International Journal of Academic Research in Accounting, Finance and Management Sciences
Vol. 9, No.4, October 2019, pp. 241–252
E-ISSN: 2225-8329, P-ISSN: 2308-0337
© 2019 HRMARS
www.hrmars.com

To cite this article: Huang, D. H., Tsai, C. H., Chueh, H. E., Wei, L.. Y. (2019). *A Hybrid Model Based on EMD-Feature Selection and Random Forest Method for Medical Data Forecasting,* International Journal of Academic Research in Accounting, Finance and Management Sciences 9 (4): 241-252

www.hrmars.com
ISSN: 2225-8329

http://dx.doi.org/10.6007/IJARAFMS/v9-i4/6841 (DOI: 10.6007/IJARAFMS/v9-i4/6841)

A Hybrid Model Based on EMD-Feature Selection and Random Forest Method for Medical Data Forecasting

Duen-Huang Huang[1], Chih-Hung Tsai[2], Hao-En Chueh[3], Liang-Ying Wei[4]

[1,2,4] Department of Information Management, Yuanpei University of Medical Technology, Hsin-Chu, Taiwan
E-mail: 4lywei@yuntech.edu.tw (Corresponding author)
[1]Department of Information Management, Hsuan Chuang University, Hsin-Chu Taiwan
[3]Department of Information Management, Chung Yuan Christian University, Tao-yuan Taiwan

Abstract

Hospital managers need to allocate emergency department (ED) resources efficiently because of the gradual aging of the population, emergency services overcrowding problem is arising. Forecasting is a vital activity that instructs decision-makers in related research fields, such as industrial scientific planning, economics, and healthcare. Scientists have applied time series methods to daily patient number forecasting at ED. Traditional time series models usually use a single variable for forecasting, but noises caused by weather conditions change and environmental factors would be included in raw data. Low forecasting performance would be generated because of using complicated raw data in time series models. Further, traditional time series models cannot be utilized in all datasets because statistics models need to meet statistical assumptions. Multi-attribute data will usually produce high-dimensional data and increase the computational complexity in the data mining procedure. For overcoming these drawbacks above, this study proposes a hybrid random-forest model based on AR (autoregressive) and empirical mode decomposition (EMD). The proposed model utilizes EMD to decompose complicated raw data into correlations frequency components and uses the feature section method to reduce high-dimensional input data generated by EMD. Then, this study combines random forest method that can surmount the limitations of statistical methods (data need to obey some mathematical distribution) to forecast daily patient volumes. To verification, daily patient volumes in an emergency are collected as experimental datasets to evaluate the proposed model. Experimental results illustrate that the proposed model surpasses the listing models.

Key words

Random Forest, Empirical Mode Decomposition (Emd), Feature Selection

Received: 25 Jan 2020
Revised: 30 Jan 2020
Accepted: 02 Feb 2020
Published Online: 04 Feb 2020

1 Introduction

The emergency department (ED) is an essential part of the hospital. Relying on professional practicians and the high technological equipment, ED can offer immediate medical care. The emergency services overcrowding problem is arising because of the gradual aging of the population. It would result in an overcrowding problem, if a hospital cannot make an efficient allocation of ED resource. ED overcrowding must affect not only patient satisfaction but also the quality of treatment and prognosis. The pivotal step for allocating ED resources is how to predict the demand for ED. The results generated by forecasting model medical would guide decision-making in many tasks such as, staff supplementation, expansions of beds. In recent years, several researchers have concerned demand forecasting for staff allocation and resources in hospitals (Asplin *et al*., 2006, Hwang and Lee, 2008, Schweigler *et al*., 2009, Georgio *et al*., 2017). Most scholars firstly consider traditional time series models as prediction models to forecast medical resource demand and several time-series models have been proposed and applied to handle the different forecasting areas (Bollerslev, 1986, Engle, 1982, Huarng, 2001, Song and Chissom, 1993, Wei *et al.,* 2017, Hamida and Scalera, 2019). Engle (1982) proposed the ARCH (p) (Autoregressive Conditional Heteroscedasticity) model that has been used by several financial and economic analysts and the GARCH (Bollerslev, 1986) (Generalized ARCH) model is the generalized form of ARCH. Box and Jenkins (1976) proposed the autoregressive moving average (ARMA) model which combines a moving average process with a linear difference equation to obtain an autoregressive moving average model, and the ARMA model performs forecasting at the linear stationary condition. Models that describe such homogeneous non-stationary behavior can be obtained by supposing some suitable differences in the process to be stationary. Therefore, the autoregressive integrated moving average model (ARIMA) (Box and Jenkins, 1976) with the assumption of linearity among variables was proposed to handle the non-stationary behavior datasets. Besides, linguistic expressions are often used to describe daily observations. Hence, Song and Chissom (1993) first proposed the original model of the fuzzy time-series and the following researcher, Chen (1996) proposed refined fuzzy time series model for enrollments forecasting. In focusing on establishing fuzzy relationships of the fuzzy time series model, Yu (2005) recommend that different weights should be set in various fuzzy relationships and proposed a weighted fuzzy time-series method to forecasting stock index. From the literature above, AR (autoregressive) is a fundamental and important method in time series models. The application of conventional time series models needs to meet statistical assumptions and not all models can be applied in all datasets (Wei, 2013). Besides, most of the traditional time series models use a single variable for forecasting. However, there are many noises involute in raw data that are caused by changes in surrounding conditions for patient volumes forecasting. The conventional time-series models which use complicated raw data would reduce the forecasting performance (Wei, 2016).

