MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190360

A Regularization-Based eXtreme Gradient Boosting Approach in Foodborne Disease Trend Forecasting

Shanen Chen[a], Jian Xu[b], Lili Chen[b*], Xi Zhang[a*], Li Zhang[b], Jinfeng Li[b]

[a] *Department of Industrial Engineering and Management, Peking University, Beijing, China*
[b] *IBM Research - China, Beijing, China*

Abstract

Foodborne disease is a growing public health problem worldwide and imposes a considerable economic burden on hospitals and other healthcare costs. Thus, accurately predicting the propagation of foodborne disease is crucial in preventing foodborne disease outbreaks. Few studies have investigated the dependencies between environmental variables and foodborne disease activity. This study develops a regularization-based eXtreme gradient boosting approach for foodborne disease trend forecasting considering environmental effects to capture dependencies hidden in foodborne disease time series. A real case in Shanghai, China was studied to validate our proposed model along with comparisons to traditional and benchmark algorithms for foodborne disease prediction. Results show that the foodborne disease prediction approach we propose achieves slightly superior performance in terms of one-day-ahead prediction of foodborne disease, and presents more robust prediction for 2-7 days ahead prediction.

Keywords:

Algorithms, foodborne diseases, machine learning

Introduction

Foodborne disease encompasses a wide spectrum of illnesses and is a growing public health concern worldwide, which is the result of eating food contaminated with bacteria, viruses, and toxic or hazardous substances [1]. It is the leading cause of symptoms of nausea, vomiting, abdominal pain, diarrhea, fever, paralysis, rash, and other symptoms across all age groups, especially vulnerable groups such as children, older people and those with impaired immunity [2]. Foodborne diseases are estimated to cause 600 million - almost 1 in 10 people in the world - fall ill and 420,000 die every year [3]. Although recent molecular methods have improved the detection, identification and characterization of foodborne disease in the environment and clinical samples, widespread emergence of the viruses still presents a challenge to the prediction technique. For healthcare authorities, developing emergency preparedness plans to manage the surge in the number of patients increase during a pandemic is important if foodborne disease could be predicted in a timely and accurate manner [4]. The reliable forecast of foodborne disease allows better coordination of mitigation and intervention resources in public health system and reduces the strain on healthcare systems [5].

A vast number of quantitative approaches have been devoted to forecasting foodborne disease [6,7,8,9]. Cawley et al. proposed a sparse Bayesian kernel survival analysis model to predict the growth of a foodborne microbial pathogen, which achieved better accuracy than models based on traditional survival analysis techniques [6]. Rasam et al. applied geographical information system (GIS) and satellite remote sensing technologies in interactive mapping of foodborne disease and analyzing dynamic changes for foodborne transmission risk factor [7]. Mohammed et al. proposed an adaptive neuro-fuzzy inference system (ANFIS) to forecast the total cases of norovirus, and the algorithm showed high accuracy and remarkable prediction of norovirus cases in drinking water [8]. Hill constructed an agent-based model to provide a reasonable representation of how the norovirus spread between students among classrooms in a school [9]. Several studies have contributed to the growing body of evidence that points to the role of environmental variables in the transmission of the foodborne disease [10,11], where linkages between certain environmental factors and specific foodborne disease type have been investigated. However, within current research there has been limited assessment of the global impact of the environment on foodborne diseases, and specifically the temporal transmission of foodborne disease. Incorporating considerations for environmental factors in foodborne disease spread may address uncertainties and non-stationarity existing in foodborne disease observations.

To address this gap, we apply a regularization-based eXtreme gradient boosting approach in the creation of a prediction model that can reveal hidden patterns in foodborne disease data related to environmental factors, particularly meteorology and air quality. The contributions of the study can be divided into three folds. *First*, environmental information is fused into the model to assess their impact on foodborne disease spread. Prediction results are compared to explain the significance in including environmental variables. *Second*, L_1-regularization (least absolute shrinkage and selection operator, LASSO) [13] is employed to identify the most explainable variables and reduce the sparse effect for multiple environmental factors in the prediction model. The LASSO has two advantages over other feature selection techniques in terms of stability and computational efficiency, which are preferable in our case to select most revelent and non-redundant environmental variables for foodborne disease prediction. *Finally*, a systematic comparison between several machine learning models and our approach is conducted through a real dataset from the Shanghai Municipal Center for Disease Control and Prevention (SHCDC) to demonstrate the performance of our proposed model.

