithms were constructed, and applied to predict seasonal influenza activity. Results of this study indicated that the ensemble elastic net regression model outperformed the compared models in monitoring seasonal influenza activity by using Baidu search engine query data.

Material and Methods

Ensemble penalized regression model.

Penalized regression model. We first considered the lasso (*L1*-penalized regression method) linear regression model[12]. We have an $n \times 1$ response vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and linearly independent predictors $\mathbf{x} = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ $(j = 1, \ldots, p)$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ be the predictor matrix. We assume that $E[y|x] = \beta_1^* x_1 + \cdots + \beta_p^* x_p$. The estimates in the lasso linear regression model are defined as (1):

$$\hat{\beta}_\lambda = \arg\min_\beta \left\| \mathrm{y} - \sum_{j=1}^{p} \mathrm{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \tag{1}$$

where $\|u\|_2^2$ denotes $\sum_1^n u_i^2$ for vector $u \in ¡^n$, and λ is the nonnegative tuning parameter. This estimation method continuously shrinks the coefficients toward 0 as λ increases, and some coefficients are shrunk to exactly 0 if λ is sufficiently large[19].

Next, we considered the lasso logistic regression setup by using the tuning parameter λ. The estimates $\hat{\beta}_\lambda$ in the model are given by (2):

$$\hat{\beta}_\lambda = \arg\min_\beta \sum_{i=1}^{n} (-y_i(\mathrm{x}_i^T \beta) + \log(1 + e^{\mathrm{x}_i^T \beta})) + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \tag{2}$$

where λ is also the tuning parameter used for shrinking coefficients in the model. Generally, the cross-validation method was proposed to select the optimal λ[20]. The ridge and elastic net penalized regression models were established using different penalties[12], and the optimal values of tuning parameters were chose by a similar way.

Ensemble penalized regression model built with a bagging strategy. To improve the performance of the conventional penalized regression model, we used a combination of bagging and a rank aggregation[21] method to develop an ensemble penalized regression model. The architecture of the model consists of a sequence of processing procedures primarily including model training, validation, evaluation and averaging, which are implemented in many random bootstrap samplings (Fig. 1). The details for the methodology are presented below.

According to Breiman, L.[22], bagging is a method of generating multiple versions of a prediction model, and these models are used to obtain an aggregated prediction, which gives substantial gains in prediction accuracy. Suppose that a training set L consisting of data $X_{n \times p}$ with known outcomes $y = (y_1, \ldots, y_n)$ that are independently drawn from the probability distribution P, then we establish a prediction model $\varphi(X, L)$. Here, n is the number of samples and p is the number of predictors. By taking repeated bootstrap samples $\{L^{(B)}\}$ from L, we formed a set of new prediction models $\varphi(X, L^{(B)})$. The final prediction of the bagging model denoted by $\varphi_A(x) = E\varphi(X, L^{(B)})$ was obtained by averaging all results for a number of sub-models. The proof of the validity of bagging on improving prediction accuracy is given in the Methods section of the Supplemental Material.

To build the ensemble model, we randomly drew several (B) bootstrap samples from the original data $\{X_{n \times p}, y_{n \times 1}\}$, trained B penalized regression models, $M^1, M^2, \ldots, M^B$, by using the bootstrap samples and combined them to obtain an aggregated prediction. To determine an optimal sub-model in the ensemble penalized regression model according to several performance measures during each random sampling, we used a multi-objective optimization method via the weighted rank aggregation[21]. First, each measure ranked the sub-models according to their performance under that particular measure and generated the ordered lists of sub-models, $R_1, \ldots, R_K$, where

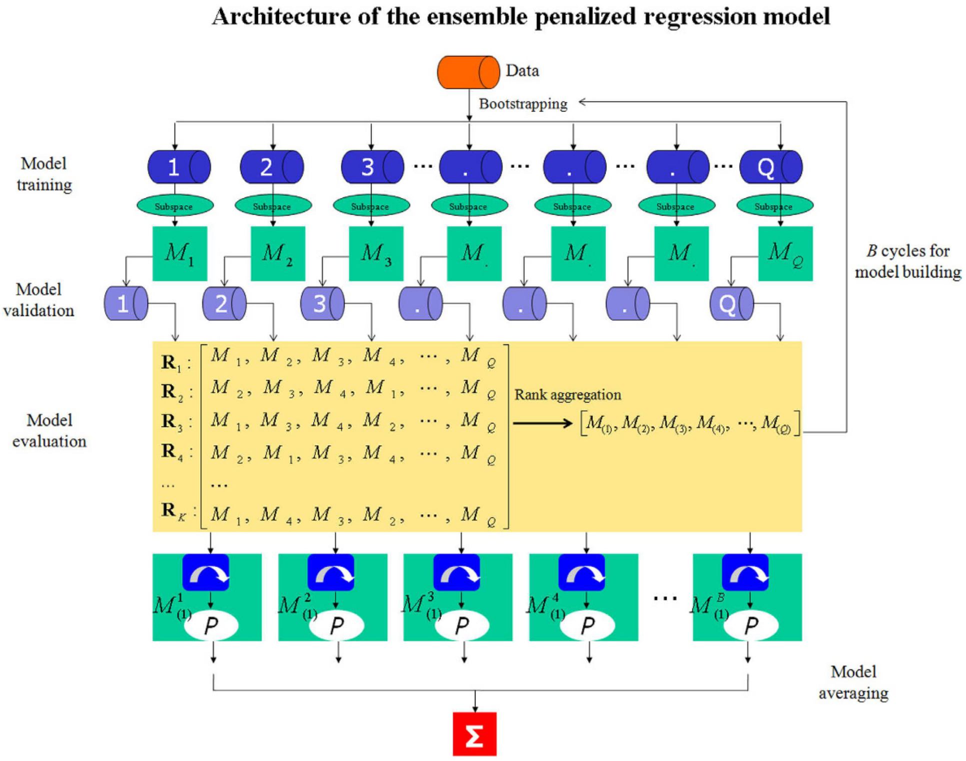
Architecture of the ensemble penalized regression model
Data
Bootstrapping
Model training
1
2
3
.
.
.
Q
Subspace
M_1
M_2
M_3
$M_.$
M_Q
B cycles for model building
Model validation
Model evaluation
R_1 : $M_1, M_2, M_3, M_4, \cdots, M_Q$
R_2 : $M_2, M_3, M_4, M_1, \cdots, M_Q$
R_3 : $M_1, M_3, M_4, M_2, \cdots, M_Q$
R_4 : $M_2, M_1, M_3, M_4, \cdots, M_Q$
R_K : $M_1, M_4, M_3, M_2, \cdots, M_Q$
Rank aggregation
$[M_{(1)}, M_{(2)}, M_{(3)}, M_{(4)}, \cdots, M_{(Q)}]$
$M^1_{(1)}$
$M^2_{(1)}$
$M^3_{(1)}$
$M^4_{(1)}$
$M^B_{(1)}$
P
Model averaging
Σ

Figure 1. Architecture of the ensemble penalized regression model. A sequence of processing procedures, primarily including model training, validation, evaluation and averaging to be implemented in random bootstrap samplings in this architecture.

