Renewable Energy 66 (2014) 132–139

Contents lists available at ScienceDirect

Renewable Energy

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

Prediction of solar radiation with genetic approach combing multi-model framework

Ji Wu*, Chee Keong Chan, Yu Zhang, Bin Yu Xiong, Qing Hai Zhang

Nanyang Technological University, School of Electrical and Electronics Engineering, Singapore

ARTICLE INFO

Article history:
Received 20 April 2013
Accepted 25 November 2013
Available online

Keywords:
Time series
Cluster
Solar Radiation
Prediction

ABSTRACT

In this paper, a genetic approach combing multi-model framework for solar radiation time series prediction is proposed. The framework starts with the assumption that there exists several different patterns in the stochastic component of the solar radiation series. To uncover the underlying pattern, a genetic algorithm is used to segment the solar series dynamically, and the subsequences are further grouped into different clusters. For each cluster, a prediction model is trained to represent that specific pattern. In the prediction phase, identifying the pattern for current period is of great importance. Thus a procedure for the pattern identification is performed to identify the proper pattern for the series belong to. The prediction result of the proposed framework is then compared to other algorithms. It shows that the proposed framework could provide superior performance compared to others.

1. Introduction

Solar radiation is an important source for many engineer applications [1]. To improve the efficiency of these related applications [1,2], an accurate prediction of solar radiation is required. Since a solar radiation series can be regarded as a time series produced by stochastic process, such an application could be regarded as a time series prediction problem [3].

One popular statistical algorithm for time series prediction is the autoregressive moving and average model (ARMA) [4]. The popularity of the ARMA model is its ability to extract useful statistical properties and the adoption of the well-known Box–Jenkins methodology [5]. ARMA models are very flexible since they can represent time series with varied statistic property by adopting different order. It has been proved to be competent in prediction when there is an underlying linear correlation structure in the time series.

In recent years, a great number of data mining technologies such as time delay neural network (TDNN) and Markov Chain are used in time series prediction by capturing the nonlinear trend lying in time series [6]. Aguiar & Collares-Pereira stochastic model is also proved to be capable for the prediction of solar radiation [7,8].

Some other research work also shows that hybrid model [9] which combines different prediction models has superior prediction performance comparatively.

Some researchers have studied the prediction performance of artificial neural network (ANN) and radial basis function network (RBF) [10]. Comparison experiment between ANN and tradition regression approach is also conducted [11]. Diagne et al. also presented an in-depth review of using different methods to obtain accurate prediction of solar radiation [12].

There are also other works that incorporate geographical coordinates and meteorological data into the prediction models to generate daily solar radiation [13–16].

The solar radiation reaching the surface of the earth could be represented in a number of ways. Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received from above by a surface horizontal to the ground. This value is of particular interest by many industrial applications. In this paper, we focus on the prediction of GHI data recorded by the local station. A novel framework is presented to conduct short term prediction of solar radiation time series. Solar radiation data is the only input parameter needed. Solar radiation series is regarded as stochastic time series. We assume that there exists several different patterns embedded in it. The value of the time series relates not only to time, but also to the pattern it belongs to. It is then segmented into subsequence. The subsequence is grouped into different clusters. Training and building prediction models are formed with the data of each cluster. Hence, when predicting the future trend of a time series, the label of cluster that the

---

* Corresponding author.
E-mail address: wuji0014@e.ntu.edu.sg (J. Wu).

0960-1481/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.renene.2013.11.064

current time series belongs to is determined first and then the corresponding model is used to do the prediction.

In the framework, the accuracy of clustering is of great importance for the prediction. The previous work just adopted a sample fixed length segmentation schema which generate subsequence of identical length [17]. However it is an over-simplified approach because a meaningful pattern might be missed if it is split over different subsequence. A dynamic segmentation is more suitable to be used in the solar radiation clustering. Genetic algorithm is applied to find the optimal segmentation.

To verify the performance of the framework described above, daily solar radiation data of Singapore is applied to it. The data is collected by an observation station located in Nanyang Technological University, Singapore and can be obtained online from http://nwsp.ntu.edu.sg. To ensure the accuracy and adaptability of the framework, 24 months' data is used in the whole process. The data of 2009 is used to train and tune the framework. And the solar radiation data of 2010 is used as validation data.

The rest of the paper is organized as follow: Section 2 details the process of building Genetic Approach Combing Multi-Model Framework (GAMMF) and the related algorithms. Section 3 discusses and analyzes the result of experiments and Section 4 is the comparison with other algorithms. Section 5 is the conclusion offers some comment on future works.