Today, information technology plays an important role in managing knowledge in the healthcare environment (Turan and Palvia, 2014, Cheng, 2012). However, there are still challenges in how to use advanced technology to create and disseminate healthcare knowledge. Information technology remains a key tool in healthcare management applications. Information technology is also used to anticipate healthcare needs, and as the demand for emergency medical services continues to increase, the use of information technology for forecasting is becoming increasingly important. Data mining technology is a kind of information technology. It is a rapidly developing technology in information processing applications. It has attracted much attention in the field of knowledge discovery. The knowledge discovery process consists of data collection, data selection, pattern recognition, and knowledge representation. Data mining consists of techniques above and has been applied to various disciplines such as business and medical data prediction (Aneeshkumar and Venkateswaran, 2015, Arslan *et al*., 2016, Cheng and Wei, 2014, Izadi and Taghva, 2017). Kim and Han (2000) proposed a genetic algorithm approach to feature discretization and the determination of connection weights for artificial neural networks (ANNs) to predict the stock price index. Roh (2007) integrated neural network and time-series model for forecasting the volatility of the stock price index. Jones *et al.* (2008), use artificial neural networks (ANNs) to forecast daily patient volumes in the emergency department. ANNs are promising forecasting methods and have been applied extensively in many domains. However, they would suffer from network construction problems and the need for large training datasets.

Besides, there are other artificial intelligence prediction methods, such as the Random Forest algorithm (RF). Random Forest is a supervised machine learning algorithm, which is a combined classification method based on statistical learning theory (Breiman, 2001). In a random forest, pluralities of random variable samples were selected as training data sets using a bagging procedure. The bagging process refers to random sampling and replacement, which is used to reduce the predicted variation and help to prevent over-fitting. A tree classifier corresponding to the selected sample is then constructed during the data training process. Decision tree classifiers and regression tree classifiers in RF are often used in tree-based classification methods. Finally, RF would combine all the classification trees by voting on each classification result and then select the final classification result based on the number of votes. Random forest method can process nonlinear data and obtained outstanding forecasting results in past researches.

To enhance prediction performances, except single forecasting algorithms hybrid models are often utilized and models that use empirical mode decomposition (EMD) have gained great attention (Chen *et al.,* 2012). EMD (Huang *et al*., 1998) is a useful method to deal with non-linear signal analysis (such as stock data) or other related fields (Vincent *et al*., 1999, Yu *et al*., 2005) and offers a new way to deal with nonlinear and non-stationary signals. EMD-based prediction methods have been used on wind speed prediction (An *et al*., 2012, Ren *et al*., 2014), industry (Feng *et al*., 2010), tourism management (Lai and Yeh, 2013), and financial time series forecasting (Fu, 2010). Based on EMD, any complicated signal can be decomposed into a finite and often small number of intrinsic mode functions (IMFs), which have simpler frequency components and stronger correlations, thus are easier and more accurate to forecast. In recent years, feature selection methods have been applied to forecasting models (Wei and Cheng 2012). The feature selection process is for evaluating features, selecting relevant features and removing redundant and/or unrelated features. Three important main advantages of feather selection are (1) model simplification, (2) easy to interpret, and (3) faster model induction and structural knowledge. The forecasting model combined with feather selection will improve prediction performance and prediction accuracy.

From the mentioned above, there are some major drawbacks in those models: (1) some traditional time series models cannot be applied to the datasets that do not follow the statistical assumptions; (2) most conventional time-series models utilize late-day data with noises as an input variable in forecasting. However, there are noises including in raw data that are generated by environment and weather conditions. To overcome the drawbacks above, this paper considers that EMD can decompose the complicated raw data into simpler frequency components and highly correlations variables. Then, this study utilizes RF as a forecasting model that can overcome the limitations of statistical methods (data need to obey some mathematical distribution). Based on the abovementioned concepts, the proposed model firstly tests the lag of the AR model. Secondly, the input variables of AR are decomposed by EMD into several IMFs and a residue. However, the IMF attribute set generated from the EMD decomposition will generate high-dimensional data and will increase the computational complexity. Therefore, this paper will utilize the feature selection method and take the advantages of feature selection to solve the problems that will arise from multi-feature data. In this study, the method of feature selection is used to reduce the IMF feature set generated by EMD decomposition. Finally, the random forest is combined with the AR method and the reduced IMF attribute set for modeling and prediction. Therefore, the proposed model can be expected to produce more accurate prediction results for solving the emergency department patient's overcrowding problem.

2. Methodology of Research

This section reviews related methodologies of the autoregressive model correlation-based feature selection, empirical mode decomposition and random forest algorithm.