K is the number of measures used. Second, the weighted rank aggregation approach was used to produce an aggregated list that ranked the sub-models according to their performance under all K measures simultaneously. To obtain the optimal ordered list of models, we defined the following objective function:

$$\Phi(\delta) = \sum_{i=1}^{K} w_i d(\delta, R_i) \tag{3}$$

where δ is an ordered list of models of size Q, d is a distance function that estimates the similarity between any two ordered lists, and w_i is a weight factor associated with each measure. The Spearman footrule distance function[23] was used to estimate the similarity between any two lists of models.

To determine an optimal model according to all K measures simultaneously, it is equivalent to seek out an optimal list δ^* to minimize the value of the objective function $\Phi(\delta)$. To determine the optimal parameter δ^*, the cross-entropy method was used for rank aggregation[24]. The algorithm of the ensemble penalized regression model is given as follows:

Algorithm. Ensemble penalized regression model.

Input:

- (X, y): training set that contains n samples and a p-dimensional vector of predictors, and $(X, y) \in ¡^{n\times(p\times1)}$.
- B: number of random bootstrap samplings.
- $n_{bootstrap}$: size of random bootstrap samples with replacement.
- Q: size of an ordered list of sub-models in the ensemble model.
- K: number of performance measures.
- RP: size of random subspace predictor.
- δ: an initial ordered list of sub-models of size L.
- $d(.)$: the Spearman footrule distance function.

Output: prediction $\psi_{average}$ of the ensemble model.

```
for $b = 1$ to $B$ do
generate bootstrap samples $(X^b, y^b) \in ¡^{n_{bootstrap}\times(p+1)}$
generate out-of-bag (OOB) samples $L_{OOB} = (X^b_{OOB}, y^b_{OOB}) \in ¡^{n_{OOB}\times(p+1)}$
for $q = 1$ to $Q$ do
randomly select $RP$ predictors as a subset from the original $P$ predictors
```

Month	Cases	Month	Cases	Month	Cases	Month	Cases	Month	Cases
2011–01	6072	2011–12	11631	2012–11	8942	2013–10	9309	2014–09	9751
2011–02	5930	2012–01	10046	2012–12	12411	2013–11	12317	2014–10	8635
2011–03	7299	2012–02	17421	2013–01	16012	2013–12	23894	2014–11	12043
2011–04	5727	2012–03	21625	2013–02	9806	2014–01	38214	2014–12	25477
2011–05	4130	2012–04	10707	2013–03	10761	2014–02	25279	2015–01	23828
2011–06	3065	2012–05	8520	2013–04	10844	2014–03	27262	2015–02	14480
2011–07	2654	2012–06	6195	2013–05	9006	2014–04	14166	2015–03	19199
2011–08	3243	2012–07	6738	2013–06	6254	2014–05	12685	2015–04	15063
2011–09	4360	2012–08	6793	2013–07	5338	2014–06	19220	2015–05	13625
2011–10	5525	2012–09	6762	2013–08	6098	2014–07	14056		
2011–11	7055	2012–10	7331	2013–09	8751	2014–08	11419		

Table 1. Data of influenza cases confirmed by laboratory test for the period January 2010 to May 2015 in China were publicly available from China's Center for Disease Control and Prevention.

```
generate a new subset of predictors $\mathring{X} = [x_1, x_2, \ldots, x_{RP}]$
generate new bootstrap samples $\mathring{L} = (\mathring{X}, y^b) \in ¡^{n_{bootstrap}\times(RP+1)}$
generate new OOB samples $\mathring{L}_{OOB} = (\mathring{X}_{OOB}, y^b_{OOB}) \in ¡^{n_{OOB}\times(RP+1)}$
establish a penalized regression model $\varphi_q(\mathring{X}, \mathring{L})$
for $k=1$ to $K$ do
compute performance measures $w_{q,k}$ based on OOB samples $\mathring{L}_{OOB}$
end
end
generate a matrix of performance measures $W_{K\times Q} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1Q} \\ w_{21} & w_{22} & \cdots & w_{2Q} \\ \cdots & \cdots & \cdots & \cdots \\ w_{K1} & w_{K2} & \cdots & w_{KQ} \end{bmatrix}$, where the measures in each row
$(w_{i1}, w_{i2}, \ldots, w_{iQ})$ were ranked in order of descending values
generate $K$ ordered list of sub-models $\{R_i = (M_1, M_2, \ldots, M_Q)^i, i = 1, \ldots, K\}$ according to $W_{K\times Q}$
establish the objective function $\Phi(\delta) = \sum_{i=1}^{K} w_i d(\delta, R_i)(w_i = (w_{i1}, w_{i2}, \ldots, w_{iQ}))$
perform the cross-entropy method for rank aggregation and to determine the optimal parameter $\delta^*$ minimizing
the value of $\Phi(\delta)$
obtain an optimal ordered list of sub-models $R_b^* = \delta^* = (M_{(1)}, M_{(2)}, \ldots, M_{(Q)})$
end
establish the ensemble penalized regression model according to $B$ optimal sub-models
$\{M_{(1)}(R_1^*), M_{(1)}(R_2^*), M_{(1)}(R_3^*), \ldots, M_{(1)}(R_B^*)\}$
produce the prediction $\psi_{average} = E\left[M_{(1)}(R^*)\right]$ via model averaging.
```

