Applied Soft Computing 48 (2016) 207–216

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel soft computing framework for solar radiation forecasting

Mahmoud Ghofrani [a,*], Mohadeseh Ghayekhloo [b], Rasool Azimi [b]

[a] School of Science, Technology, Engineering and Mathematics (STEM), University of Washington, Bothell, USA
[b] Young Researchers and Elite Club, Qazvin Branch, Islamic Azad University, Qazvin, Iran

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form 13 June 2016
Accepted 14 July 2016
Available online 18 July 2016

Keywords:
Clustering
Data preprocessing
Forecasting
Soft computing
Solar radiation
T.S.C K-means

ABSTRACT

Accurate forecasting of renewable-energy sources plays a key role in their integration into the grid. This paper proposes a novel soft computing framework using a modified clustering technique, an innovative hourly time-series classification method, a new cluster selection algorithm and a multilayer perceptron neural network (MLPNN) to increase the solar radiation forecasting accuracy. The proposed clustering method is an improved version of K-means algorithm that provides more reliable results than the K-means algorithm. The time series classification method is specifically designed for solar data to better characterize its irregularities and variations. Several different solar radiation datasets for different states of U.S. are used to evaluate the performance of the proposed forecasting model. The proposed forecasting method is also compared with the existing state-of-the-art techniques. The comparison results show the higher accuracy performance of the proposed model.

1. Introduction

The number of solar energy plants has increased spectacularly in recent years. However, the intermittent and unpredictable nature of renewable energy such as solar and wind leads to defects on reliability and stability of power grid systems while increasing its integration and operational costs [1]. This requires strategies to enable integrating renewable energy sources with the traditional resources of energy production [2]. The need for forecasting energy generated by these plants has become a decisive factor in competitive power markets. Solar irradiance forecasting requires further investigation to provide more accurate and efficient methods. Therefore, we propose a hybrid forecasting framework to enhance the accuracy of predicting solar photovoltaic (PV) generation to reduce the associated costs to enable a more economic and reliable operation of electric power systems. Analyzing solar irradiance time series using pattern recognition techniques can significantly improve the accuracy of the forecast. Clustering analysis is one of the most common unsupervised machine learning methods used in pattern recognition. The main goal of clustering is to generate compact groups of objects or data that share similar patterns within the same cluster, and isolate these groups from those which contain elements with different characteristics. In the field of renewable energy forecasting, this technique allows handling groups of data separately, which provides a better understanding of the collected information and improves the accuracy of the final forecast results. Several clustering methods have been used to identify patterns and provide a pattern-based prediction technique for renewable energy resources such as solar irradiance [3–19]. In [3] an optimization of interval type-2 fuzzy integrators was proposed for the Mackey-Glass time series forecasting. ANFIS (adaptive neuro-fuzzy inferences systems) models and genetic algorithms (GAs) were combined for this purpose. Type-2 fuzzy logic was also used in [4] to better manage uncertainties of real data for enhancing the learning process of the backpropagation method and improving the forecast accuracy. Two novel evolutionary neural network approaches including sparsely connected evolutionary ANN (SEANN) and time lag feature selection EANN (TEANN) were proposed in [5] for time series forecasting. The SEANN approach provides more flexible ANN structures for multi-step ahead forecasting. The TEANN approach adjusts the ANN parameters with a set of time lags which are fed into the forecasting model. A new time series forecasting model was proposed in [6] which is based on the belief rule (BR) inference methodology. The efficiency of the proposed method depends on both the structure and the belief degrees. Akaike's information criterion (AIC) was used to provide the appropriate delay step to determine the structure. A novel approach based on neuro-fuzzy modeling was proposed in [7] for time series prediction. The training patterns are extracted after selecting the proper relevant variables. A set of Takagi-Sugeno-Kang (TSK) fuzzy rules is constructed based on the extracted training patterns. The

* Corresponding author at: UWBB room 227, 18807 Beardslee Blvd, Bothell, WA 98011, USA.
E-mail address: mghofrani@uwb.edu (M. Ghofrani).

