Knowledge-Based Systems 88 (2015) 12–23

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion

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

[a] Department of Computer Science and Technology, East China Normal University, Shanghai 200241, PR China
[b] College of Physics and Engineering, Qufu Normal University, Qufu 273165, PR China

ARTICLE INFO

Article history:
Received 24 December 2014
Received in revised form 2 August 2015
Accepted 4 August 2015
Available online 25 August 2015

Keywords:
Diarrhoea outpatient visits
Prediction model
Time series decomposition
Ensemble empirical mode decomposition
Generalized regression neural networks
Multi-predictor fusion

ABSTRACT

Accurate and reliable prediction of diarrhoea outpatient visits is necessary for the health authorities to ensure the appropriate action for the control of the outbreak. In this study, a novel method based on time series decomposition and multi-local predictor fusion has been proposed to predict the diarrhoea outpatient visits. For time series decomposition, the Ensemble Empirical Mode Decomposition with Adaptive Noise (EEMDAN) is used to decompose diarrhoea outpatient visits time series into a finite set of Intrinsic Mode Function (IMF) components and a residue. The IMF components and residue are modeled and predicted respectively by means of Generalized Regression Neural Network (GRNN) as local predictor. Then the prediction results of all components are fusioned using another independent GRNN as fusion predictor to obtain final prediction results. This is the first study on using a EEMDAN and GRNN to constructing an prediction model for diarrhoea outpatient visits prediction problems. The pre-procession and post-processing techniques are used to take into account the seasonal and trend effects in the datasets for improving the prediction precision of proposed model. The performance of the proposed EEMDAN–GRNN model has been compared with Seasonal Auto-Regressive Moving Average (SARIMA), Single GRNN, Wavelet-GRNN and also with EEMD–GRNN by applying them to predict four real world diarrhoea outpatient visits. The results indicate that the proposed EEMDAN–GRNN model provides more accurate prediction results compared to the other traditional techniques. Thus EEMDAN–GRNN can be an alternate tool to facilitate the prediction of diarrhoea outpatient visits.

1. Introduction

Diarrhoea is the passage of three or more loose or liquid stools per day, or more frequently than is normal for the individual [1]. Diarrhoeal caused by a variety of bacterial, viral and parasitic organisms is usually a symptom of gastrointestinal infection. Infection is spread through contaminated food or drinking-water, or from person to person as a result of poor hygiene [1]. Diarrhoeal disease is a serious threat to the health and well-being of the citizens of the world, especially in the developing countries like China, African countries and India. Thousands of people, in particular for children less than five years old, suffer from this disease every year. In spite of many studies on the diarrhoea still there were nearly 1.7 billion cases of diarrhoeal disease every year [1]. Diarrhoeal disease is the second leading cause of death in children under five years old, and is responsible for killing around 760,000 children every year. Government authorities are incurring huge cost to control and eliminate the outbreaks of diarrhoea. Thus, accurately and timely predict diarrhoea outpatient visits in advance outbreaks is an important issue for various national governments and international organizations. 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 [2].

However, many problems have occurred in prevention and control of diarrhoeal disease. One of the main problems associated with prediction of diarrhoeal disease is the complexity and diversity of influence factors that affect the diarrhoeal incidences, such as malnutrition, meteorological, living surroundings, and living habits. Especially due to global warming, rapid climate changes are occurring which result in the increase of diarrhoeal disease incidence depending upon the specific micro-climate of that particular region [56–58]. The complexity and diversity of the influence factors make it is great challenge for the researchers to predict the diarrhoeal disease outbreaks in advance. In the absence of knowledge about probabilistic attack of these diseases and exogenous factors are often limited by the availability of data,

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

http://dx.doi.org/10.1016/j.knosys.2015.08.013
0950-7051/

government fails to provide adequate treatment facility on time. Thus, it is necessary to forecast the occurrence of these diseases in advance so that its devastating impact on the society can be reduced.

There have been wide attempts to capture the relationship between the available information using some straightforward linear regression models like Auto-Regressive Integrated Moving Average (ARIMA) [3–10]. Such traditional techniques do require minimum computational efforts to set up prediction models which are considered to be an advantage. However, with non-linear nature of diarrhoeal disease it becomes difficult to use these techniques. In recent years, many recent studies focus on the use of machine learning techniques, such as Artificial Neural Networks (ANNs), to build a prediction model for time series prediction problems. Unlike traditional statistical models, ANNs are data-driven and non-parametric models. They do not require strong model assumptions and can map any nonlinear function without a priori assumption about the properties of the data, even though the underlying relationships are unknown or hard to describe. Related works have shown that machine learning techniques outperform many traditional models [11–15].