The rest of this paper is organized as follows. Methods presents the model details of the proposed regularization-based eXtreme gradient boosting approach for foodborne disease trend forecasting. Results presents a real case study, where foodborne disease time series obtained from Shanghai is applied to demonstrate the performance. Finally, the discussion provides the summary and concluding remarks.

Methods

Data Description

A national foodborne disease monitoring and surveillance system was established in 2010, covering all 31 provinces, major municipalities and autonomous regions in Mainland China. Among those covered regions, Shanghai municipality has been confronted with the challenge of reducing the burden of foodbore disease. To determine a diagnosis of foodborne illness, clinicians review history and and underlying condition for patients who report gastrointestinal symptoms, vomiting, fever, or drowsiness. In collaboration with SHCDC, the daily number of total foodborne disease cases, along with five other channels of foodborne disease such as the daily numbers of positive cases and foodborne cases conducting laboratory test, was collected from 1st-Jan-2014 to 25th-Oct-2017. During this period of 1394 days a total of 52163 cases were reported by selected survillance hospitals. The meteorology data including daily reports of sunshine, humidity, ground temperature, pressure, rainfall, air temperature, wind speed, max wind speed and wind direction from 2014 to 2017 were collected from Shanghai Meteorological Bureau. The daily concentrations of ambient air pollutants from 2014 to 2017 were also obtained from Shanghai Environmental Protection Bureau website, where average values of sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), particular matter less than 2.5 μm in diameter ($PM_{2.5}$), particulate matter less than 10 μm in diameter (PM_{10}), carbon monoxide (CO) and air quality index (AQI) from different monitoring stations were calculated. The missing values of each variable were imputed with the mean computed from the remaining records of that variable. After preprocessing, these three sources of databases were aggregated according to the recorded dates. With consideration of periodicity such as date, week, month, year, day of week, day of month and day of year, the dataset was composed of a total of 29 variables. For our study, the daily number of total foodborne disease cases was the predictive variable of all models, and meteorology data and ambient air pollutants data were the input variable.

KPSS and Kruskal-Wallis test

The stationarity and linearity analysis of the foodborne disease series were performed. *First*, the non-stationary characteristic of the time series was detected by means of KPSS (Kwiatkowski, Phillips, Schmidt, and Shin) test [14]. Then, Kruskal-Wallis test [15] was applied to confirm the existence of non-linearities, which was performed by splitting the time series into five packages of data and detecting the equality of these five distribution functions to learn whether those samples were from the same distribution.

L_1-regularization for Important Environmental Factor Selection

The prediction task can be challenging given limited data and multiple environmental factors due to overfitting and strong correlations among different environmental factors. Variable selection methods are commonly used to in such scenarios. LASSO was applied in our study to identify an appropriate variable set of environmental factors for improving the prediction accuracy of foodborne disease.

Proposed by Tibshirani, LASSO is widely employed to estimate parameters and select variables in regression analysis [12]. The LASSO estimator is a particular case of the penalized least squares regression using L_1-penalty to shrink regression coefficients toward zero. Thus the most explainable variables whose coefficients are not equal to zero are selected. Typically, we assume that $X \in N \times p$ is the input data matrix, for which the first column is time index, and the remaining $(p-1)$ columns are predictors including environmental components and other channels of foodborne diseases for continuous N days. $Y \in N \times 1$ represents daily number of total foodborne disease cases during the same continuous N days. Given the standardized predictors x_{ij} for $i = 1,2,\dots,N$ and $j = 1,2,\dots,p$, and the response values y_i, the LASSO estimator is described as

$$x_{ij}\hat{\beta}_{LASSO} = arg\,min\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|\} \quad (1)$$

