Applied Soft Computing 35 (2015) 280–290

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Artificial neural networks for infectious diarrhea prediction using meteorological factors in Shanghai (China)

Yongming Wang[a,*], Jian Li[b], Junzhong Gu[a], Zili Zhou[c], Zhijin Wang[a]

[a] Department of Computer Science & Technology, East China Normal University, Shanghai 200241, China
[b] Department of Acute Infectious Disease, Shanghai Municipal Center for Disease Control & Prevention, Shanghai 200336, China
[c] School of Physics and Engineering, Qufu Normal University, Qufu, Shandong 273165, China

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form
23 December 2014
Accepted 17 May 2015
Available online 3 July 2015

Keywords:
Artificial neural networks
Forecasting model
Infectious diarrhea
Meteorological factors
Sensitivity analysis

ABSTRACT

Infectious diarrhea is an important public health problem around the world. Meteorological factors have been strongly linked to the incidence of infectious diarrhea. Therefore, accurately forecast the number of infectious diarrhea under the effect of meteorological factors is critical to control efforts. In recent decades, development of artificial neural network (ANN) models, as predictors for infectious diseases, have created a great change in infectious disease predictions. In this paper, a three layered feed-forward back-propagation ANN (BPNN) model trained by Levenberg–Marquardt algorithm was developed to predict the weekly number of infectious diarrhea by using meteorological factors as input variable. The meteorological factors were chosen based on the strongly relativity with infectious diarrhea. Also, as a comparison study, the support vector regression (SVR), random forests regression (RFR) and multivariate linear regression (MLR) also were applied as prediction models using the same dataset in addition to BPNN model. The 5-fold cross validation technique was used to avoid the problem of overfitting in models training period. Further, since one of the drawbacks of ANN models is the interpretation of the final model in terms of the relative importance of input variables, a sensitivity analysis is performed to determine the parametric influence on the model outputs. The simulation results obtained from the BPNN confirms the feasibility of this model in terms of applicability and shows better agreement with the actual data, compared to those from the SVR, RFR and MLR models. The BPNN model, described in this paper, is an efficient quantitative tool to evaluate and predict the infectious diarrhea using meteorological factors.

1. Introduction

Infectious diarrhea which can be caused by a variety of bacterial, viral and parasitic organisms remains a major public health problem around the world [1]. As a kind of common and important infectious disease, infectious diarrhea has a serious threat to human health and leads to one billion disease episodes and 1.8 million deaths each year. Infectious diarrhea in young children is a killer illness, especially in developing countries [2,3]. In China, infectious diarrhea is a notified infectious disease which is the biggest developing country. In Shanghai of China, the incidence of infectious diarrhea has significant seasonality throughout the year and is particularly high in the summer and autumn of recent years. Moreover the incidence of infectious diarrhea has been the highest among the class A and B intestinal infectious diseases. Infectious diarrhea has become the focus of the prevention and control of infectious diseases over the years. Health forecasting is a novel area of forecasting, and a valuable tool for predicting future health events or situations such as demands for health services and healthcare needs. It facilitates preventive medicine and health care intervention strategies, by pre-informing health service providers to take appropriate mitigating actions to minimize risks and manage demand [4]. Hence, a robust short-term (week-ahead) prediction model for infectious diarrhea incidence is necessary for decision-making in policy and public health.

Infectious diseases have a closely relation with meteorological factors [5] and can affect infectious diseases in a linear or nonlinear fashion [6]. Over the past couple of decades, there has been a large scientific and public debate on climate change and its direct as well as indirect effects on human health [7]. The effects of meteorological factors, such as temperature, rainfall and relative humidity, on diarrhea diseases incidence have got much more concerning recently. As far as we are concerned with the prediction of diarrhea diseases in literature, many forecasting models

* Corresponding author. Tel.: +86 13482813260.
E-mail address: ymwang819@gmail.com (Y. Wang).

http://dx.doi.org/10.1016/j.asoc.2015.05.047
1568-4946/

based on statistical methods for diarrhea diseases forecasting have been devoted. Alexander et al. [8] evaluated monthly reports of diarrheal disease among patients presenting to Botswana health facilities and compared this to climatic variables (rainfall, temperature, and vapor pressure). Kolstad et al. [9] combined a range of linear regression coefficients to compute projections of future climate change (temperature)-induced increases in diarrhea. Zhao et al. [10] established multiple regression model rolling forecast of daily incidence of infectious diarrhea. Chou et al. [11] applied a climate variation-guided Poisson regression model to predict the dynamics of diarrhea-associated morbidity. The results indicated that the maximum temperature and extreme rainfall days were strongly related to diarrhea-associated morbidity. In Hashizume et al. [12], weekly rainfall, temperature and number of hospital visits for non-cholera diarrhea were analyzed by a Poisson regression model. McCormick et al. [13] studied of temporal and spatial patterns of diarrheal disease and construct a spatial panel regression model suing contemporary acute diarrhea disease and climatic data (temperatures and precipitation) and found there is a strong association between daily mean temperature and precipitation and the incidence of hospitalization. Singh et al. [14] examined diarrhea notifications in Fiji in relation to estimates of temperature and rainfall, using Poisson regression analysis of monthly data for 1978–1998. Result indicated that there were positive associations between diarrhea reports and temperature and between diarrhea reports and extremes of rainfall. Lloyd et al. [15] undertook a global cross-sectional study of diarrhea incidence in children under 5, and assessed the association with climate variables (temperature and rainfall) by linear regression method.