From our experience, the model performed similarly when parameter B was large, for example, $B=100$. The value of $n_{bootstrap}$ was set to the size of the original data. The size of the ordered list of sub-models Q was set to 10 to ensure efficiency and fast convergence[21]. Previous studies[25,26] suggested that the random subspace method usually produced an improved ensemble model. Thus, we constructed the ensemble model by using a random subset of predictors, $p^* = \sqrt{p}$, as proposed by Breiman, L.[27]. To assess the contribution of each predictor in the ensemble model, we used a permutation method to estimate the importance of each predictor as follows:

$$I_j = \frac{1}{B}\sum_{i=1}^{B}\left(e_i^{X_{(j)}^{OOB}} - e_i^{X^{OOB}}\right) \tag{4}$$

where $\boldsymbol{I_j}$ is the importance score of predictor j, $X_{(j)}^{OOB}$ represents the OOB samples with the j^{th} predictor randomly permuted, $\boldsymbol{X^{OOB}}$ is the non-permuted samples, and $\boldsymbol{e_i}$ is the error rate of prediction. The architecture of the ensemble penalized regression model is depicted in Fig. 1.

Model evaluation. To widen the application of the ensemble model, we considered two set-ups of the model including the logistic and linear regression models for monitoring influenza epidemics. For the logistic regression model, we used five performance measures, including accuracy, sensitivity, specificity, AUC[15] and KIA[16]. For the linear regression model, we used relative error (RE), root mean square error (RMSE), mean absolute error (MAE) and symmetric mean absolute percentage error (SMAPE)[28] to assess performance.

Application to monitor seasonal influenza activity.

Data sources. This study used monthly case counts of influenza occurring from January 2011 to May 2015 in China for testing the model. These laboratory-confirmed cases of influenza were reported by physicians to the notifiable disease-monitoring system managed by China's Center for Disease Control and Prevention, and the data are publicly available on the official website (http://www.moh.gov.cn/). The influenza surveillance data for the studied period corresponded to a total of 53 months of influenza cases. Table 1 shows the details of monthly influenza case counts used in this study.

Search query data were obtained from the Baidu Index website, which contains logs of online search query volume for numerous keywords searched by Baidu users. Since the search query data were available on a daily basis, we converted the data to monthly counts over the study period for analysis.

Keyword selection, crawling and filtering. Previous studies generally chose the names or clinical symptoms of the studied diseases as the primary terms to find more related keywords[11,29,30]. From this idea, we used the term "influenza" ("流感" in Chinese) as a primary keyword to search for more keywords associated with the studied disease on a Chinese website (http://tool.chinaz.com/baidu/words.aspx). The recommended keywords were comprehensively extracted from different sources, including Baidu, portal websites, and blogs[11]. On typing in the primary keyword, a total of 100 related keywords were obtained for further analysis (Table 2). After determining the related keywords, we established an auto-crawler by using Python and used it to collect search volume data for the keywords. The framework of an auto-crawler is depicted in Fig. 2. The Python scripts could be available from the authors for academic usage.

Because some recommended keywords were not necessarily related to influenza epidemics, we further filtered the keywords in three steps: first, the selected search keywords should represent factors that might affect the influenza epidemic; second, the search volume data for each keyword could be presented as a sequential time series with a specific resolution of time (e.g., daily, weekly or monthly); third, the time series of selected keywords should have a maximum cross-correlation coefficient of at least 0.4 with the influenza case data. These filtering approaches were also proposed in previous studies[11,30].

We considered two scenarios of model validation. First, the influenza case surveillance data were divided into a fitting and validation dataset. Models were fitted by using data from January 2011 to June 2014, and the remaining part of the data was used for model validation. Second, to compare the models for monitoring a high level of influenza epidemics, we investigated three cases of high incidence thresholds defined as the median, 75th and 90th percentiles of number of influenza cases over the study period, and evaluated their performance. The receiver operating characteristic (ROC) curve was used to assess the predictive ability of the models.

Results

On the basis of our filtering steps, 19 of the 100 keywords were not related to influenza epidemics, 8 keywords did not have sequential time series due to low search volume, and a set of only 58 keywords was retained for building the compared models (Table 2). Taking into account the delayed effects of predictors, we considered time lags of 0 to 1 month and the autoregressive term of influenza case number in the previous month. In total, 117 predictors were used for building the prediction models. In this case, the number of predictors was more than the length of time series of influenza cases ($117 > 53$). Thus, the penalized estimation of parameters in the model was necessary in this study.

In general, influenza causes annual epidemics that peak during the spring and winter in China. Most of our selected search keywords captured the peaks and troughs of the time series curves of influenza cases, so they were good indicators for monitoring influenza epidemics in the country (Figures S1–S5).

Comparison of prediction performance of different penalized regression models and the algorithms in the proposed ensemble framework is shown in Table 3. For the prediction of seasonal influenza case counts in the period between July 2014 and May 2015, the ensemble framework improved the performance of the conventional lasso, ridge and elastic net regression models. Among the models, the ensemble elastic net regression model outperformed the others since it had the smallest prediction errors (Table 3). Regardless of the periods for model fitting and prediction, the ensemble elastic net regression model was able to capture the peaks and troughs of the time series curves of influenza cases (Fig. 3). The forecast intervals given by the ensemble model well covered the actual epidemic curve of influenza cases.

For monitoring a high level of influenza epidemics, this study integrated the set-up of logistic regression models in the ensemble prediction framework. We studied three situations of high incidence thresholds defined as the median, 75th and 90th percentiles of number of influenza cases over the study period. The performance of the models to detect a large number of influenza cases was assessed using the measures including accuracy, sensitivity, specificity, AUC and KIA (Table 4). Overall, the ensemble elastic net regression model had the largest average AUC of 0.97, and thus outperformed the others, irrespectively of thresholds of influenza incidence used. In addition, it suggested that the predictability of the conventional lasso, ridge and elastic net models was consistently improved by the ensemble framework (Fig. 4).

Figure 5 shows the estimated importance score for the top 25 keywords contributing to the prediction of the ensemble model. The keyword, "*type a flu*" (variable $X39$), was the most significant factor predicting influenza epidemics over the study period. In addition, the keywords "*saying type a h1n1 flu*" (variable $X99$), "*the toll of swine flu-related death*" (variable $X52$) and "*flu symptom*" (variable $X47$) played important roles in the internet search queries-based surveillance model we established. The ensemble elastic net regression model performed similarly with a large number of random bootstrap samplings, for example, with $B = 100$ (Figure S6). It also guaranteed that the prediction of the ensemble model converged to a stable result.