http://dx.doi.org/10.1016/j.asoc.2016.07.022
1568-4946/

related rule parameters are then refined by the learning method. The refined fuzzy rules are then used for time series forecasting. A hybrid of linear autoregressive integrated moving average (ARIMA) and nonlinear artificial neural network (ANN) models was used in [8] for one-step and multi-step ahead time series forecasting. A hybrid model of self-organizing maps (SOM), support vector regression (SVR) and particle swarm optimization (PSO) was proposed in [9] to forecast hourly global solar radiation. SOM algorithm was applied in the first step to divide the entire input space into several disjointed regions or clusters. SVR was then used to model each cluster for detecting characteristic correlation between the predicted and the actual values. In order to deal with the volatility of SVR with different parameters, PSO algorithm was used to improve the forecasting performance of the SVR models. However, the SOM algorithms may converge to non-optimal clustering results depending on the initialization and learning rate considered for the algorithm. In addition, neighborhood violations occur if the output space topology does not match with the data shape. A new model based on fuzzy set theory was developed in [10] for forecasting hourly global solar irradiation. A Fuzzy C-Means clustering was used for establishing the membership functions from the input variable's attributes. Performance and limitations of the model were tested using data from ten stations with latitudes between 40° and 50° north. It was concluded that the accuracy of the model is adequate for forecasting the performance of photovoltaic systems in this band of latitudes. A solar radiation forecast technique based on the fuzzy logic and neural network was proposed in [11]. The clustering process was improved by using fuzzy logic and the number of sky classes was reduced via the clustering characteristics of the uniform process with more intuitive interpretation. In [12] a new time series clustering technique was proposed to reduce the computational complexity of smart grid optimization problems in applications such as demand forecasting and renewable energy prediction. A simple spectral clustering was applied to the household's electricity load and supply data to capture their behavioral patterns. A hybrid forecasting technique was presented in [13] to predict the hourly power output for a photovoltaic (PV) system. The proposed hybrid method combined the generalized radial basis function network (GRBFN), deterministic annealing (DA) clustering technique, weight decay (WD) method and an improved version of the particle swarm optimization (PSO) algorithm for the prediction. SOM algorithms and wavelet neural networks were employed in [14] to develop a short-term PV generation forecasting model. The SOM was used for clustering the weather data to account for seasonal variations, and the wavelet neural network was utilized to build prediction models for each cluster sample. However, limitations of the SOM mentioned above still held true for this study.

A method based on clustering and learning vector quantity (LVQ) was presented in [15] to forecast solar flares. First, the K-means algorithm was used to cluster the dataset with a skewed data distribution. This resulted in more balanced training datasets for LVQ to build a training model that can classify data into class labels indicating whether the solar flare occurs in the next 24 h or not. However, the accuracy of the LVQ method is highly dependent on the initialization of the model and the learning parameters used as well as the class distribution of the training data. A combined K-means clustering and nonlinear auto regressive (NAR) neural network model was introduced in [16] to improve forecast accuracy of the hourly solar irradiance. A forecasting technique based on the K-means clustering algorithm was presented in [17] for wind and PV energy production. However, the initial cluster centroids randomly selected in the first phase of the algorithm may provide incorrect results for K-means clustering. A combination of an enhanced version of K-means algorithm and L-method approach was used in [18] to develop a new forecasting technique that provides maps to determine suitable locations for the deployment of solar power plants. An efficient combination of the affinity propagation (AP) and the K-means clustering algorithms was employed to address the sensitivity of the K-means to the initial cluster centroids. However, the inherent weakness of the AP method for not detecting the non-spherical clusters makes it more effective for the synthetic dataset (which are generally spherical shape) rather than the real dataset. K-means algorithm was combined with cumulative probability distribution functions to forecast the hourly solar irradiance using a daily clearness index [20]. The performance of the K-means based clustering is highly dependent on the initial cluster centroids. The random selection of the initial cluster centroids for the K-means algorithm may result in non-optimal solutions [20]. Different variants of k-means algorithm have been proposed to address this limitation [21–24]. A global K-means algorithm was developed to provide optimal solution for clustering problems [21]. K-means++ initialization algorithm was presented in [22] to obtain an initial set of centroids that is near-optimal. A combination of a new initialization technique, K-means algorithm and a data transformation approach was proposed to solve the K-means clustering problems [23]. A novel data clustering algorithm was introduced in [24] that combines granular theory and Fuzzy C-Means (FCM) clustering algorithm for more reflective data structuring.