2.1. Autoregressive Model

In the time series forecast, predictions are practically obtained by forecasting a value at the next period based on a specific prediction algorithm. Besides, forecasting non-periodic short-term time series is much more difficult than that for long-term time series. The autoregressive moving-average (ARMA) is a traditional method that is very suitable for forecasting regular periodic data such as seasonal or cyclical time series (Chang, 2008).

Box and Jenkins (1976) developed a general linear stochastic model by assuming that time-series data can be generated by a linear aggregation of random shocks. In this study, we focus on AR model, which is a model includes one or more past values of the dependent variable among its explanatory variables, and a simplest AR(1) is defined as:

$$y_t = \phi_1 y_{t-1} \tag{1}$$

When the random error and the constant term are taken into account, the modified AR(1) model becomes

$$y_t = \mu + \phi_1 y_{t-1} + u_t \tag{2}$$

Where $\Phi_1$is the first-order autoregression coefficient and u_t is the white noise viewed as a random error. An autoregressive model is simply a linear regression of the current value of the series against one or more prior values of the series. In the AR(1) model, it can be thought like that for a given value *y* in period *t* that has a relationship with period *t*–1. If there is an autoregressive model of order *p*, an AR(*p*) model can be expressed as

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + u_t \tag{3}$$

2.2. Correlation-based Feature Selection (CFS)

Correlation-based Feature Selection (CFS) is proposed by Hall (1998). A central problem in machine learning is to identify a set of representative features from data to builds a classification model for a specific task. The CFS method mainly expects to use the correlation-based method for dealing with feature selection problems. The major concept is that a good feature set contains highly features related to decision feature, but is not related to each other. Based on the proposition that the feature evaluation formula comes from the test theory, an operational definition of this hypothesis is provided. CFS (Based on relevance attribute filtering) is an algorithm that couples this evaluation formula with appropriate relevance measures and heuristic search strategies. The CFS method has a function that can quickly identify and screen irrelevant and redundant features. In practical applications, CFS usually eliminates more than half of the input attributes that are independent of the output attributes. In most cases, the forecasting accuracy of the classifier which utilizes the attributes filtered by the CFS method is higher than that produced by using a prediction model that takes all attributes as input variables. In general, the CFS method outperforms the wrapper attribute filtering method on small data sets. CFS executes many times faster than the wrapper, which allows it to scale to larger data sets.

2.3. Random Forest Algorithm

Random forest algorithm (RF)(Breiman, 2001) is to build a forest randomly. The forest is composed of many decision trees (DT). There is no correlation between each decision tree in the random forest. After obtaining the algorithm structure of the entire forest, if new data must be classified, the attribute value of this data is input to each decision tree in the forest, and each decision tree will return a classification value for the category voting. The classification value predicted by the RF is the result of the most voting among all DT-related categorical variables in the forest. Each DT has the following characteristics:

1. If *N* is the number of instances in the data set, RF chooses a random sample and replaces *N* instances from the original data. This sample will be used as the training set for the forest.

2. If *M* is the number of features in the dataset, assign a value *m* << *M*. During the forest construction process, this value of *m* remains unchanged.

3. On each node of the tree:

3.1 Randomly select *m* attributes from original *M* features.

3.2 The segmentation criterion is calculated based on these features. Features with classification ability will be used to split node.

There is no pruning process after RF builds each decision tree. In the original random forest paper, Breiman (2001) used two RF methods:

Random Forests Using Random Input Selection: The simplest random forest with random features is formed by selecting at random, at each node, a small group of input variables to split on. Grow the tree using classification and regression tree (CART) methodology to maximum size and do not prune. Denote this procedure by Forest-RI. The size *F* of the group is fixed. Two values of *F* were tried. The first used only one randomly selected variable, i.e., *F*=1. The second took *F* to be the first integer less than $log_2^{(M)} + 1$, where M is the number of inputs.

Random Forests Using Linear Combinations of Inputs: If there are only a few inputs, say *M*, taking F an appreciable fraction of M might lead to an increase in strength but higher correlation. Another approach consists of defining more features by taking random linear combinations of a number of the input variables. That is, a feature is generated by specifying *L*, the number of variables to be combined. At a given node, *L* variables are randomly selected and added together with coefficients that are uniform random numbers on [-1, 1]. *F* linear combinations are generated, and then a search is made over these for the best split. This procedure is called Forest-RC.

The method chosen in this paper is random forests using random input selection, which is the most common method. The number of features selected for segmentation is the first integer value to be determined, and its value is less than $log_2^{(M)}$ +1, where *M* is the number of input variables (attributes). It also uses Information Gain as the segmentation criterion. The Information Gain algorithm relies on the so-called "Entropy". Its formula is:

Entropy = $-p * log_2^p - q * log_2^q$ (4)

p: the probability of success (or probability of true) *q*: the probability of failure (or probability of false).