With regard to the fact that number of meteorological factor that effect infectious diarrhea are too much and the inter-relation among them is also very complicated, prediction models based on traditional statistics methods may not be fully suitable for such type of problems. In recent years, artificial intelligence techniques such as artificial neural network (ANN) and support vector machine (SVM) have been employed for developing predictive models in complex prediction problems. Nowadays, ANNs are considered to be one of the intelligent tools to understand the complex problems. As a powerful computational method, ANNs have been widely used in the medical and health field, such as medical diagnosis and disease prediction [16–23]. Moreover, ANN-based hybrid prediction model (hybrid neural networks) have been widely used in different fields and obtained the very good prediction result [24–27]. Compared with other forecast methods, ANNs are advantageous in terms of high data error tolerance, easy adaptability to online measurements.

The feed forward back propagation neural network (BPNN) as a typical ANNs is essentially a mapping function from input to output vector(s) without knowing the correlation between the data. BPNN can implement any complex nonlinear mapping function proved by mathematical theories, and approximate an arbitrary nonlinear function with satisfactory accuracy [28]. However BPNN are characterized by very poor convergence. Several improvements for BPNN were developed. The Levenberg–Marquardt back propagation neural network is a powerful optimization technique that was introduced to the neural net research because it provided methods to accelerate the training and convergence of the algorithm [29].

Based on a literature review, so far as I know to the best knowledge of the authors, there is no works has been carried out to utilize the ANNs method in predicting diarrhea disease, to say nothing of infectious diarrhea. In this paper, an attempt had been made to establish a new BPNN model to predict the weekly number of infectious diarrhea (*WNID*) in Shanghai of China with a set of meteorological factors as input variable. Also, as a comparison SVR, RFR and MLR models were developed for the same purpose in addition to BPNN model. Finally, since one of the drawbacks of ANNs is the interpretation of the final model in terms of the importance of variables, the calculation is performed using sensitivity analysis.

Table 1
Summary of population in Shanghai [30].

Year	Population (million)	Population growth rate (%)
2004	13.52	
2005	13.60	0.582
2006	13.68	0.575
2007	13.79	0.788
2008	13.91	0.883

The rest of this paper is organized as follow. Study area and dataset that is used in this study are briefly described in Section 2. The prediction methods and performance evaluation criteria which are used in this paper are introduced in Section 3. These methods are ANN SVR, RFR and MLR. ANN and SVR, RFR, MLR models are developed in Sections 4 and 5, respectively. In order to investigate the performance of the established model, the prediction results of the ANN model are reported in comparison with the SVR, RFR and MLR models as discussed in Section 6. Section 7 illustrated the sensitivity analysis, and conclusions of this paper are concluded in Section 8.

2. Study area and diarrhea-meteorological data

In this section, we present the study area in Section 2.1. We also describe the dependent and independent variables used in this paper in Section 2.2.

2.1. Study area

Shanghai is located in the eastern part of China which is the largest developing country in the world, and the city has a mild subtropical climate with four distinct seasons and abundant rainfalls. It is the most populous city in China comprising urban/suburban districts and counties, with a total area of 6340.5 km^2 and had a population of 13.9 million by the end of 2008 (Table 1) [30]. Generally, in epidemiological studies, the incidence is an important index that has become the focus for many researchers. The incidence is calculated by dividing the disease number with the total population. Since the study population was relatively stationary during the time period from 2005 to 2008 with the annual growth rate below 1% (Table 1), the trend of incidence during that time period could be similarly prescribed by the trend of disease cases number. Hence we used the number of infectious diarrhea as the response variable in our models. Description of the input (meteorological factors) and output (the weekly number of infectious diarrhea) parameters for constructing the BPNN model have been given in Table 2.

2.2. Dependent and independent variables

The *WNID* cases data from 2005.1.3 to 2009.1.4 were collected from National Disease Supervision Information Management System, which is a modifiable diseases reporting system of real-time, online, based-on-case information. The infectious diarrhea cases were all clinical or laboratory-confirmed cases of infectious diarrhea and reported by hospital diagnostic. In choosing meteorological factors, an important consideration is of course whether a factor is likely to have a significant influence on the infectious diarrhea. The Spearman Rank Correlation Analysis (ARCA) was carried to determine the effects of nine meteorological factors on infectious diarrhea. According to the ARCA results (see Table 3), T_{max}, T_{min}, T_{avg} and AP_{avg} are statistically significant correlation ($p < 0.01$) and RH_{min}, RH_{avg}, SD, WS_{avg} and R are statistically correlation ($p < 0.05$) with infectious diarrhea. This implies that it is entirely feasible to