In this paper, we propose a hybrid forecasting method that combines a time-series analysis, a novel clustering technique for solar time series data, a new cluster selection algorithm and MLPNNs to predict solar radiations. The paper includes two significances: improving clustering algorithm as an important method in the field of data mining and enhancing solar forecasting as a planning strategy for renewable energy integration. The hybrid forecasting method provides a soft computing framework with the following contributions:

1 An improved version of K-means algorithm, named time series clustering (T.S.C) K-Means, is proposed to provide fixed, definitive clustering results for different runs of the algorithm as opposed to different answers obtained for different runs of the original K-Means.
2 A modified initialization is proposed to select initial centroids which are closer to the optimum centroids' locations to solve the problem of empty cluster generation that frequently occurs due to the random initialization of K-means based clustering methods.
3 A new classification approach is developed to classify hourly solar data based on clustering results of the average daily data to better characterize irregularities and variations of solar radiation. This leads to deep learning of neural networks to provide more accurate forecasts as compared to direct clustering of hourly data. In addition, using the average daily data rather than the hourly data in the classification process significantly decreases the associated computational time.
4 After splitting the training data into multiple sub-trains by the proposed classification approach, a new method based on Principal Component Coefficients (PCCs) and correlation analysis is used to select the most appropriate sub-train as the input to the neural network. This significantly accelerates the forecast process as the most appropriate portion of the data rather than the whole data is used for the NN training. The proposed technique is particularly important for very short-term forecasting where the forecast horizon can be as short as a few seconds ahead.

The proposed clustering algorithm addresses the limitations of the K-means based clustering and provides a faster processing than the existing state-of-the-art clustering methods. It also offers more accurate clustering results than the original K-means algorithm and its variants on real time-series data such as solar data. The envisioned widespread application of the developed clustering in pattern recognition constitutes the desired and expected broader impact of the proposal. The proposed forecasting method facilitates the solar power integration by reducing the forecast errors associated with the existing forecasting methods. The developed forecasting framework enables independent system operators (ISOs), generation companies (GENCOs) and electric utility systems to establish more appropriate operational practices and bidding strategies and to schedule adequate energy transactions. The associated energy savings also benefit the end users by reducing the integration and operational costs of renewable integration.

The rest of the paper is organized as follows. Section 2 provides a brief description of the K-means algorithm. It also explains the proposed T.S.C K-means, and the hybrid solar irradiance forecasting method. Section 3 demonstrates a case study where the clustering errors are calculated for different algorithms. The performance of the developed forecasting method is also evaluated with respect to different aspects including different seasons with different kinds of weather conditions and different prediction horizons (for 1 h, 24 h, and 48 h) in this section. Finally, Section 4 concludes the paper.

2. Methodology

2.1. K-means algorithm

K-means is a well-known, low complexity algorithm utilized for data-partitioning [25]. The algorithm starts running after an input of K clusters is given. It outputs the cluster centroids through iterations. Let $X = \{x_1, x_2, \ldots, x_N\}$ be the set of N points to be grouped into K cluster sets as $C = \{C_k\}$ $k = 1, 2, \ldots, K$. Using the Euclidean distance, the algorithm assigns each data point to its closest centroid C_k. After the first run, the algorithm calculates the mean of the data points in each cluster c_k using (1). It then starts a new iteration by selecting this value as a new cluster centroid. Once new clusters are selected, a new mean value is also obtained.

$$c_k = \left(\frac{1}{n_k}\right) \cdot \sum_{i=1}^{n_k} x_i^{(k)} \tag{1}$$

$X_i^{(k)}$ is the i-th data point and N_k is the number of data points in cluster k.