2.4. Empirical Mode Decomposition

The empirical mode decomposition (EMD) technique, proposed by Huang et al. Empirical mode decomposition (Huang *et al.,* 1998), is a form of adaptive time series decomposition technique using the Hilbert-Huang transform (HHT) for nonlinear and non-stationary time series data. The basic principle of EMD is to decompose a time series into a sum of oscillatory functions, namely, intrinsic mode functions (IMFs). In the EMD, the IMFs must satisfy two conditions: (1) the number of extreme (sum of maxima and minima) and the number of zero-crossing differs only by one, and (2) the local average is zero. The condition that the local average is zero implies that envelope means of the upper envelope and lower envelope is equal to zero. The first condition is similar to the traditional narrowband requirements for a stationary Gaussian process (Huang *et al.*, 1998). The second condition modifies classical global requirement to a local one; it is necessary so that the instantaneous frequency will not have the unwanted fluctuations induced by asymmetric waveforms (Huang *et al.,* 1998). The detailed algorithm for EMD is shown as follows (Huang *et al.*, 1998):

Step 1: Identify local extreme in the experimental data {x (t)}. All the local maxima are connected by a cubic spline line $U(t)$, which forms the upper envelope of the data. Repeat the same procedure for the local minima to produce the lower envelope $L(t)$. Both envelopes will cover all the data between them. The mean of the upper envelope and lower envelope $m_1(t)$ is given by:

$$m_1(t) = U(t) + L(t)/2 \tag{5}$$

Subtracting the running mean $m_1(t)$ from the original time series $x(t)$, we get the first component $h_1(t)$,

$$h_1(t) = x(t) - m_1(t) \tag{6}$$

The resulting component $h_1(t)$ is an IMF if it is symmetric and has all maxima positive and all minima negative. An additional condition of intermittence can be imposed here to sift out waveforms with a certain range of intermittence for physical consideration. If $h_1(t)$ it is not an IMF, the sifting process has to be repeated as many times as it is required to reduce the extracted signal to an IMF. In the subsequent sifting process steps, $h_1(t)$ is treated as the data to repeat steps mentioned above,

$$h_{11}(t) = h_1(t) - m_{11}(t) \tag{7}$$

Again, if the function $h_{11}(t)$ does not yet satisfy criteria for IMF, the sifting process continues up to k times until some acceptable tolerance is reached:

$$h_{1k}(t) = h_{1_{(k-1)}}(t) - m_{1k}(t) \tag{8}$$

Step 2: If the resulting time series is an IMF, it is designated as

$c_1 = h_{1k}(t)$.

The first IMF is then subtracted from the original data, and the difference r_1 given by

$$r_1(t) = x(t) - c_1(t) \tag{9}$$

is the residue. The residue $r_1(t)$ is taken as if it were the original data, and we apply to it again the sifting process of Step 1.

Following the above procedures, we continue the process to find more intrinsic modes ci until the last one. The final residue will be a constant or a monotonic function which represents the general trend of the time series. Finally, we obtain

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n$$

$$r_{i-1}(t) - c_i(t) = r_i(t) \tag{10}$$

Where r_n is a residue

Thus, the residue $r_n(t)$ is the mean trend of x(t) $x(t)$. The IMFs $c_1(t), c_2(t), \ldots, c_n(t)$ include different frequency bands ranging from high to low. The frequency components contained in each frequency band are different and they change with the variation of signal $x(t)$, while $r_n(t)$ represents the central tendency of the signal $x(t)$.

3. Proposed Model

In this section we illustrate the datasets collected in this study. Next section shows the proposed algorithm of this study.

3.1. Data Source

This study collects datasets from the hospital information system of a regional hospital. Emergency department has three diverse divisions (internal medicine, surgical, pediatrics), and all daily patient volumes (all datasets) of the emergency department contain patients of three different divisions. Patients visit the division of internal medicine and division of surgical every day but patients visit the pediatrics division just in a few days. Thus, this study only extracts daily patient volumes in the internal medicine division and surgical as two sub-datasets. All patient volumes of ED are denoted as dataset I and patient volumes of internal medical division and surgical division are denoted as dataset II and dataset II, respectively. There are 731 observations from July 2010 to June 2012 in each experimental dataset, data from July 2010 to December 2011 is selected as training data and testing data is extracted from January 2012 to June 2012. The detailed information of these datasets is showed in Table 1.

Table 1. Comparisons of experimental datasets

	Dataset I	**Dataset II**	**Dataset III**
Dataset description	All patient volumes of ED	Patient volumes of internal medical division	Patient volumes of surgical division
Max value	159	128	41
Min value	18	5	4
Mean	42.98	23.94	18.9
Standard deviation	12.8	10.95	5.52

3.2. Proposed Algorithm

Based on the research concepts in Section 1, this paper proposes a hybrid time series model which considers the EMD method, the AR model, and combines CFS to reduce IMFs generated by EMD decomposition. Further, the proposed model utilizes a random forest method to forecast daily patient volumes in an emergency. This study firstly tests the lag of AR by statistical analysis and then uses EMD to decompose input variables of AR, and generated IMFs set is reduced by the CFS method. Finally, the proposed model applies the random forest method to forecast the patient number. The overall process of the proposed model is shown in Figure 1.