LASSO is implemented with training data to select the most robust predictors using least angle regression algorithm by employing the LARS package in the R (Version 3.3.2). The following methods are also applied with R based on the corresponding R packages, such as xgboost, stats, adabag, gbm, etc.

eXtreme Gradient Boosting Approach

With the predictor subset selected by LASSO, machine learning models are applied to forecast foodborne disease cases. The gradient boosting model proposed by Friedman has been empirically proven to have high efficiency [13]. As one of popular gradient boosting models, tree boosting is a highly effective and widely used machine learning method. In this paper, a scalable end-to-end tree boosting approach, namely eXtreme gradient boosting (XGBoost) has been applied to perform foodborne disease forecasting. XGBoost is widely used by data scientists to achieve state-of-art results on difficult machine learning challenges [14].

For a given data set with N samples and p features $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}$ $(|\mathcal{D}| = N, x_i \in \mathbb{R}^p, y_i \in \mathbb{R})$, a tree ensemble model uses K additive functions (K trees) to predict the output.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \quad (2)$$

where $\mathcal{F} = \{f(x) = \omega_{q(x)}\ (q: \mathbb{R}^p \to T, \omega \in \mathbb{R}^T)$ stands for the space of regression trees (also known as CART). Here q represents the structure of each tree that maps an instance to the corresponding leaf index. T is the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and leaf weights ω. For each instance, the decision rules in the trees (given by q) are applied to classify it into the leaves and calculate the final prediction by summing up the weight in the corresponding leaves (given by ω). To learn the set of functions used in the model, we minimize the following objective in equation (3) .

$$L(\emptyset) = \sum_{i=1}^{N} loss(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad (3)$$

where the first part in equation (3) is the differentiable convex training loss that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The second term Ω is the complexity of the trees. The model complexity $\Omega(f_k)$ is defined by the following equation,

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \left|\left|\omega_j\right|\right|^2, \quad (4)$$

where T is the number of leaves in the tree, λ represents for the regularization weight, γ is the minimum loss reduction and $||\omega_j||$ stands for the score for corresponding leaves. The tree $f_k(x)$ is defined in the following equation (5),

$$f_k(x) = \omega_q(x), \omega \in \mathbb{R}^T, q: \mathbb{R}^d \to 1,2,\dots,T. \quad (5)$$

The tree ensemble model in equation (3) includes function as parameters and cannot be optimized using traditional optimization methods in Euclidean space. Instead, the model is trained in an additive manner. Formally, let $\hat{y}_i^{(t)}$ represents

the prediction of i-th training sample at the t-th iteration, we will need to add f_t to minimize the following objective.:

$$L^{(t)} = \sum_{i=1}^{N} loss\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (6)$$

The f_t that most improves the models is greedily added to the model structure. By embedding the regularization term with equation (4) and (6) it can be further approximated based on the Taylor's expansion as follows:

$$\tilde{L}^{(t)} \approx \sum_{i=1}^{N} \left[loss\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} ||w_j||^2, \quad (7)$$

where $g_i = \partial loss(y_i, \hat{y}_i^{(t-1)}) / \partial \hat{y}^{(t-1)}$, $h_i = \partial^2 loss(y_i, \hat{y}_i^{(t-1)}) / \partial \hat{y}^{(t-1)}$ are the first and second order gradient statistics on the loss function .

Define $I_j = \{i | q(x_i = j)\}$ as the instance set in leaf j. We can rewrite equation (7) as follows.

$$\tilde{L}^{(t)} \approx \sum_{j=1}^{T} [\left(\sum_{i \in I_j} g_i\right) \omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) \omega_j^2] + \gamma T. \quad (8)$$

Given a fixed decision rule $q(x)$, the optimal weight w_j^* of leaf j can be computedby setting the first-order derivatives of $\tilde{L}^{(t)}$ equal to zero, obtaining the following expression:

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \quad (9)$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. The corresponding optimal value is calculated as

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (10)$$

The optimal XGBoost classifier is obtained with the sequential optimization at each scenario.