The algorithm halts once the sum of the squared error over K clusters is minimized. However, with respect to its computational complexity and the initialization step, the K-means approach presents some limitations [26].

i) The number of clusters, k, must be known before the first iteration of the algorithm.
ii) The K-means algorithm is very sensitive to the initial cluster centroids. The more the selected initial clusters are distant from the optimal cluster centroids, the more iteration will take for the algorithm to converge.
iii) K-means is also strongly sensitive to noisy data, which may affect the accuracy of the final forecasts [27].

2.2. The proposed initialization technique for K-means

An improved version of K-means algorithm, time series clustering (T.S.C) K-means, is proposed in this section. The proposed clustering algorithm uses a new technique to select the initial cluster centroids to solve the shortcoming of the existing K-means algorithms and provide a better performance.

Let $X = [x_1,\ldots,x_n]$ be a set of n data vectors. The steps of the proposed initialization technique to select K initial centroids are as follows:

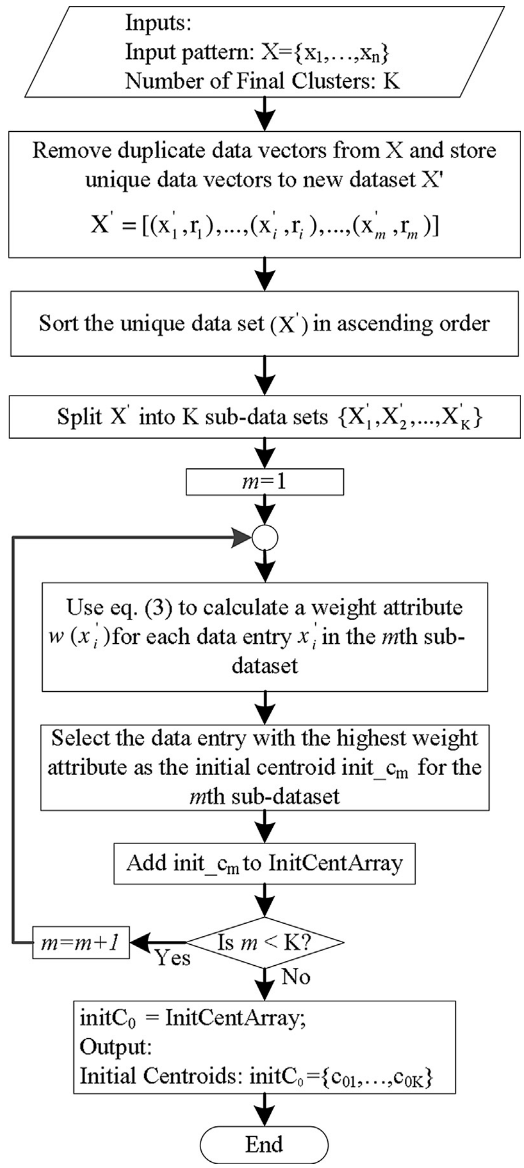

Fig. 1. Flowchart of the proposed method for the initial centroids selection.

1. Remove duplicate data and store them to new dataset $\mathbf{X}' = [(\mathbf{x}_1', \mathbf{r}_1), \ldots, (\mathbf{x}_m', \mathbf{r}_m)]$ where r_i is the repetition number for each non-repetitive data vector $(\mathbf{x}_i)$ in the new dataset $\mathbf{X}'$ $(i \le m \le n)$.

2. Sort the data vectors in the dataset X' in ascending order based on the length (norm) of the vector [28]. The vector length is the Euclidean length or Euclidean norm.

3. Divide the dataset $\mathbf{X}'$, consisting of m data, into K sub-datasets, with (at most) $P = \lceil m/K \rceil$ data, according to Eq. (2), such that the data elements of $\mathbf{X}'$ are distributed among the sub-datasets $\mathbf{X}_1'$ to $\mathbf{X}_K'$.

$$\begin{aligned}
&\mathbf{X}_1' = [(\mathbf{x}_1', \mathbf{r}_1), \ldots, (\mathbf{x}_P', \mathbf{r}_P)],\\
&\mathbf{X}_2' = [(\mathbf{x}_{P+1}', \mathbf{r}_{P+1}), \ldots, (\mathbf{x}_{2P}', \mathbf{r}_{2P})],\\
&\vdots\\
&\mathbf{X}_K' = [(\mathbf{x}_{(K-1)\times(P)+1}', \mathbf{r}_{(K-1)\times(P)+1}), \ldots, (\mathbf{x}_{KP}', \mathbf{r}_{KP})].\\
&\mathbf{X}' = \bigcup_{k=1}^{K} \mathbf{X}_k'
\end{aligned} \tag{2}$$