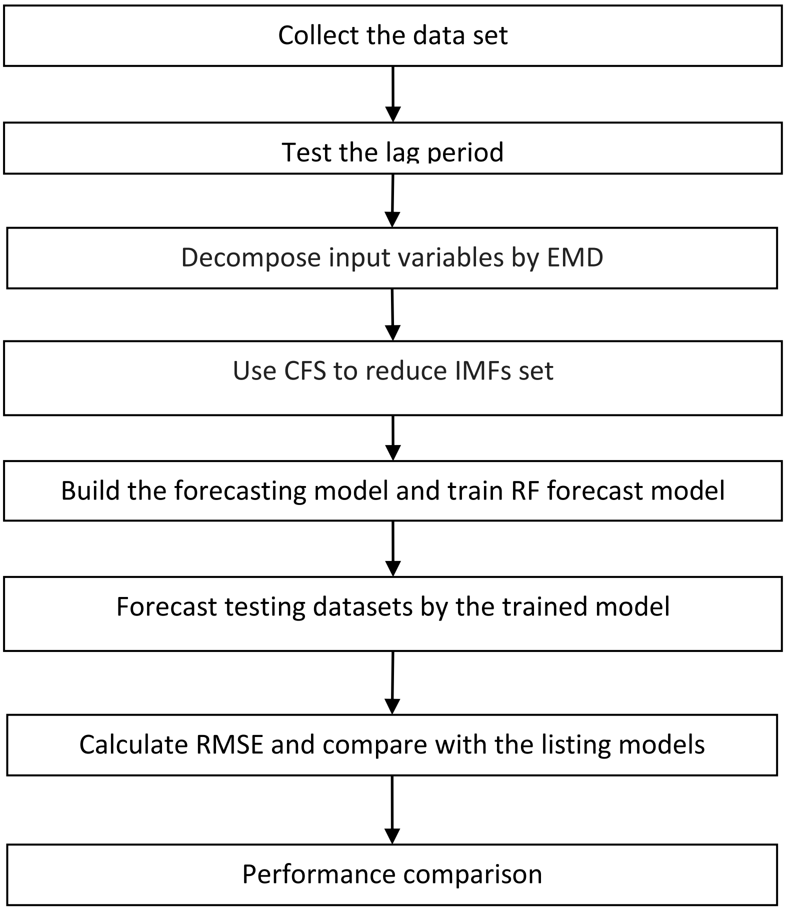

Collect the data set
Test the lag period
Decompose input variables by EMD
Use CFS to reduce IMFs set
Build the forecasting model and train RF forecast model
Forecast testing datasets by the trained model
Calculate RMSE and compare with the listing models
Performance comparison

Figure 1. Flowchart of the proposed procedure

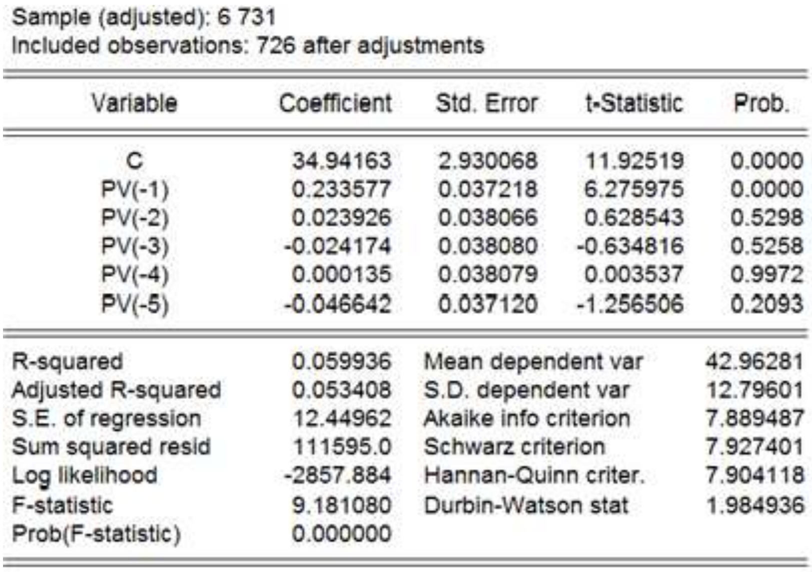
This section uses practically collected data (Dataset I) as the example step by step to show the core concept of the proposed algorithm as follows.

Step1: collect the data set

All patient volumes of ED (Dataset I) from July 2010 to June 2012 are collected to demonstrate the proposed model. There are 549 training data (from July 2010

Step 2: Test the lag period

E-Views software package is utilized to fit the AR model for orders and different lags of patient volumes (PV). There are five linear regression variables in dataset I, (i.e., from PV (t − 1) to PV (t − 5)) are selected to be estimated and tested. If the p-value is less than the significant level of 0.05, then reject the null hypothesis. Take Dataset I as an example, Figure 2 illustrates that the p-value (0.0000) for PV (t−1)) is less than the significant level of 0.05 among five variables, from PV (t -1) to PV (t −5). Further, the variable PV (t -1) is not equal to zero. Therefore, the order of AR is one.



```
Sample (adjusted): 6 731
Included observations: 726 after adjustments
====================================================================
     Variable        Coefficient    Std. Error    t-Statistic   Prob.
====================================================================
        C             34.94163      2.930068      11.92519     0.0000
      PV(-1)          0.233577      0.037218      6.275975     0.0000
      PV(-2)          0.023926      0.038066      0.628543     0.5298
      PV(-3)         -0.024174      0.038080     -0.634816     0.5258
      PV(-4)          0.000135      0.038079      0.003537     0.9972
      PV(-5)         -0.046642      0.037120     -1.256506     0.2093
====================================================================
R-squared             0.059936    Mean dependent var        42.96281
Adjusted R-squared    0.053408    S.D. dependent var        12.79601
S.E. of regression    12.44962    Akaike info criterion     7.889487
Sum squared resid     111595.0    Schwarz criterion         7.927401
Log likelihood       -2857.884    Hannan-Quinn criter.      7.904118
F-statistic           9.181080    Durbin-Watson stat        1.984936
Prob(F-statistic)     0.000000
====================================================================
```