Discussion

We used bagging and a multi-objective optimization technology to establish a novel ensemble elastic net penalized regression model to detect seasonal influenza epidemics in China. The results revealed high performance and small fluctuation of extrapolating ability for the proposed model as a Baidu search engine queries-based surveillance framework. The empirical analysis demonstrated that monitoring seasonal influenza epidemics was better with our ensemble models than the conventional penalized regression models.

Variable	Chinese name	English name	Include (√)/exclude (×)	Variable	Chinese name	English name	Include (√)/ exclude (×)
X1	香港流感	Hongkong flu	√	X51	***甲型h1n1流感病毒***	***type a h1n1 flu virus***	×
X2	***流感***	***Flu***	×	X52	猪流感死亡人数	the toll of swine flu-related death	√
X3	***流感症状***	***flu symptom***	×	X53	北京 流感	Beijing (space) flu	√
X4	流感疫苗	flu vaccine	√	X54	流感 韩国	flu (space) Korea	√
X5	***猪流感***	***swine flu***	×	X55	猪流感的症状	the symptom of swine flu	√
X6	流感病毒	flu virus	√	X56	**狗流感**	**dog flu**	×
X7	流感疫苗有必要打吗	the flu vaccine is necessary to play	√	X57	**关颖 上流感**	**Guanying up flu**	×
X8	甲型流感	influenza A	√	X58	季节性流感	seasonal influenza	√
X9	流感的症状	the symptom of flu	√	X59	流感疫苗价格	influenza vaccine price	√
X10	西班牙流感	Spanish flu	√	X60	猪流感怎么治疗	how to treat swine flu	√
X11	*美国流感疫情蔓延*	*US flu spread*	×	X61	*流感疫苗接种时间*	*influenza vaccination time*	×
X12	甲型h1n1流感	type a h1n1 flu	√	X62	h7n9流感症状	h7n9 flu symptom	√
X13	**流感 韩国电影**	**flu Korean movies**	×	X63	甲型h1n1流感不可怕	type a h1n1flu is not terrible	√
X14	流感吃什么药	influenza drugs	√	X64	***流感传播途径***	***the routes of flu transmission***	×
X15	**流感概念股**	**influenza concept stock**	×	X65	流感的预防	the prevention of flu	√
X16	流感嗜血杆菌	haemophilus influenzae	√	X66	**上流感 关颖**	**up flu Guanying**	×
X17	预防流感	prevent flu	√	X67	甲型h7n9流感症状	type a h7n9 flu symptom	√
X18	流感丸	flu pill	√	X68	猪流感疫情	swine flu epidemic	√
X19	北京流感	Beijing flu	√	X69	a型流感病毒	type a flu virus	√
X20	**情流感**	**love flu**	×	X70	甲型h1n1流感症状	type a h1n1 flu symptom	√
X21	**情流感菌**	**love flu strain**	×	X71	*甲型h3n2流感病毒*	*type 1 h3n2 flu virus*	×
X22	流感疫情	influenza epidemic	√	X72	**甲型hini流感症状**	**type a hini flu symptom**	×
X23	***超级流感***	***super flu***	×	X73	流感大流行	influenza pandemic	√
X24	**三月 情流感**	**March love flu**	×	X74	流感的传播途径	The route of transmission of flu	√
X25	台湾流感	Taiwang flu	√	X75	流感的预防措施	the prevention measures of flu	√
X26	流感疫苗副作用	side effects of flu vaccine	√	X76	美国 流感	US (space) flu	√
X27	***猪流感症状***	***swine flu symptom***	×	X77	猪流感的治疗	the therapy of swine flu	√
X28	如何预防流感	how to prevent flu	√	X78	***猪流感疫苗***	***swine flu vaccine***	×
X29	h7n9流感	h7n9 flu	√	X79	*h3n2流感病毒*	*h3n2 flu virus*	×
X30	病毒性流感	viral flu	√	X80	甲型流感 症状	type a flu symptom	√
X31	上海流感	Shanghai flu	√	X81	*甲型流感疫苗*	*type a flu vaccine*	×
X32	怎样预防流感	how to prevent flu	√	X82	*怎样预防甲型流感*	*how to prevent type a flu*	×
X33	甲型流感病毒	type a flu virus	√	X83	*18号欧洲流感*	*18 European flu*	×
X34	***h1n1流感***	***h1n1 flu***	×	X84	*2012欧洲流感*	*2012 European flu*	×
X35	流感预防	flu prevention	√	X85	h1n1甲型流感	h1n1 type a flu	√
X36	h1n1流感症状	h1n1 flu symptom	√	X86	***h1n1流感的症状***	***the symptom of h1n1 flu***	×
X37	甲型流感症状	the symptom of type a flu	√	X87	**h1n1流感手抄报**	**h1n1 Shouchao Bao**	×
X38	**情流感的症状**	**the symptom of love flu**	×	X88	***h1n1流感有什么症状***	***what are the symptoms of h1n1 influenza***	×
X39	a型流感	type a flu	√	X89	h1n1流感预防知识	h1n1 knowledge of influenza prevention	√
X40	流感治疗	flu therapy	√	X90	***h1n1流感最新疫情***	***the latest outbreak of h1n1 influenza***	×
X41	美国流感	US flu	√	X91	**hini流感**	**hini flu**	×
X42	新型流感	new type flu	√	X92	**hini流感症状**	**the symptom of hini flu**	×
X43	**qq流感大盗**	**qq flu game**	×	X93	**qq流感大盗下载**	**qq flu game download**	×
X44	**上流感**	**up flu**	×	X94	北京甲流感	Beijing influenza a	√
X45	甲流感	A influenza	√	X95	**本山快乐营猪流感**	**Benshan Happy camp swine flu**	×
X46	流感 症状	flu (space) symptom	√	X96	***得了猪流感怎么办***	***how to do when having swine flu***	×
X47	香港流感病毒	Hongkong flu virus	√	X97	***关于甲型h1n1流感***	***about type a h1n1 influenza***	×
X48	***新流感***	***new flu***	×	X98	**关于猪流感的作文**	**about the composition of swine flu**	×
X49	**羊流感**	**goat flu**	×	X99	话说甲型h1n1流感	saying type a h1n1 influenza	√
X50	副流感病毒	parainfluenza virus	√	X100	季节性流感疫苗	seasonal influenza vaccine	√

Table 2. Search keywords from Baidu search engine used in this study. The search terms in bold were excluded at filtering step (i), search terms in italics were excluded at filtering step (ii) and search terms in bold and italics were excluded at filtering step (iii). Step (i): the selected search terms should represent factors that might influence the influenza epidemics. Step (ii): the search query data for each term was a sequential time series with a daily, weekly or monthly resolution. Step (iii): the time series of selected search terms should have a maximum cross-correlation coefficient of at least 0.4 with the influenza case data.