4. Now, we have K sub datasets where each one is used to determine only one of the K initial centroids. Eq. (3) is used to calculate a weight attribute $w(x_i^{'})$ for each data entry $x_i^{'}$ with the repetition number r_i in each of the K sub datasets$\{\mathbf{X}_1^{'}, \mathbf{X}_2^{'}, \ldots, \mathbf{X}_K^{'}\}$.

$$w(x_i^{'})_m = \frac{1}{\frac{1}{P}\sum_{j=1}^{\mathbf{P}} dist(x_i^{'}, x_j^{'})} \cdot (r_i)_m, (1 \le m \le K) \tag{3}$$

where $w(x_i^{'})_m$ is the weight attribute for $x_i^{'}$ in the m-th sub-dataset.

5. In each of the K sub datasets, the data entry with highest weight attribute is elected as the initial centroid.

Fig. 1 shows the flowchart of our proposed method for selecting initial centroids.

The proposed initialization approach guarantees that the K-means algorithm does not generate empty clusters during data clustering process. The Silhouette method [29] is used to estimate the number of cluster K in the proposed approach.

Solar data cannot be adequately characterized by clustering alone. This is due to irregular variations of the solar data and its dependence on meteorological parameters that are highly intermittent. The solar data could be better characterized by classifying the data in advance to their clustering. An efficient data classification approach is proposed in Section 2.3 (steps ii–iv) to provide a representative data for each day rather than the 24 hourly data. The classified data are then used as the inputs for clustering. The clustering outputs are further processed according to steps v–ix to provide the forecasting results.

2.3. The proposed hybrid solar forecasting method

Neural networks and machine learning as the principal constituents of soft computing are useful for applications such as forecasting, classification, clustering and optimization. A hybrid soft computing framework based on neural networks and machine learning approaches is therefore developed in this section for solar forecasting. The proposed method uses the T.S.C K-Means, a new hourly time-series classification method based on clustering results of the daily data, a cluster selection algorithm and a MLPNN. Fig. 2 shows the flowchart of the proposed forecasting method. The proposed hybrid forecasting is outlined as follows:

i) The hourly global horizontal irradiance (GHI) or direct normal irradiance (DNI) is used as the input data with 80% allocated for training and the remaining 20% for testing.

ii) The training data are classified into 24-h daily data to form the hourly data package for day one to n where n is the total number of days. The average of the hourly data for each day (i) is then calculated as the day's representative by:

$$DailyAvg_i = 1/24\sum_{t=1}^{24} hourlydata_t^i \tag{4}$$

where $hourlydata_t^i$ is the solar radiation data for hour t of day i.

iii) The proposed T.S.C K-means algorithm groups the daily average data into K clusters with $n_1, n_2, \ldots, n_k$ numbers of data within clusters $1, 2, \ldots, K$, respectively where $n_1 + n_2 + \ldots + n_k = n$. $DailyAvg_{n_k}^k$ denotes the n_k-th representative value in cluster K.

iv) Clusters 1 to K are classified as sub-trains 1 to K by including the associated hourly data in each cluster. $HourlyDataPackage_{n_k}^k$ represents the 24 h data for the n_k-th day within cluster K.

v) For each sub-train and testing input data, the length of the sliding window used for the lagged data ($L = 5$) is calculated. The calculated sliding window is then used to construct the time series structure for the hourly data within each sub-train. The data include N rows of training and N' rows of testing.

vi) The principal component coefficients (PCCs) are calculated for the training input of each sub-train and the testing input as follows:

By considering X as an $(M \times N)$matrix whose (i, j)th element is defined with $i = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, N$, the covariance matrix C with dimension $N \times N$ is formed by [30]:

$$\mathbf{C} = \sum_{i=1}^{M} \mathbf{X}_i \mathbf{X}_i^{\mathbf{T}} \tag{5}$$

M is the total number of observations, N is the number of variables and T represents an implicit transposition. Eigenvalues of C (λ) are calculated by solving Eq. (6) where I is the identity matrix.