Figure 2. Testing the lag period of PV in Dataset I

Step 3: Decompose input variables by EMD

From step 2, the lag period test results shows that the order of AR is one. Therefore, the input variable (PV(t)) is decomposed by EMD into a finite set of IMFs (the residual $r_{n+1}(t)$ also be considered as an IMF). There are ten IMFs and one residue generated from PV(t) in Dataset I.

Step 4: Use CFS to reduce IMFs set

In this step, this study uses the feature selection method (CFS method) to reduce the attributes of the IMF attribute set generated in the third step. Seven IMFs are selected by CFS method from 11 IMFs generated in the previous step.

Step 5: Build the forecasting model and train RF forecast model

Then, this paper uses *PV (t)* and the 7 IMFs selected by the CFS in step 4 as the input attributes of the prediction model and uses *PV (t + 1)* (the next day's *PV* value) as the output variable. Then, this paper applies a random forest method to build a prediction model.

Step 6: Forecast testing datasets by the trained model

The random forecast parameters of the forecasting models are determined when the stopping criterion is reached from step 5, then the training forecasting model is used to forecast *PV(t+1)*, for the target testing datasets.

Step 7: Calculate RMSE and compare with the listing models

Calculate RMSE values in testing datasets by Equation (11). Then the RMSE is taken as an evaluation criterion to compare with the listing models.

$$\text{RMSE} = \sqrt{\sum_{t=1}^{n} | \, actual(t) - forecast(t) \, |^2 \Big/ n} \qquad (11)$$

Where *actual(t)* denotes the real PV value, *forecast(t)* denotes the predicting PV value and *n* is the number of data.

Step 8: Performance comparison

RMSE values in testing datasets are calculated by equation (11). Then, the RMSE is taken as the evaluation criterion to compare with the listing models.

4. Experiments and Comparisons

In this section, the RMSE is taken as the evaluation criterion to evaluate forecasting accuracy. The Dataset I, II, III are used as the experimental datasets to verify the proposed model. Training Data are selected from July 2010 to December, and those January 2012 to June 2012 are selected for testing in each dataset. Further, this paper compares the forecasting accuracy of the proposed model with the traditional time-series model (AR (1) model (Engle, 1982)), fuzzy time series model (Chen's model (Chen, 1996), Yu's model (Yu, 2005)) and random forest (Breiman, 2001) model. AR-EMD-RF model (IMFs set without being reduced by feature selection) is also as comparison model. This study tests the lag period of PV in Dataset II and II and the order of AR for two datasets is all one. The number of decomposed IMFs in Dataset II and III are all 9. The number of IMFs selected by CFS method in Datasets II and III reduced are all 8.

The performances of the listing models above used to forecast PV are compared to the proposed model. The forecasting performances of the AR (1) model, Chen model, Yu model, random forest model, AR-EMD-RF model, and the proposed model are listed in Table 2. From Table 2, results show that the proposed model outperforms the other five models in each dataset. These PV forecasting performance evaluations illustrate the excellent performance of the proposed model.

Table 2. The results of different models for testing data in RMSE

Models	**Dataset**		
	I	**II**	**III**
Chen's model	17.33	26.4	6.12
Yu's model	14.65	12.06	6.63
AR(1)	12.4	10.31	5.43
RF	12.9	11.23	5.47
AR-EMD-RF	10.15	8.35	4.61
Proposed model	9.87 [a]	8.16 [a]	4.60 [a]

[a] The best performance among six models

This paper uses a nonparametric statistical method, the Friedman test (Friedman 1937), to verify that the proposed model is superior to the other methods. Using the data from Tables 2, a chi-square test is used to test the hypothesis H_0: equal performance. The results (p=0.012) concerning this hypothesis, which rejects H_0= 0, are listed in Tables 3. Tables 4 shows the mean rank by Friedman test, demonstrating that the proposed model (mean rank = 1) outperforms the other models. Based on Tables 3-4, the difference in performance is significant.