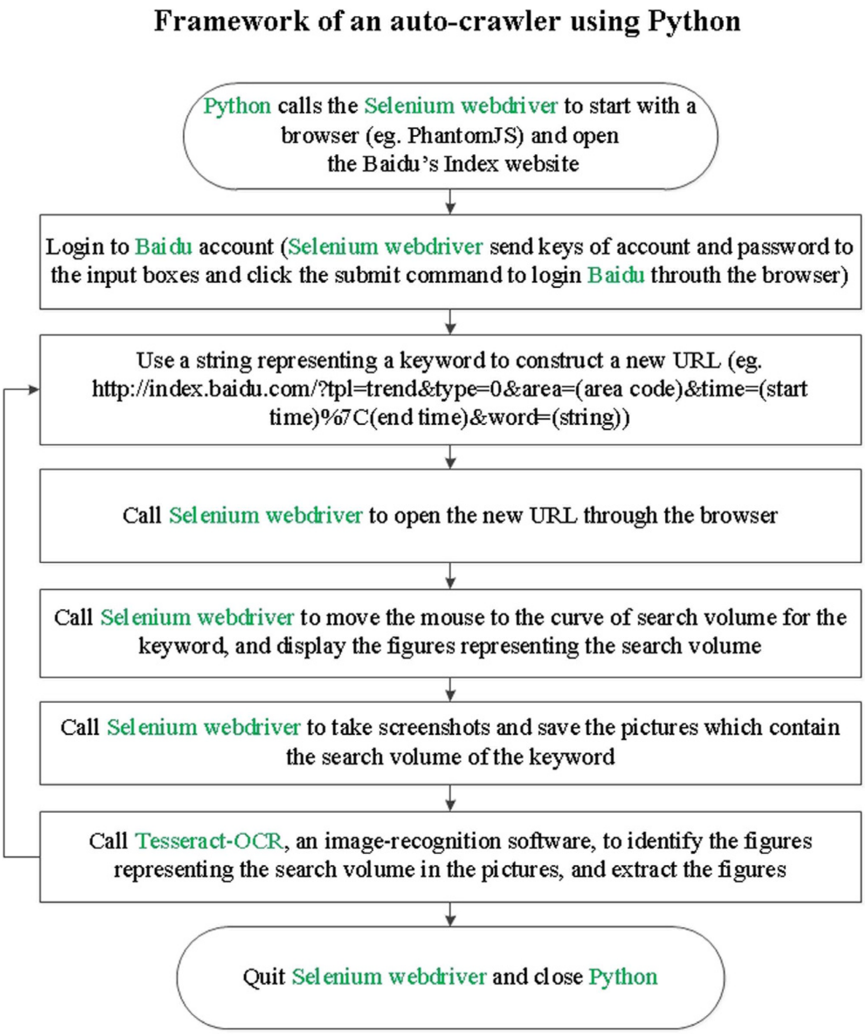
Framework of an auto-crawler using Python
Python calls the Selenium webdriver to start with a browser (eg. PhantomJS) and open the Baidu's Index website
Login to Baidu account (Selenium webdriver send keys of account and password to the input boxes and click the submit command to login Baidu throuth the browser)
Use a string representing a keyword to construct a new URL (eg. http://index.baidu.com/?tpl=trend&type=0&area=(area code)&time=(start time)%7C(end time)&word=(string))
Call Selenium webdriver to open the new URL through the browser
Call Selenium webdriver to move the mouse to the curve of search volume for the keyword, and display the figures representing the search volume
Call Selenium webdriver to take screenshots and save the pictures which contain the search volume of the keyword
Call Tesseract-OCR, an image-recognition software, to identify the figures representing the search volume in the pictures, and extract the figures
Quit Selenium webdriver and close Python

Figure 2. Framework of an auto-crawler using Python to collect search query data from the Baidu Index website. The Selenium webdriver was mainly used with Python for automatic crawling of search query data. The software and search engine used for this analysis are in green.

Prediction period	Model	RMSE	MAE	RE	SMAPE
2014/07–2015/05	ridge	5283.87	4045.57	26.41%	26.53%
	lasso	4702.71	3419.20	25.16%	23.60%
	elastic net	3396.04	2799.94	21.29%	21.19%
	ensemble ridge	4148.28	3285.35	22.73%	21.74%
	ensemble lasso	3756.31	2897.91	20.42%	19.24%
	ensemble elastic net	3488.42	2650.17	19.09%	17.55%

Table 3. Prediction performance of different penalized regression algorithms (lasso, ridge and elastic net) and the algorithms in the proposed ensemble framework was compared using the number of influenza cases during the period of July 2014 to May 2015. Measures including relative error (RE), root mean square error (RMSE), mean absolute error (MAE) and symmetric mean absolute percentage error (SMAPE) were used to assess the predictions.

Recently, Salathé M. *et al.*[31] discussed the importance of digital disease surveillance for rapid disease outbreak detection and proposed it as a powerful tool to complement traditional approaches. In fact, internet search query data is being explored as a low-cost approach to providing near real-time estimates of disease activity and is becoming widely used for disease surveillance[11,18,29,30]. In China, influenza activity based on routine surveillance data from the ministry of health of China was usually reported with a 1 to 2-week lag. Hence, as a convenient source for timely estimating of influenza activity and detecting an epidemic, search query data can contribute to improve the results of traditional disease surveillance.