Benchmark Algorithms for Comparative Analysis

We implemented benchmark algorithms with their standard implementations and compared them to the XGBoost.

- Linear Regression (LR): The dependent variable is considered as categorical in nature while classification is done by calculating the error [16].
- Support Vector Regression with Linear Kernel (SVR): Attempts to minimize the generalization error bound so as to achieve generalized performance. SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear kernel [17].
- Bagging (Treebag): A method for generating multiple versions of a predictor to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome [18].
- Gradient Boosting Model (GBM): Combines the results of different prediction models by iteratively learning from the losses which occurr at the previous step through calculation of negative gradient [13].

Evaluation

The two indices, root mean square error (RMSE) and mean absolute percentage error (MAPE), are selected for evaluation of the errors. The RMSE is a frequently used measure of the differences between values predicted by a model and the values observed, which is the square root of the average of squared errors and scale-dependent. The MAPE is also a popular measure of prediction accuracy of a prediction method in statistics. The absolute value in the calculation is summed for every forecasted point in time and divided by the number of fitted points. Compared to the RMSE, the MAPE is scale-independent. The formulas for calculation are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \quad (11)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i}, \quad (12)$$

where N is the number of data samples, $y_i (i = 1,2, \dots, N)$ is the observed number of foodborne cases and $\hat{y}_i (i = 1,2, \dots, N)$ is the predicted daily number of total foodborne disease cases.

Results

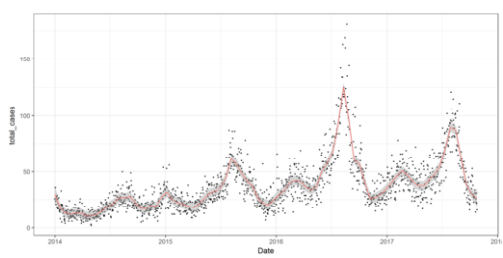
The descriptive statistics summary of the daily number of foodborne disease cases and environment factors are given in Table 1.

Table 1– Decriptive Statistics of the Data (Daily Based)

Variable	Mean	SD	Min	Max
Total foodborne cases	37.30	24.07	2.00	181.00
Sunshine(h)	4.49	4.00	0.00	12.70
Relative humidity(%)	73.17	12.56	35.00	98.00
Ground temperature(℃)	18.66	9.25	-2.70	40.20
Pressure (Pa)	1015.90	8.88	992.80	1039.70
Rainfall (mm)	3.98	12.71	0.00	155.40
Air temperature(℃)	17.71	8.55	-6.10	34.80
Wind speed(m/s)	2.55	0.92	0.50	7.40
Max wind speed(m/s)	4.97	1.22	2.50	10.40
$PM_{2.5}$($\mu g/m^3$)	47.31	30.68	5.00	216.00
PM_{10}($\mu g/m^3$)	66.39	37.06	8.00	256.00
SO_2($\mu g/m^3$)	15.35	8.55	6.00	75.00
CO($\mu g/m^3$)	0.80	0.27	0.40	2.20
NO_2($\mu g/m^3$)	43.38	19.38	5.00	143.00
O_3($\mu g/m^3$)	106.50	44.54	11.00	286.00
AQI	83.14	37.89	23.00	266.00

In time series plot of the daily number of total foodborne disease cases (Figure 1), a moderate peak is observed in 2015, with two sharp peaks followed in 2016 and 2017 respectively. Before 2015, steady growth in the daily number of total foodborne cases was recognized, which demonstrated the nonlinearity existing in foodborne disease occurrence.