$$|\mathrm{C} - \lambda \mathrm{I}| = 0 \tag{6}$$

The eigenvector (E) of the covariance matrix C is calculated as follows:

$$\mathrm{CE} = \lambda \mathrm{E} \tag{7}$$

where, E is an $N \times N$ matrix. The PC (Z) produces linear combinations of the original variables. The original variables are determined taking the elements of X and calculating the coefficients of Z as follows [31].

$$\mathrm{Z} = \mathrm{XE} \tag{8}$$

The Principal Component Coefficients (PCCs) determine the influence of each variable $x_{i,1}, x_{i,2}, \ldots, x_{i,j}$on each principal component $z_{i,1}, z_{i,2}, \ldots, z_{i,j}$ (Eq. (8)). Rows of the PCC matrix correspond to the variables while its columns correspond to the components. Elements of the PCC matrix are calculated by [30]:

$$\mathrm{PCC(i,j)} = \frac{e_{i,j}}{\sqrt{\mathrm{Var}(\mathrm{x}_{i,j})}} \tag{9}$$

where $\mathrm{e}_{i,j}$ represents the element of the i^{th} row and j^{th} column in the eigenvector matrix E and Var $(\mathrm{x}_{i,j})$ represents the variance of $\mathrm{x}_{i,j}$.

The PCCs for the training input matrix for the k-th subtrain $(Training - Input^k)$ and the input matrix for the testing data $(Testing - Input)$ are calculated as follows:

$$\begin{aligned} \mathbf{PCC} - Tr_{L\times L} &= \mathrm{PCC}(Training - Input_{N\times L}^k) \\ \mathbf{PCC} - Ts_{L\times L} &= \mathrm{PCC}(Testing - Input_{N'\times L}) \end{aligned} \tag{10}$$

vii) For each sub-train, the correlation coefficient between the calculated PCCs is determined by:

$$\mathrm{CC}^k = abs(Corrcoeff(PCC - Tr_{L\times L}, PCC - Ts_{L\times L})) \tag{11}$$

viii) The algorithm is repeated for all sub-trains and the subtrain with the maximum coefficient is selected as the best training input into the MLPNN.

ix) Once the MLPNN is trained using the selected training data, it uses the solar radiation testing input data as well as the air temperature, wind speed, wind direction and extraterrestrial radiation data to calculate the testing outputs as forecasts. The forecasts are represented by $S(t+1), S(t+2), \ldots, S(t+N')$ in Fig. 2.

3. Case studies

In this section, we first examine the performance of the proposed T.S.C K-means clustering algorithm and then evaluate the developed hybrid solar forecasting method for 1 h, 24 h, and 48 h prediction horizons. The datasets used in the experiment are available online [32].