Due to the nonlinear and non-stationary characteristic, accurate prediction diarrhoea outpatient visits by the building a single global predictor is often not possible. One of the best ways to solve these problems is using decomposition-and-ensemble principle. Firstly, a complex problem is decomposed into a set of sub-problems according to the inherent class relations among training data, then gives a local predictor to learn each of the sub-problems, finally, combination the multi-local predictors into a solution to the original problem [16–19]. In the area of time series forecasting (TSF), decomposition-and-ensemble principle has proven to be a method superior to single global predictor. Many recent studies in the different area have been shown to perform better than single models [20–28]. However, so far as I know, this is the first study on using a decomposition-and-ensemble principle to constructing a prediction model for diarrhoea outpatient visits prediction problems in health forecasting.

In order to decompose a complex time series prediction task into several relative simple subtasks, time series decomposition methodologies have been widely used in different studies. These techniques can divide the data into local characteristic time scale and extracting meaningful features embedded implicitly in data. The Empirical Mode Decomposition (EMD) [34] and its improved version named Ensemble Empirical Mode Decomposition (EEMD) [35], have been widely used as a promising alternative for nonlinear and non-stationary time series and successfully applied to different areas [22–24,29–33]. In this paper, a variation of the EEMD algorithm, called Ensemble Empirical Mode decomposition with Adaptive Noise (EEMDAN) [36] is used for the time series decomposition. The EEMDAN provides an exact reconstruction of the original signal and a better spectral separation of the modes, with a lower computational cost.

Among ANN models, multilayer perceptron (MLP) trained by the standard Back-Propagation (BP) learning algorithm is popular ANNs for predicting time series. Despite the advantages of MLP, they have several weaknesses (e.g., a large number of design parameters, long training time, and suffering from local minima) which make modeling more difficult [37]. In this paper, we attempt to develop local predictors using generalized regression neural networks (GRNN), a special type of ANNs. GRNN has only a single design parameter and is simple and fast in training. Our effort in this paper focuses on designing a modeling scheme to take full advantage of EEMDAN and GRNN properties for diarrhoea outpatient visits prediction. Therefore, a novel prediction algorithm call EEMDAN–GRNN is proposed. To increase the accuracy of the prediction, we perform data pre-processing techniques such as data transformation, detrending and deseasonalizing. In order to improve the robustness and error tolerance of proposed model, a trainable fusion method (another independent GRNN predictor) is used to fusion the prediction results of multi-local predictors. The strength of the proposed prediction method is tested on four real word monthly diarrhoea outpatient visits time series datasets from three different geographical location of China.

In summary, the primary innovation and contributions of our study can be outlined as follows:

(1) Based on a literature review, there is no works has been carried out to utilize the decomposition-and-ensemble principle method in predicting diarrhoea outpatient visits. In this study, following the decomposition-and-ensemble principle, a novel model based on time series decomposition and multi-local predictor fusion has been proposed to predict the diarrhoea outpatient visits.
(2) A variation of the EEMD algorithm, called EEMDAN is used for the diarrhoea outpatient visits time series decomposition. The GRNN is used as local predictor. So far, EEMDAN and GRNN have not been used in this direction. The proposed EEMDAN–GRNN algorithm adequately makes use of the advantages of the EEMDAN decomposition method and GRNN and integrates them well, which conduce to boosting the model prediction ability and enhancing prediction efficiency.
(3) Our proposed EEMDAN–GRNN algorithm uses a dynamic nonlinear weighted scheme to fusion the multi-local predictor into a single predictor. Each local GRNN predictor independently predicts the output. A fusion predictor is then trained to predict the final output from the outputs of local predictor. Consequently, fusion predictor can capture interactions among local predictors.

The remainder of this paper is organized as follow. The methodologies that are used in this study are briefly described in Section 2. The proposed EEMDAN–GRNN modeling framework is presented in detail in Section 3. Section 4 illustrates the experimental design and methodologies implementation in details. Following that, in Section 5, the experimental results obtained from four real diarrhoea outpatient visits datasets are presented and discussed. Finally, the study is concluded in Section 6.

2. Methodologies

Before starting to present the proposed method, it is necessary to describe the theory of the acquired methodologies in the proposed approach. In this section, the decomposition technique of EEMDAN and the theory of GRNN algorithm are briefly introduced.

2.1. EEMD with adaptive noise

The EMD [34] is an adaptive signal processing technique introduced to analyze non-linear and non-stationary time series. It consists in a local and fully data-driven separation of a time series in fast and slow oscillations. However, EMD experiences some problems, such as the presence of oscillations of very disparate amplitude in a mode, or the presence of very similar oscillations in different modes, named as mode mixing [36]. To overcome these problems, the EEMD method was proposed [35]. It performs the EMD over an ensemble of the signal plus Gaussian white noise. The addition of white Gaussian noise solves the mode mixing problem by populating the whole time–frequency space to take advantage of the dyadic filter bank behavior of the EMD [38]. However it creates some new ones. Indeed, the reconstructed signal includes