Figure 1– Time Series Plot of Daily Foodborne Disease Cases

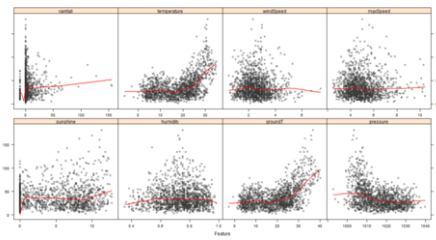


Correlation, Stationarity and Linearity Analysis Between Disease and Environment

The Pearson correlation analysis was implemented to explore the relationship between external factors and foodborne disease cases. Daily numbers of the positive cases, the laboratory testing cases and the norovirus testing cases, as well as the daily percentage of total positive cases were found to be strongly correlated with the daily number of total foodborne disease cases, with positive correlation values of 0.71, 0.65, 0.65 and 0.44 respectively. The qualitative relationships between daily foodborne disease cases and weather conditions were exploited with paired scatter plots as shown in Figure 2. For each subfigure in Figure 2, the x axis denotes the corresponding environmental factor, while the y axis represents the daily number of total foodborne disease

cases. As seen from the figure, weather conditions such as ground temperature, air temperature and pressure were highly

Figure 2- Scatter Plot Matrix Between Foodborne Disease and Environmental Factors



correlated with the daily number of total foodborne disease cases. These findings are further verified through quantitative analysis of Pearson correlation calculation, where the correlations between ground temperature, air temperature, pressure and daily number of total foodborne disease cases reached 0.48, 0.47 and -0.38 respectively. Weak correlations were also observed between air quality conditions and daily number of total foodborne disease cases. Except for O_3, SO_2, NO_2, $PM_{2.5}$, PM_{10}, CO and AQI were negatively correlated with the daily number of total foodborne disease cases, with a maximum absolute value of 0.29 found at NO_2.

The null hypothesis that the foodborne disease series was level (or trend) stationary was rejected by KPSS test results, with p value smaller than 0.01 (p_{trend}<0.01, p_{level}<0.01). The Kruskal-Wallis test also depicited that nonlinearity existed in foodborne disease occurrence with an extreme small p value (p=9.43 x 10^{-144}).

Trend Prediction with Regularization Based XGBoost

The collected foodborne disease cases and environmental datasets were employed to validate the effectiveness of regularization based XGBoost methods. The entire dataset was divided into two partitions, with observations from 1st , Jan, 2014 to 19, Jan, 2017 (1116 samples) for training and observations from 20, Jan, 2017 to 25, Oct, 2017 (279 samples) for testing. The seasonal ARIMA (SARIMA) models were first implemented to encompass the potential critical predictors for instance and the historical daily number of total foodborne disease cases, and then applied as a benchmark model without considering any environmental factors. The parameters in $SARIMA(p,d,q)(P,D,Q)_s$, which represented seasonal factors (s), autoregressive (AR) order (p), degree of differentiating (d), moving average (MA) order (q), seasonal AR order (P), seasonal differential order (D) and seasonal MA order (Q), were selected using both partial autocorrelation and iterative searching approach. Through the comparisons of RMSE and MAPE on the validation dataset (the last 20% of training samples in training dataset), the prediction results showed that $SARIMA(6,0,0)(3,0,0)_7$ achieved best generalization ability with a minimum MAPE of 18.04% on validation dataset compared with other parameter combinations. The selected model ($SARIMA(6,0,0)(3,0,0)_7$) was employed to predict the daily number of total foodborne disease occurrence in 1-7 days' advance with test dataset. The predictive performance steadily decreased when the forecasting horizon increased due to uncertainty and complexity in future foodborne disease occurrence. The best prediction was achieved at 1-day ahead prediction, with a minimum RMSE and MAPE of 11.72 and 21.34% respectively. To illustrate the effects of environmental factors and other monitoring channels of foodborne disease in foodborne disease forecasting, XGBoost method was applied to predict the cases. Before model building, the LASSO was firstly adopted to derivative of 27 external variables for selecting the most explainable features. For 1-day ahead prediction, 12 of 27 variables including daily numbers of other foodborne diseases (positive cases, laboratory testing cases, norovirus positive cases), daily percentage of total positive cases, air temperature, sunshine, humidity, wind speed, maximum wind speed, CO, year, and day of week were found to be critical in improving the performance of XGBoost. With proposed regularization based XGBoost method, we were able to obtain a RMSE of 11.71 and a MAPE of 21.15% on the same test dataset with SARIMA model, slightly improving the predictive performance.