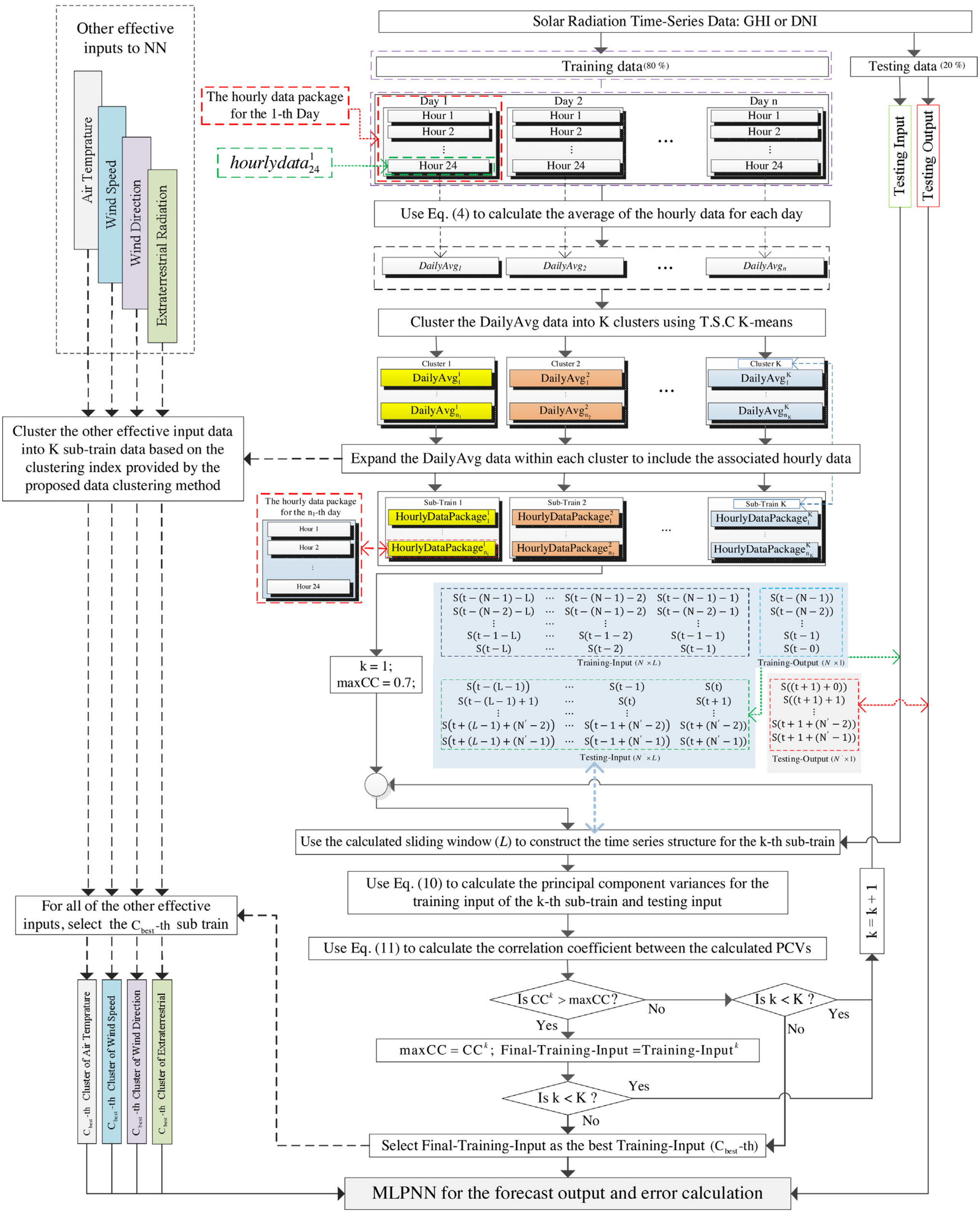

Fig. 2. The proposed forecasting method.

3.1. Evaluation of the proposed T.S.C K-means algorithm

This section evaluates the accuracy of the proposed clustering method (T.S.C K-means). More information regarding the data collection is presented in Fig. 3. The mean squared error (MSE) is used to determine clustering errors as follows:

$$MSE = \frac{1}{K.N}\sum_{k=1}^{K}\sum_{i=1}^{N}\|X_i^{(k)} - C_k\|^2 \quad (12)$$

where N is the number of data points in cluster k, and $X_i^{(k)}$ is the ith data point in cluster k. It should be mentioned that the testing data are normalized in the range of [−1, 1]. Table 1 shows the MSE

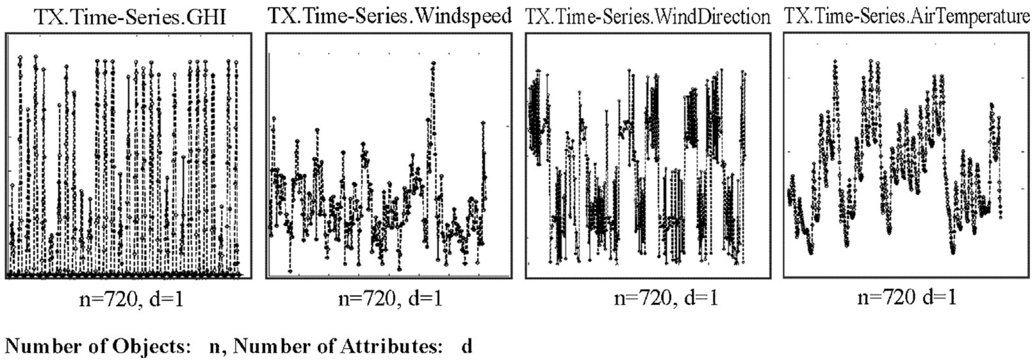

Fig. 3. Datasets used to compare the performance of data clustering techniques.