2. The proposed methodology

2.1. Segment & clustering

During the segmentation phase, the solar radiation series is segmented and grouped into different clusters. The data of different cluster is used to train a prediction model. In the final prediction phase, the appropriate prediction model is identified in the current time, and then the actual data is fed to the chosen model to conduct the prediction. The segmentation in the second phase adopts a fixed length. This might be oversimplified because a meaningful pattern in the time series may appear with different lengths. Fixed length segmentation may split the pattern which across more than one subsequence. Thus a segmentation approach with dynamic length is introduced to improve the clustering accuracy.

Since the segmentation is so important, there have been a lot of works on it, e.g. Ref. [18]. This problem can easily be converted to an optimization problem. Genetic algorithm (GA) is a popular optimization algorithm and has been widely used in time series analysis. A research based on GA and enable the user to search predefined patterns in time series has been proposed [19]. Tseng et al. have shown GA to be able to address the time series segment problem to find a meaningful pattern in 2006 [20].

In this paper, we segment the solar radiation series into subsequences with dynamic lengths instead of fixed lengths. Combined with cluster algorithm, GA is used to find the optimal segmentation scheme. The proposed segmentation algorithm generates a population of chromosomes first. Each chromosome represents a possible segmentation of the series. After segmentation, K-means is used to group the subsequences into K different clusters. Then each cluster only contains subsequences of similar feature. To compare subsequence of different length, Dynamic Time Warping (DTW) is used as a distance measurement. The Silhouette index is used to evaluate the accuracy of clustering [21]. The evaluation result of clustering is used as cost function to choose appropriate chromosome for further mutation and mating. Finally, the optimal chromosome is obtained. The whole process of searching optimal segmentation is shown in Fig. 1. Some details will be discussed in the following sections.

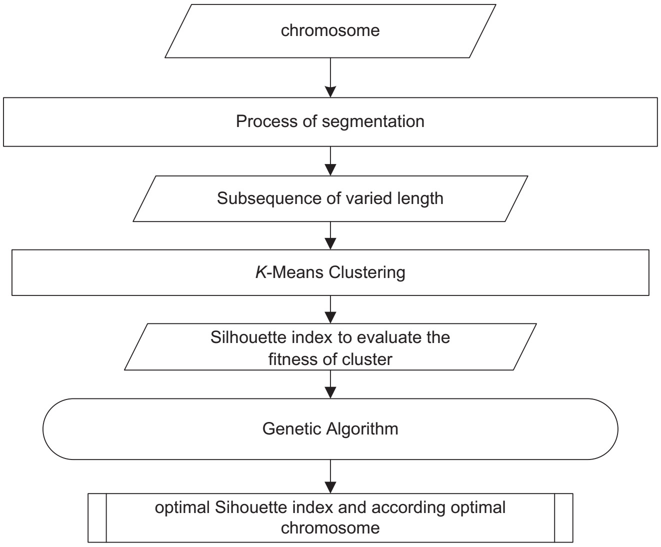

Fig. 1. The flow chart of using GA to find the optimal segment.

2.1.1. Chromosome representation

The segmentation scheme must be encoded as chromosome first. There are two popular encoding methods for GA. One is real-number encode method; the other is bit-string method. In this paper, bit-string encode method is adopted.

A time series can be represented by a vector $\{x_1, x_2, \dots, x_k, \dots, x_n\}$, x_k is the k-th point of the time series. The chromosome is a bit string which has the same length as the time series to process. It could be represented as a vector as $\{c_1, c_2, \dots, c_k, \dots, c_n\}$. c_k is the k-th point of the chromosome. If $c_k = 1$, it means x_k is a segment point. For example, if c_3 and c_7 equals one, that means $\{x_3, x_4, x_5, x_6, x_7\}$ is a subsequence generated by the chromosome. Other segment will be generated similarly. To prevent generating meaningless subsequence, there are two constraints for the chromosome: first, the beginning and ending points of the chromosome must be segment points. Second, there cannot be two adjacent points in chromosome to be 1. This kind of chromosome generates null subsequence. Each chromosome represents a possible segmentation for the time series. Then the genetic algorithm is used to find the optimal chromosome for the segment of the residual series.

2.1.2. Initial population

The genetic algorithm needs an initial population of chromosome to start the iteration process. Each chromosome is a feasible segmentation for the time series. In the experiments, the initial chromosomes are randomly generated ones that satisfy these constrains mentioned above.

2.1.3. Clustering algorithm

To derive a proper pattern of the residual solar radiation series, K-means algorithm is used to group these subsequences into different clusters. Because of adopting a dynamic segment approach, the subsequences are of varying length. Thus when comparing the similarity of two subsequences, they must firstly be converted to the same length. The polynomial interpolation is used to normalize subsequences into identical length. The length is pre-defined. K-means algorithm is used to cluster the normalized subsequences. Each cluster represents a pattern lying in the time series. The number of clusters K also needs to be pre-defined. To evaluate the accuracy of clustering result, the well-known silhouette index is used. Silhouette index has proved its ability in evaluating the performance of clustering in many experiments [21]. $a(i)$ represents the average distance