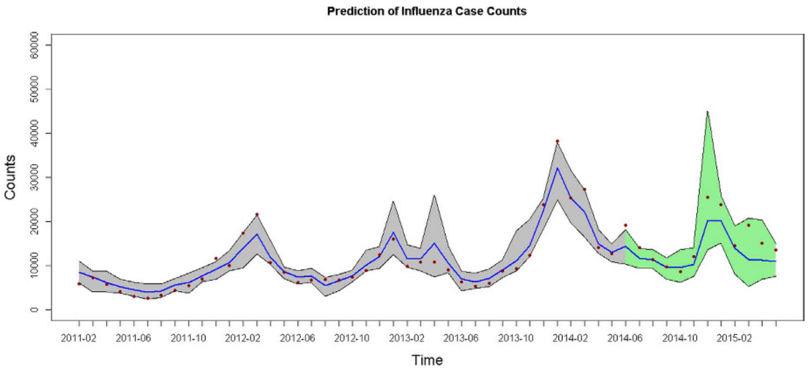

Figure 3. Predictions of influenza cases according to the ensemble elastic net regression model for the period of July 2014 to May 2015. Dark-red dots represent the actual counts of influenza case, blue line represents the fitted counts, and the 95% prediction interval is presented, respectively; the grey areas correspond to the periods used for model fitting and the green areas for prediction.

Percentile cutoff	Model	Acc	Sen	Spe	AUC	KIA
median	ridge	0.92	0.96	0.88	0.92	0.85
	lasso	0.92	0.93	0.92	0.92	0.85
	elastic net	0.98	1.00	0.96	0.98	0.96
	ensemble ridge	0.94	0.96	0.92	0.94	0.88
	ensemble lasso	1.00	1.00	1.00	1.00	1.00
	ensemble elastic net	1.00	1.00	1.00	1.00	1.00
75th	ridge	0.85	0.43	1.00	0.71	0.52
	lasso	0.96	0.86	1.00	0.93	0.90
	elastic net	0.96	0.86	1.00	0.93	0.9
	ensemble ridge	0.87	0.50	1.00	0.75	0.59
	ensemble lasso	0.98	0.93	1.00	0.96	0.95
	ensemble elastic net	1.00	1.00	1.00	1.00	1.00
90th	ridge	0.90	0.17	1.00	0.58	0.26
	lasso	0.92	0.33	1.00	0.67	0.47
	elastic net	0.94	0.50	1.00	0.75	0.64
	ensemble ridge	0.90	0.17	1.00	0.58	0.26
	ensemble lasso	0.96	0.67	1.00	0.83	0.78
	ensemble elastic net	0.98	0.83	1.00	0.92	0.90

Table 4. Comparison of different penalized regression algorithms (ridge, lasso and elastic net) and the algorithms in the proposed ensemble framework in predicting influenza epidemics, by using three cases of high incidence thresholds defined as the median, 75th and 90th percentiles of number of influenza cases over the study period. Measures including accuracy (Acc), sensitivity (Sen), specificity (Spe), area under the receiver operating characteristic curve (AUC) and kappa index of agreement (KIA) were used to assess the predictions.

In a newly released report[32], about 87% of Chinese internet users preferred Baidu to search for any information, so it is the most popular search engine in China. With the wide use of the Baidu search engine, the search volume of Baidu naturally reflects Chinese online behavior[30]. Therefore, data from Baidu are more representative of search queries in China for this analysis. Many search keywords are more likely to be captured with this search engine to build a Baidu search engine queries-based surveillance model.

The data for the surveillance model must be automatically fetched over the internet. To achieve this goal, we established an auto-crawler by using Python to collect search volume data for the keywords obtained. The auto-crawler was mainly completed by using the Selenium package within Python. The framework of the auto-crawler included calling the tool of the Selenium webdriver[33] to start with a browser and open the Baidu Index website, construct a new uniform resource locator (URL) using a keyword, call the Selenium webdriver to open the URL and take screenshots that containing the figures of search volume, and call Tesseract-OCR to extract the data (Fig. 2).

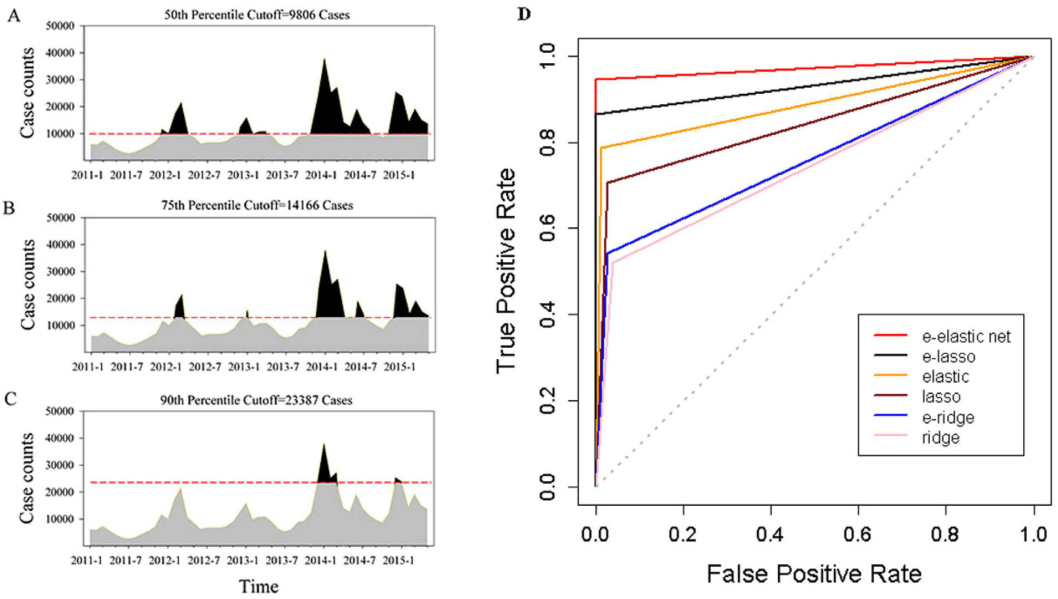

Figure 4. Performance of different penalized regression algorithms (ridge, lasso and elastic net) and the algorithms in the proposed ensemble framework in predicting influenza epidemics. (**A**) High incidence threshold defined as the median percentile of influenza case counts over the study period. (**B**) High incidence threshold defined as the 75th percentile of influenza case counts. (**C**) High incidence threshold defined as the 90th percentile of influenza case counts. (**D**) Comparison of performance of the six prediction models using the receiver operating characteristic (ROC) curve. The e-elastic net, e-lasso and e-ridge models represent the ensemble elastic net, lasso and ridge regression models, respectively.