Discussion

We further compared the daily foodborne disease forecasting performances obtained from other machine learning models including linear regression (LR), support vector regression with linear kernel (SVR), bagging (treebag), gradient boosting model (gbm), and XGBoost model. Table 2 shows the forecasting performances obtained on test dataset.

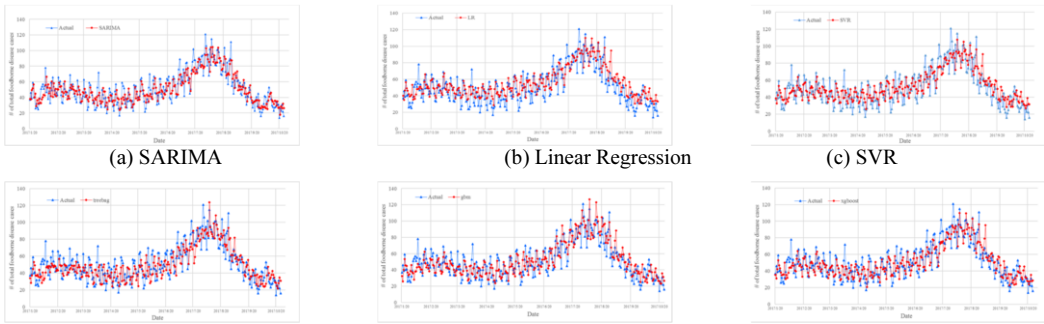
Table 2– Performance of Mutiple Models on Test Dataset

Model	Horizon	RMSE	MAPE (%)	Model	Horizon	RMSE	MAPE (%)
SARIMA	1	11.72	21.34	treebag	1	13.07	23.32
	2	12.49	22.44		2	14.72	25.02
	3	13.37	24.06		3	14.18	24.44
	4	13.98	24.42		4	14.79	25.40
	5	14.89	24.92		5	15.30	27.03
	6	15.79	26.24		6	14.95	26.21
	7	16.88	27.70		7	15.57	26.41
LR	1	12.29	25.43	gbm	1	11.85	21.09
	2	14.06	29.47		2	13.47	24.23
	3	13.61	28.16		3	12.98	23.83
	4	15.49	31.54		4	13.84	26.04
	5	14.87	30.53		5	15.45	27.44
	6	15.35	31.60		6	15.89	28.67
	7	15.78	33.80		7	16.17	27.63
SVR	1	11.65	22.78	XGBoost	1	11.71	21.15
	2	13.21	26.58		2	13.21	24.14
	3	12.88	25.89		3	12.85	25.30
	4	13.98	27.99		4	13.98	27.14
	5	14.16	28.41		5	14.21	26.60
	6	14.31	29.17		6	14.14	26.25
	7	14.51	30.77		7	13.59	25.51

Varying with the forecasting horizon, both SARIMA and XGBoost achieved a stable and robust predictive precision in terms of RMSE and MAPE. However, when the forecasting horizon increased, the predictive errors of XGBoost remained steady, but had a rising trend with SARIMA model. The performances of 1-day ahead prediction on test dataset are further provided in Figure 3 as an example, of which x axis denotes date and y axis is the daily number of foodborne disease cases. From these performances we suggest when the forecasting horizon is within 4 days, SARIMA model can be deployed for real application. When the forecasting horizon is greater than 4, regularization based XGBoost model is recommended.

Conclusion

This paper presents a regularization based model based on eXtreme gradient boosting approach for foodborne disease trend forecasting. Environmental factors including weather conditions and air quality conditions were incorporated into the prediction model and the LASSO technique was applied to shrink the predictors and identify the critical factors that would affect foodborne disease progagation. The real case demonstrated a satisfactory and robust performance in terms of 5-7 day ahead prediction, with a lowest MAPE of 25.52% for 7-day ahead forecasting. However, the model does not significantly improve the predictive performance within 4 days of environmental factors and foodborne disease.



