VOL. 10, NO 21, NOVEMBER, 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

FORECASTING OF MONTHLY TEMPERATURE VARIATIONS USING RANDOM FORESTS

Wai Yan Nyein Naing and Zaw Zaw Htike
Department of Mechatronics Engineering, Faculty of Engineering, International Islamic University Malaysia, Malaysia

ABSTRACT

This study utilized a random forest model for monthly temperature forecasting of KL by using historical time series data of (2000 to 2012). Random Forest is an ensemble learning method that generates many regression trees (CART) and aggregates their results. The model operates on patterns of the time series seasonal cycles which simplifies the forecasting problem especially when a time series exhibits nonstationarity, heteroscedasticity, trend and multiple seasonal cycles. The main advantages of the model are its ability to generalization, built-in cross-validation and low sensitivity to parameter values. As an illustration, the proposed forecasting model is applied to historical load data in Kuala Lumpur (2000 to 2012) and its performance is compared with some alternative models such as K-Nearest Neighbours , Least Medium square Regression , RBF (Radial Basic Function) network and MLP (Multilayer Perceptron) neural networks. Application examples confirm good properties of the model and its high accuracy.

Keywords: random forests, monthly temperature forecasting, K-Nearest neighbours, least medium square regression, RBF (Radial Basic Function) network, MLP (Multilayer Perceptron) neural networks.

1. INTRODUCTION

A. Overview

There are various approaches available in weather forecasting, from relatively simple observation of the sky to highly complex computerized mathematical models. Among the weather forecasting, the prediction of temperature variation condition is essential for various applications. Some of them are climate monitoring, drought detection, severe weather prediction, agriculture and production, planning in energy industry, aviation industry, communication, pollution dispersal, and so forth [1]. Temperature forecasting is the most important services provided by the meteorological profession to protect life and property of residents in their respective counties and to improve the efficiency of operations, and by individuals to plan a wide range of daily activities [2]. Long range of forecasting temperature variation is a very challenging task. For instance, although we can forecast tomorrow's weather with a certain degree of accuracy, it is highly challenging to forecast monthly temperature accurately even for the professionally trained meteorologists. By directing outgrowth of time series forecasting technology, the notable improvement of accuracy has been achieved in prediction of environmental factors including temperature forecasting, flood detection, amount of rainfall estimation. A basic concept of time series forecasting is a prediction of some future event or events, a sequence of observations over time, yi ∈ R that can define the mathematically function in Equation (1) [3].

$$Yi = \{yi \in R \mid I = 1, 2, 3 \ldots n\} \quad (1)$$

Where Yi is treated as random variables, i represent the time index; n represents the number of sample or observation. For time series forecasting, different kinds of algorithms and data analysis methods are accessible for various purposes. In the past few years, machine learning algorithms have been advanced and become serious opponents to traditional statistical models for time series forecasting. The fundamental concept of machine learning is the domain of computational intelligence which is concerned with the question of how to construct computer programs that automatically improve with experience [3]. To perform time series prediction, various machine learning algorithms have been proposed. This paper is utilized one of the machine learning model called Random Forest to forecast temperature variation of Kuala Lumpur based on meteorological time series data of (2000 to 2012) [4].

B. Related works

Machine learning algorithms have been advanced and become serious opponents to traditional statistical models for various science and engineering fields in the past few years. Especially, neural network models have been remarkable developments, both in the amount and variations of the models established and the theoretical understanding of the models in the last few years. Ricardo [5] have proposed ANN (Artificial Neural Network) model for simulation of daily temperature for climate change over Portugal. In their study, they compared performances of linear models and non-linear ANN using a set of rigorous validation techniques to construct scenarios of daily temperature at the present day (1970–79) and for a future decade (2090–99). Charles Jones etc. [6] also developed ANN model in their research to predict

air surface temperature over the city at the University of California, Santa Barbara. Chaudhuri etc [7] came with a Feed forward multi-layered artificial neural network model to estimate the maximum surface temperature and relative humidity. Their result has been stated that one hidden-layer neural network is an efficient forecasting tool for predicting maximum surface temperature and relative humidity. Amanpreet Kaur etc., [7] have been tested Artificial Neural Network in forecasting minimum temperature, they have used multilayer perceptron architecture to model the forecasting system and back propagation algorithm is used to train the network. Their result found that minimum temperature can be predicted with reasonable accuracy using ANN model. For the issue of time series forecasting of temperature variation , one of the machine learning model called neural network have been commonly used in the last decades. Then, research observation have been extended to new models, such as, ARIMA models, the exponential smoothing, decision trees, self-organizing maps (SOMs) and others [8].