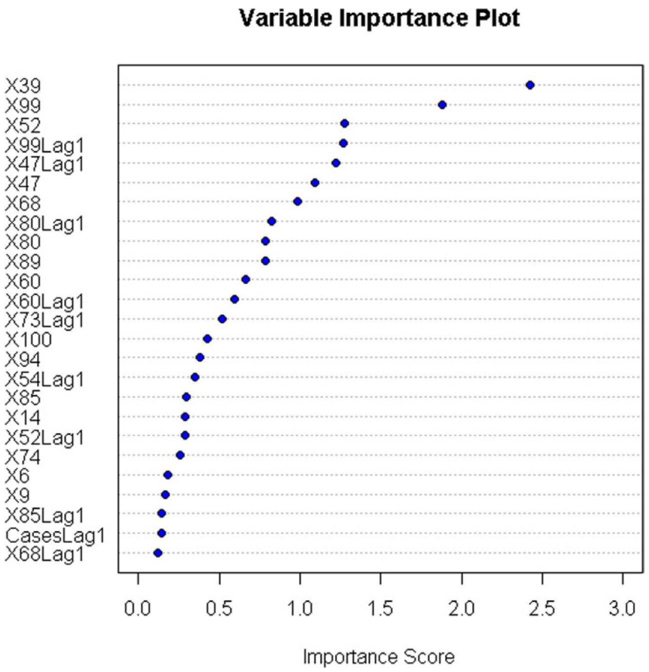

Figure 5. Contribution of each predictor to the prediction in the ensemble elastic net regression model. Only the top 25 significant predictors are shown, and their meanings are described in Table 2. For example, the variable *X39* represents the keyword "*type a flu*" and *X99* denotes the keyword "*saying type a h1n1 flu*".

For our empirical analysis, the number of search terms used for predicting influenza epidemics was greater than the sample size (117 > 53) (Table 2). Beyond the use of a linear regression model using a stepwise fashion for significant variable selection and model prediction[11], this study utilized penalized regression approaches[12] to establish prediction models with various search keywords. With a large number of predictors in the model, we would prefer to search for a smaller subset that has the strongest effects. A feature of the penalized regression models is a tuning parameter, λ, that controls the amount of shrinkage applied to the coefficients. By shrinking variables with very unstable estimates towards zero, the approach can effectively exclude some irrelevant variables and produce a subset of variables with strong effects. Regarding the tuning parameter, the traditional way of choosing the optimal λ is to use the cross-validation method. However, the robustness of variable selection is affected by the fold assignment used for cross-validation to some extent[34]. This situation results in estimating the model parameters with a degree of variability. To enhance the predictability of penalized regression models, we combined the methods of bagging and multi-objective optimization to construct the ensemble penalized regression models. Bagging can substantially improve the accuracy of an instable prediction model[22]. Our study suggested that the proposed ensemble framework significantly improved the performance of the conventional lasso, ridge and elastic net regression models, and the ensemble elastic net regression model was optimal in estimating influenza activity.

We found high correlations between specific search terms of Baidu and seasonal influenza incidence. We developed an index of importance score to estimate the contribution of each search term to the prediction of influenza epidemics. Breiman, L.[27] introduced a practical approach to measure variable importance based on computationally intensive permutations. We adopted this idea and assessed the contribution of each predictor in the ensemble model. For the performance, our predictions of time periods with high influenza incidence based on the ensemble elastic net regression model were very accurate, for different thresholds of high incidence (Table 4). Together, these results demonstrate the viability of the presented ensemble model in supporting influenza surveillance. The ensemble model performed similarly when the number of bootstrap replicates was large. The results of the empirical study indicated that the ensemble model was robust.

Although China has established a notifiable infectious disease monitoring system nationwide, reported influenza cases are available to the public with a delay of about 1 to 2 weeks. The rapid expansion of the geographical distribution and genetic diversity of novel influenza viruses poses a direct challenge to current disease control systems in China[35]. Potentially, influenza may become a long-term threat to public health in this country. Predictive search term-based models were found to perform better than a model using only reported cases to predict future cases[7,8,11]. Specifically, an internet search-term model returns results more quickly and with better performance[18]. Our study also suggested that most of the selected search keywords captured the peaks and troughs of the time series curves of influenza cases. Our ensemble elastic net regression model predicted seasonal influenza epidemics with high performance. Thus, in China, this internet search term-based system might be used as a supplement to existing surveillance systems. However, we should note that surveillance models based on internet search query data like Google Flu Trends have substantial flaws including missing the first wave of the 2009 influenza H1N1 pandemic and overestimating the intensity of the H3N2 epidemic during the 2012/2013 season in United States[36]. It means that there is room to improve the performance of surveillance models based on internet search query data and provide reliable surveillance for seasonal or pandemic influenza[36]. In addition, because Google has pulled out of mainland China since 2010, search query data from Google during the study time period of 2011–2015 are not publicly available in mainland China. Therefore, an overall comparison between the algorithm proposed in this study and that of Google Flu Trends cannot be made. All of these drive us to further validate the performance of the proposed algorithm by ongoing studies in the future.

Several limitations of this study should be mentioned. In fact, different people may use different words to search for the same information, especially when searching in Chinese, which has various ways of expression. Thus, search keywords should be carefully selected to reflect terms most likely associated with influenza epidemics. As well, internet searching behavior was susceptible to the impact of media reports, which might affect the performance of the internet search term-based system[37]. Third, in the empirical study, 100 bootstrap replicates were used for building the ensemble model. With this setting, the ensemble prediction was converged to a stable result but required much time to generate an aggregated prediction. This issue was also discussed by Breiman, L.[27]. A procedure for parallel computing integrated into the ensemble model to speed up the analysis would be practical. Hence, the computing efficiency needs to be improved.