(a) SARIMA (b) Linear Regression (c) SVR

(d) Treebag (e) GBM (f) XGBoost method

Figure 3– Foodborne Disease Prediction with Different Machine Learning Techniques

Future works still need to be done to investigate relationship between specific type of foodborne disease (e.g norovirus foodborne disease) and environmental factors. Moreover, spatial factor should also be modeled for accelerating the investigation of foodborne disease outbreak.

Acknowledgements

The authors would like to thank National Meteorological Information Center and who provided open, online access to weather and meteorological data (https://www.aqistudy.cn/)

References

[1] R.V. Tauxe, M.P. Doyle, T. Kuchenmüller, J. Schlundt, and C.E. Stein, Evolving public health approaches to the global challenge of foodborne infections, *Int J Food Microbiol* **139** (2010), S16-S28.

[2] National Institute of Diabetes and Digestive and Kidney Diseases, Foodborne Illnesses, 2014, Retrieved from https://www.niddk.nih.gov/health-information/digestive-diseases/foodborne-illnesses on 30 October 2018.

[3] WHO News, Food Safety, 2017, Retrieved from http://www.who.int/news-room/fact-sheets/detail/food-safety on 30 October 2018.

[4] D.G. Newell, M. Koopmans, L. Verhoef, E. Duizer, A. Aidara-Kane, H. Sprong, M. Opsteegh, M. Langelaar, J. Threfall, F. Scheutz, J.v. der Giessen, and H. Kruse, Food-borne diseases - The challenges of 20 years ago still persist while new ones continue to emerge, *Int J Food Microbiol* **139** (2010), S3-S15.

[5] M. Thakur, S. Olafsson, J.-S. Lee, and C.R. Hurburgh, Data mining for recognizing patterns in foodborne disease outbreaks, *J Food Eng* **97** (2010), 213-227.

[6 G.C. Cawley, N.L.C. Talbot, G.J. Janacek, and M.W. Peck, Sparse bayesian kernel survival analysis for modeling the growth domain of microbial pathogens, *IEEE Trans Neural Netw* **17** (2006), 471-481.

[7] A.R. Abdul Rasam and A.M. Mohd Noor, Contribution of GIS and remote sensing technologies for managing foodborne diseases in Malaysia, *Control and System Graduate Research Colloquium. IEEE* (2012), 258-261.

[8] H. Mohammed, I.A. Hameed, and R. Seidu, Adaptive neuro-fuzzy inference system for predicting norovirus in drinking water supply, *International Conference on Informatics, Health & Technoogy. IEEE* (*2017)*, 1-6.

[9] A.L. Hill, Norovirus outbreaks: Using agent-based modeling to evaluate school policies, *Winter Simulation Conference. IEEE* (2016),1265-1276.

[10] M.S. Park, K.H. Park, and G.J. Bahk, Combined influence of multiple climatic factors on the incidence of bacterial foodborne diseases, *Sci Total Environ* **610** (2017), 10-16.

[11] R.S. Kovats, S.J. Edwards, S. Hajat, B.G. Armstrong, K.L. Ebi, and B. Menne, The effect of temperature on food poisoning: a time-series analysis of salmonellosis in ten European countries, *Epidemiol Infect* **132** (2004), 443-453.

[12] Tibshirani, Robert, Regression shrinkage and selection via the lasso: A retrospective, *J R Stat Soc Ser B-Stat Methodol* **73** (2011), 273-282.

[13] J.H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann Stat* **29** (2001),1189-1232.

[14] D. Kwiatkowski, P.C.B. Phillips, P. Schmidt, and Y. Shin, Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root, *J Econometrics* **54** (1992), 159-178.

[15] W.H. Kruskal and W.A. Wallis, Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47** (1952), 583-621.

[16] K. Farid, Applied Linear Regression Models, *Technometrics*,**32**(2004), 352-353.

[17] A.J. Smola and B. Schölkopf, A tutorial on support vector regression, *Stat comput* **14** (2004), 199-222.

[18] L. Breiman, Bagging predictors, *Mach Learn* **24** (1996), 123-140.

Address for correspondence

Lili Chen, chenlili1002@126.com;
Xi Zhang, xi.zhang@coe.pku.edu.cn.