C. Outline

In this study, we will utilize Random Forest model to improve forecasting accuracy of maximum and minimum temperature by using meteorological data of Kuala Lumpur city located in Malaysia for one month ahead forecasting of temperature of this area. The remains content of the sections are organized as follows. Section 2 will explain the Random Forest model for forecasting purpose. In Section 3, we will present the experimental results of the simulated model and section 5 some concluding remarks are summarized.

2. RANDOM FORESTS MODEL

Random Forests (RF) is the most popular methods in data mining. The method is widely used in different time series forecasting fields, such as biostatistics, climate monitoring, planning in energy industry and weather forecasting. Random forest (RF) is an ensemble learning algorithm that can handle both high-dimension classification as well as regression. RF is a tree-based ensemble method where all trees depend on a collection of random variables. That is, the forest is grown from many regression trees put together, forming an ensemble [9]. After individual trees in ensemble are fitted using bootstrap samples, the final decision is obtained by aggregating over the ensemble, i.e. by averaging the output for regression or by voting for classification. This procedure called bagging improves the stability and accuracy of the model, reduces variance and helps to avoid overfitting. The bias of the bagged trees is the same as that of the individual trees, but the variance is decreased by reducing the correlation between trees (this is discussed in [10]). Random forests correct for decision trees' habit of overfitting to their training set and produce a limiting value of the generalization error [11]. The RF generalization error is estimated by an out-of-bag (OOB) error, i.e. the error for training points which are not contained in the bootstrap training sets (about one-third of the points are left out in each bootstrap training set). An OOB error estimate is almost identical to that obtained by *N*-fold cross-validation. The large advantage of RFs is that they can be fitted in one sequence, with cross-validation being performed along the way. The training can be terminated when the OOB error stabilizes [12]. The algorithm of RF for regression is shown in Figure-1 [10].

1. For $k = 1$ to K:
   1.1. Draw a bootstrap sample L of size N from the training data.
   1.2. Grow a random-forest tree T_k to the bootstrapped data, by recursively repeating the following steps for each node of the tree, until the minimum node size m is reached.
      1.2.1. Select F variables at random from the n variables.
      1.2.2. Pick the best variable/split-point among the F.
      1.2.3. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_k\}_{k=1, 2, \ldots, K}$.

To make a prediction at a new point $\mathbf{x}$:

$$f(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} T_k(\mathbf{x}) \quad (1)$$

Figure-1. Algorithm of RF for regression [7].

Where K represents the number of trees in the forest and F represents the number of input variables randomly chosen at each split respectively. The number of trees can be determined experimentally. And, we can add the successive trees during the training procedure until the OOB error stabilizes. The RF procedure is not overly sensitive to the value of F. The inventors of the algorithm recommend F = n/3 for the regression RFs. Another parameter is the minimum node size m. The smaller the minimum node size, the deeper the trees. In many publications m = 5 is recommended. And this is the default value in many programs which implement RFs. RFs show small sensitivity to this parameter.

Using RFs we can determine the prediction strength or importance of variables which is useful for ranking the variables and their selection, to interpret data and to understand underlying phenomena. The variable importance can be estimated in RF as the increase in prediction error if the values of that variable are randomly permuted across the OOB samples. The increase in error as a result of this permuting is averaged over all trees, and divided by the standard deviation over the entire ensemble. The more the increase of OOB error is, the more important is the variable.

3. EXPERIMENT

A. Results and discussions

We tested our proposed Random Forests model using temperature data of twelve years are collected from 'Monthly Maximum and Minimum Temperature of Kuala Lumpur" dataset. This dataset contains maximum temperature of KL and minimum temperature of KL for (2000 to 2012). This data set contains 154 instances. The chosen temperature are were divided into two randomly selected groups, the training group, corresponding to 70% of the patterns, and the test group, corresponding to 30% of patterns; so that the generalization capacity of stimulated model could be checked after training phase. Also four random days were selected as unseen data. We used the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as a measure of error made by Random Forest.

MAE is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation. It usually expresses accuracy as a percentage, and can define mathematically function as follow:

$$\text{MAE} = \frac{1}{M_{total}} \sum_{i=1}^{M_{total}} |P_i - P_i^*|$$

Where P_i represents exact values, P_i^* presents predicted values and M_{total} represents as total number of the test data respectively.