In conclusion, this present study developed a novel ensemble elastic net penalized regression model by combining bagging and a multi-objective optimization method to monitor seasonal influenza activity. The approach provided a useful tool in support of the public health response to influenza and other infectious diseases in China.

References

1. World Health Organization. Influenza (Seasonal) http://www.who.int/mediacentre/factsheets/fs211/en/ (Date of access: 26/01/2017) (2014).
2. He, Q. *et al.* Effectiveness of seasonal influenza vaccine against clinically diagnosed influenza over 2 consecutive seasons in children in Guangzhou, China: a matched case-control study. *Human Vaccines & Immunotherapeutics* **9,** 1720–1724 (2013).
3. H, Y. *et al.* Regional variation in mortality impact of the 2009 A(H1N1) influenza pandemic in China. *Influenza & Other Respiratory Viruses* **7,** 1350–1360 (2013).
4. Freifeld, C., Mandl, K., Reis, B. & Brownstein, J. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association* **15,** 150–157 (2008).
5. Chew, C. & Eysenbach, G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *Plos One* **5,** e14118 (2010).
6. Brownstein, J. S., Freifeld, C. C. & Madoff, L. C. Digital disease detection–harnessing the Web for public health surveillance. *New England Journal of Medicine* **360,** 1656–1658 (2009).
7. Eysenbach, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA Annual Symposium Proceedings*. **244,** 244–248 (2006).
8. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457,** 1012–1014 (2009).
9. Ou, C., Deng, Z. & Yang, L. Prediction of Influenza-like Illness Using Auto-regression Model. *Chinese Journal of Health Statistics* **24,** 569–571 (2007).

10. Zhao, Y. U., Fang, Q. S., Zhou, M., Lian-Hong, L. I. & Wang, W. Surveillance of influenza in Zhejiang, 2008–2012. *Disease Surveillance* **27,** 1003–9961 (2012).
11. Yuan, Q. *et al.* Monitoring Influenza Epidemics in China with Search Query from Baidu. *Plos One* **8,** e64323–e64323 (2013).
12. Tibshirani, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society* **73,** 273–282 (2011).
13. Guo, P. *et al.* Improved Variable Selection Algorithm Using a LASSO-Type Penalty, with an Application to Assessing Hepatitis B Infection Relevant Factors in Community Residents. *Plos One* **10,** e0134151 (2015).
14. Efron, B. & Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *American Statistician* **37,** 36–48 (2012).
15. Guo, P. *et al.* Gene expression profile based classification models of psoriasis. *Genomics* **103,** 48–55 (2014).
16. Cohen, J. A coefficient of agreement of nominal scales. *Educational and Psychological Measurement* **20,** 37–46 (1960).
17. Datta, S., Pihur, V. & Datta, S. An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data. *BMC Bioinformatics* **11,** 427 (2010).
18. Althouse, B. M., Ng, Y. Y. & Cummings, D. A. T. Prediction of Dengue Incidence Using Search Query Surveillance. *Plos Neglected Tropical Diseases* **5,** e1258–e1258 (2011).
19. Zou, H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101,** 1418–1429 (2012).
20. Guo, P. *et al.* Blood lead levels and associated factors among children in Guiyu of China: a population-based study. *Plos One* **9,** e105470–e105470 (2014).
21. Pihur, V., Datta, S. & Datta, S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* **23,** 1607–1615 (2007).
22. Breiman, L. Bagging predictors. *Machine Learning* **24,** 123–140 (1996).
23. Fagin, R., Kumar, R. & Sivakumar, D. Comparing top k lists. *SIAM Journal on Discrete Mathematics* **17,** 28–36 (2003).
24. Pihur, V., Datta, S. & Datta, S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* **10,** 62 (2009).
25. Hoens, T. R. & Chawla, N. V. Generating Diverse Ensembles to Counter the Problem of Class Imbalance. *Advances in Knowledge Discovery and Data Mining* **6119,** 488–499 (2010).
26. Panov, P. & Džeroski, S. Combining Bagging and Random Subspaces to Create Better Ensembles. *Advances in Intelligent Data Analysis VII* **4723,** 118–129 (2007).
27. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).
28. Makridakis, S. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting* **9,** 527–529 (1993).
29. Kang, M., Zhong, H., He, J., Rutherford, S. & Yang, F. Using Google Trends for influenza surveillance in South China. *Plos One* **8,** e55205–e55205 (2012).
30. Gu, Y. *et al.* Early detection of an epidemic erythromelalgia outbreak using Baidu search data. *Scientific Reports* **5,** 12649 (2015).
31. Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F. & Brownstein, J. S. Influenza A (H7N9) and the importance of digital epidemiology. *New England Journal of Medicine* **369,** 401–404 (2013).
32. China Internet Network Information Center. The Chinese search engine market research report in 2013 http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/ (Date of access: 26/01/2017) (2013).
33. npm Enterprise. selenium-webdriver https://www.npmjs.com/package/selenium-webdriver (Date of access: 26/01/2017) (2016).
34. Roberts, S. & Nowak, G. Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis* **70,** 198–211 (2014).
35. Lam, T. T. *et al.* Dissemination, divergence and establishment of H7N9 influenza viruses in China. *Nature* **522,** 102–105 (2015).
36. Olson, D. R., Konty, K. J., Paladini, M., Viboud, C. & Simonsen, L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLOS Computational Biology* **9,** e1003256 (2013).
37. Valdivia, A. *et al.* Rapid communications Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks -results for 2009–10. *Euro Surveill* **15,** 2–7 (2010).

Acknowledgements

We thank the China's Center for Disease Control and Prevention for providing publicly available data of reported influenza cases. This study was supported by the Department of Education, Guangdong Government under the Top-tier University Development Scheme for Research and Control of Infectious Diseases (2015022 and 2015023). We thank Mrs Laura Smales (BioMedEditing, Toronto, Canada) for English language editing. We really thank the editor and two anonymous reviewers for their professional suggestions which greatly improve the manuscript.

Author Contributions

P.G. and Q.Y.Z. conceived and designed the study. P.G., J.J.Z., L.W., S.Y.Y., G.F.L., C.Y.D., Y.W. and Q.Y.Z. collected and cleaned the data. P.G. and Q.Y.Z. analyzed, interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Guo, P. *et al.* Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. *Sci. Rep.* **7**, 46469; doi: 10.1038/srep46469 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017