Table 3. Results of the Friedman test

Parameter	
n	3
Chi-Square	14.619
df	5
Asymp. Sig.	0.012

Note: n is the number of data points; df denotes the degree of freedom

Table 4. Mean rank of Friedman Test

Models	**Mean Rank**
Chen	5.67
Yu	5.33
AR(1)	3.00
SVR	4.00
AR-EMD-SVR	2.00
Proposed model	1.00

5. Findings

A novel model, based on AR-EMD and feature selection method joining to fusion random forest algorithm, has been proposed to forecast patient volumes in Taiwan. Further, the proposed model is compared with five forecasting models, Chen's model Yu's model, AR(1) model, AR-EMD-RF model, and random forest model, to evaluate the performance of proposed model. After verification and comparison, the proposed method surpasses the listing methods. There are three findings from the experimental results in this paper as follows:

(1) *The advantage of the hybrid model*: According to Table 2, it is evident that the hybrid models (AR-EMD-RF model and proposed model) are superior to the single methods (Chen model, Yu model, AR model, random forest model) in condition of RMSE. The major reason is that the hybrid models take into account AR-EMD method with random forest learning for PV forecasting, integrating the advantage of random forest, which offer efficient estimates of the test error without incurring the cost of repeated model training associated with cross-validation.

(2) EMD superiority: From Table 2, it is shown that the performances of the proposed model and AR-EMD-RF are better than the random forest model. EMD methods would decompose the noise raw data into highly correlations and input variables simpler frequency components, and reduce prediction error more efficaciously.

(3) Reducing IMFs set for forecasting performance: Table 2 discloses that the proposed model with reducing IMFs set process performs better than the AR-EMD-RF model. This indicates that the feature selection process of the proposed model could improve forecasting capability.

6. Conclusions

In recent years, the number of emergency department patients has continued to increase, and the rapid changes in the supply and demand of emergency room resources have made the issue of how to effectively allocate emergency department resources more and more important. For the field of clinical research, scholars have proposed many models to predict the daily number of patients in the emergency department. However, there are still some shortcomings in the previous prediction models: (1) some data sets do not meet statistical assumptions, and some traditional time series models in past are not suitable to analyze these data sets; (2) in most traditional time series models, The latest period of data is used as the input variable of the prediction model to predict the value of the next day, but there is often noise in the input original data value.

To solve the deficiencies of the traditional prediction models in the past, this paper proposes a new hybrid model which combines AR and EMD model (EMD can decompose the original data which includes noise into IMFs having strong correlation with the output value), and the random forest algorithm can overcome problem that traditional statistical methods cannot be applied to datasets without obeying the statistical assumptions, and consider advantage that reducing IMFs attributes could improve the performance of the prediction model.

Experimental results show that the proposed model could reliably and accurately predict the number of patients admitted to the emergency department each day. The reason why the proposed model is better than other prediction models used for comparison in this research is that EMD is used as a preprocessor to remove noise from the original signal. IMF has a specialty which is a simple frequency component and a high correlation with the output value. Besides, the proposed model uses the method of feature selection to further reduce the IMF attribute set to enhance prediction performance.

Through this pre-processing mechanism, the proposed model not only promotes the simplification of the random forest modeling process but also has more accurate prediction performance than other prediction methods (based on the RMSE evaluation criteria). Therefore, this method is very suitable for analyzing nonlinear and noise including data and is an effective method for predicting the number of patient visits. Besides, the results of this paper are useful and feasible for policymakers and scholars in related research fields in the future. Hospital managers can use this predictive model to discover useful relevant knowledge in the field of medical research.

In subsequent related work, more patient visit data can be collected as an experimental data set to verify the stability of the proposed model. In the future, other traditional time series methods can be integrated into the new prediction model to improve prediction performance. Other feature selection methods can be combined with the forecasting model to enhance prediction performance. Further, researchers can also consider data mining methods that can generate rules to be used as a forecasting model, and it is expected that the new model can generate useful decision rules for hospital managers or related decision-makers.

References

1. An, X., Jiang, D., Zhao, M., Liu, C. (2012) Short-term prediction of wind power using EMD and chaotic theory, Commun. Nonlinear Sci. Numer. Simul. 17 1036-1042,

2. Aneeshkumar, A. S., Venkateswaran, C. J. (2015) Reverse Sequential Covering Algorithm for Medical Data Mining, Procedia Computer Science, 47, 109-117,

3. Arslan, A. K., Colak C., Sarihan, M. E. (2016) Different medical data mining approaches based prediction of ischemic stroke, Computer Methods and Programs in Biomedicine, 130, 87-92,

4. Asplin, B. R., Flottemesch, T. J., Gordon, B. R. (2006) Developing models for patient flow and daily surge capacity research. Acad Emerg Med; 13: 1109-1113.

5. Bollerslev, T. (1986) Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics. 31 307-327.

6. Box, G., Jenkins, G. (1976) Time series analysis: Forecasting and control, San Francisco: Holden-Day.

7. Breiman, L. (2001). Random forests, Mach. Learn. 45 (1) 5–32,.

8. Chang, B. R. (2008) Resolving the forecasting problems of overshoot and volatility clustering using ANFIS coupling nonlinear heteroscedasticity with quantum tuning. Fuzzy Sets and Systems, 159(23) 3183-3200.

9. Chen, S. M. (1996) Forecasting enrollments based on fuzzy time series, Fuzzy Sets Systems, 81, 311-319.

10. Chen, K. L., Yeh, C. C., Lu, T. (2012) Forecasting the output of Taiwan's integrated circuit (IC) industry using empirical mode decomposition and support vector machines, Int. J. Phys. Sci. 3 (78) 5460-5467.