Root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and prediction values) predicted by a model or an estimator and the values actually observed.

RMSE can also define the mathematically function as follow:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y)^2}{n}}$$

Where $\hat{y}_t$ represents predicted values for times *t*, y represents a "regression's dependent variable" and which is computed for *n* different predictions as the square root of the mean of the squares error.

The comparison between optimal structure of Random Forests and some alternative machine learning models for obtaining minimum prediction error are as shown in Table-1.

Table-1. Results of forecasting error

Model		Maximum Temperature				Minimum Temperature	
	MAE		RMSE		MAE		RMSE
KNN (K-nearest neighbor)	0.6457		0.8331		0.6761		0.7892
Least Med square	0.9288		1.1147		0.6609		0.7869
MLP (Multilayer Perceptron)	1.2506		1.4767		1.9321		2.375
RBF (Radial Basic Function)	0.6924		0.8596		0.6995		0.795
Random Forests	0.5597		0.5021		0.6072		0.583

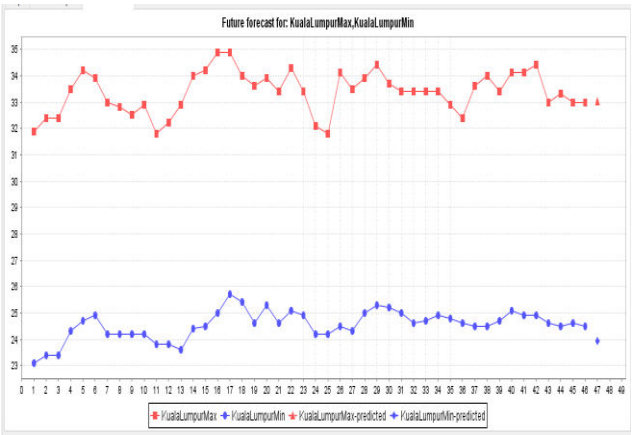

Figure-1. Future forecast of Kuala Lumpur maximum& minimum temperature.

As a result of Table-1 and Figure-1, it is observed that the predicted values of temperature variation are good agreement with exact values and the predicted error is very less. Therefore the proposed Random Forests (RFs) model can perform good prediction with least error than other alternative techniques.

4. CONCLUSIONS

The result of Random Forests model used for one month ahead temperature forecast in the Kuala Lumpur, Malaysia, shows that Random Forests Model has a good performance and reasonable prediction accuracy was achieved for this model. Its forecasting reliabilities were evaluated by computing the mean absolute error and root mean square error between the exact and predicted values. The results suggest that this random forests model could be an important tool for temperature forecasting.

REFERENCES

[1] Kapoor, P. and S.S. Bedi. 2013. Weather Forecasting Using Sliding Window Algorithm. ISRN Signal Processing.

[2] Kenitzer, S., J. Rosenfeld, and K. Heideman. 2007. The American Meteorological Society: Annual Report 2006. Bulletin of the American Meteorological Society. 88(4): p. 1.

[3] Kane, M.J., *et al.*, 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC bioinformatics. 15(1): 276.

[4] Htike, Z.Z. 2013. Multi-horizon ternary time series forecasting. in Signal Processing: Algorithms, Architectures, Arrangements and Applications (SPA). IEEE.

[5] Trigo, R.M. and J.P. Palutikof. 1999. Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach. Climate Research. 13(1): p. 45-59.

[6] Jones, C., P. Peterson, and C. Gautier. 1999. A new method for deriving ocean surface specific humidity and air temperature: An artificial neural network approach. Journal of Applied Meteorology. 38(8): p. 1229-1245.

[7] Shrivastava, G., *et al.*, 2012. Application of artificial neural networks in weather forecasting: a comprehensive literature review. International Journal of Computer Applications (0975-8887).

[8] Ahmed, N.K., *et al*. 2010. An empirical comparison of machine learning models for time series forecasting. Econometric Reviews. 29(5-6): 594-621.

[9] Adriansson, N. and I. Mattsson. 2015. Forecasting GDP Growth, or How Can Random Forests Improve Predictions in Economics?

[10] Hastie T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learnin. New York: Springer.

[11] Breiman, L. 2001. Random forests. Machine learning. 45(1): 5-32.

[12] Dudek, G. Short-Term load forecasting using random forests, in Intelligent Systems' 2014. 2015, Springer. pp. 821-828.