11. Cheng, C. H., Wei, L. Y. (2014) A novel time series model based on empirical mode decomposition for forecasting TAIEX, Economic Modelling, 136-141

12. Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimator of the variance of United Kingdom inflation. Econometrica. 50(4) 987-1008.

13. Feng, F., Zhu, D., Jiang, P., Jiang, H. (2010) GA-EMD-SVR condition prediction for a certain diesel engine, 2010 Progn. Syst. Heal. Manag. Conf. PHM '10,

14. Fu, C. (2010) Forecasting exchange rate with EMD-based Support Vector Regression, in: 2010 Int. Conf. Manag. Serv. Sci. MASS,

15. Friedman, M. (1937), The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc.; 675-701.

16. Georgio, G., Guttmann, A., Doan, Q. H. (2017). Emergency Department Flow Measures for Adult and Pediatric Patients in British Columbia and Ontario: A Retrospective, Repeated Cross-Sectional Study, The Journal of Emergency Medicine, In press.

17. Hall, M. A. (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

18. Hamida, H. B. H., Scalera, F. (2019)Threshold Mean Reversion and Regime Changes of Cryptocurrencies using SETAR-MSGARCH Models International Journal of Academic Research in Accounting, Finance and Management Sciences Vol. 9, No.3, pp. 221–229

19. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng Q. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis, in: Proceedings of the royal society of London series a-mathematical physical and engineering sciences, series A, 454 903-995.

20. Huarng, K. H. (2001) Effective lengths of intervals to improve forecasting in fuzzy time series, Fuzzy Sets and Systems. 123 155-162.

21. Hwang, S. W., Lee, H. J. (2008) Development of a revisit prediction model for the outpatient in a hospital. J Korean Soc Med Inform; 14: 137-145.

22. Izadi, M., Taghva, M. R. (2017)Using AHP Technique and Fuzzy VIKOR Technique to Select a Dynamic Enterprise Resource Planning System in Rayan Pardazesh Co. Case study, International Journal of Academic Research in Accounting, Finance and Management Sciences Vol. 7, No.2, April, pp. 64–75

23.Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., Snow, G. L. (2008) Forecasting Daily Patient Volumes in the Emergency Department, Academic Emergency Medicine; 15:159–170

24.Kim, K., Han, I. (2000) Genetic algorithms approach to feature discretization in artificial neural networks for prediction of stock index. Expert System with Applications. 19 125-132.

25.Lai, M. C., Yeh, C. C. (2013) A hybrid model by empirical mode decomposition and support vector regression for tourist arrivals forecasting, J. Test. Eval. 41.

26.Ren, Y., Suganthan, P. N., Srikanth, N. (2014) A novel empirical mode decomposition with support vector regression for wind speed forecasting, IEEE Trans. Neural Networks Learn. Syst. 1-6.

27.Roh, T. H. (2007) Forecasting the volatility of stock price index. Expert Systems with Applications. 33 916-922.

28.Schweigler, L. M., Desmond, J. S., McCarthy, M. L., Bukowski, K. J., Ionides, E. L., Younger, J. G. (2009). Forecasting models of emergency department crowding. Acad Emerg Med; 16: 301-308.

29.Song, Q., Chissom, B. S. (1993) Forecasting enrollments with fuzzy time series Part I, Fuzzy Sets and Systems. 54 1-10.

30.Turan, A. H., Palvia, P. C. (2014) Critical information technology issues in Turkish healthcare, Information & Management, 51, (1), 57-68,

31.Vincent, H. T., Hu S. L. J., Hou, Z. (1999) Damage detection using empirical mode decomposition method and a comparison with wavelet analysis, in: Proceedings of the second international workshop on structural health monitoring, Stanford 891-900.

32.Wei, L. Y. **(**2013) A GA-weighted ANFIS model based on multiple stock market volatility causality for TAIEX forecasting , Applied Soft Computing, 13 911-920

33.Wei, L. Y. **(**2016). A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting, Applied Soft Computing, 368-376

34.Wei, L. Y., and Cheng, C. H. (2012) A HYBRID RECURRENT NEURAL NETWORKS MODEL BASED ON SYNTHESIS FEATURES TO FORECAST THE TAIWAN STOCK MARKET, International Journal of Innovative Computing, Information and Control, 8 (8), pp. 5559-5571

35.Wei, L. Y., Tsai, C. H., Chung, Y. C., Liao, K. H., Chueh, H. E., Lin, J. S. (2017) A Study of the Hybrid Recurrent Neural Network Model for Electricity Loads Forecasting, International Journal of Academic Research in Accounting, Finance and Management Sciences Vol. 7, No.2, pp. 21–29

36.Yang, J. J., Li, J., Mulder, J., Wang, Y., Pan, H. (2015) Emerging information technologies for enhanced healthcare, Computers in Industry, 69, 3-11,

37.Yu, D. J., Cheng, J. S., Yang, Y. (2005) Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings, Mech. Syst. Signal Process. 19 (2) 259-270.

38.Yu, H. K. (2005) Weighted fuzzy time series models for TAIEX forecasting, Physica A. 349 609-624.