Table 1
MSE measures for different clustering techniques (The best results are shown in bold).

Dataset	T.S.C K-means	K-means*	K-means++	K-means	SOM	FCM	K-Medoids
TX-GHI	**0.0044**	0.0047	0.0084	0.0063	0.0075	0.0049	0.0192
TX-W.S	**0.0079**	0.0091	0.0130	0.0115	0.0119	0.0085	0.0362
TX-W.D	**0.0102**	0.0137	0.0164	0.0157	0.0139	0.0114	0.0419
TX-A.T	**0.0129**	0.0148	0.0155	0.0154	0.0167	0.0132	0.0132

values for different clustering techniques including the proposed T.S.C K-means, K-means*, K-means++, K-means, K-Medoids, SOM [33], and FCM [34]. The calculated MSE values show that the T.S.C K-means algorithm improves the quality of clustering in comparison with other clustering methods. This guarantees the efficiency of the developed approach for the solar forecasting.

3.2. Evaluation of the proposed solar forecasting method

This section evaluates the developed hybrid solar forecasting method with the proposed T.S.C K-means algorithm. The proposed forecasting is tested on several different solar datasets to provide a comprehensive performance analysis. The forecast accuracy is evaluated by the mean absolute error (MAE), normalized MAE (nMAE), root mean square error (RMSE) and normalized RMSE (nRMSE) as follows:

$$MAE = \frac{1}{K'}\sum_{n=1}^{K'} |\hat{S}(n) - S_{Actual}(n)| \tag{13}$$

$$nMAE(\%) = \left({}^{MAE}\!/\!{S_{max} - S_{min}} \right) \times 100 \tag{14}$$

$$RMSE = \sqrt{\frac{1}{K'}\sum_{n=1}^{K'} \left(\hat{S}(n) - S_{Actual}(n)\right)^2} \tag{15}$$

$$nRMSE(\%) = \left({}^{RMSE}\!/\!{S_{max} - S_{min}} \right) \times 100 \tag{16}$$

where K' is the total number of testing outputs, $\hat{S}(n)$ and $S_{Actual}(n)$ are the solar radiation forecast and the actual solar radiation for hour n. The datasets are from different states with different solar radiation characteristics. More information regarding the solar radiation data is presented in Table 2. Fig. 4 shows the performance of the proposed solar radiation forecasting model for spring, summer, fall and winter between 2006 and 2009 in Arizona State, based on RMSE measures. Table 3 provides the performance indicators of the GHI and DNI forecasts, for different forecast horizons (1-h, 24-h and 48-h ahead). The performance of the proposed forecasting method with the daily data-based clustering is compared to the results of the same forecasting method with direct clustering of hourly data. Table 4 provides the performance indicators and the processing time of the proposed forecasting method using this two

Table 2
The solar radiation data used for the evaluation of the proposed forecasting method.

State	Longitude	Latitude
Arizona (AZ)	112° W	34° N
California (CA)	120° W	37° N
Colorado (CO)	105° W	39° N
Iowa (IA)	93° W	42° N
Kansas (KS)	98° W	38.5° N
Michigan (MI)	85.01° W	43.6° N
Montana (MT)	110° W	47° N
Nevada (NV)	117° W	39° N
New Mexico (NM)	106° W	34° N
New York (NY)	74° W	40.71° N
North Dakota (ND)	100° W	47° N
Oklahoma (OK)	98° W	35.5° N
Oregon (OR)	120.5° W	44° N
South Dakota (SD)	100° W	44.5° N
Texas (TX)	100° W	31° N
Washington (WA)	120.5° W	47.5° N

clustering techniques. The accuracy performance of the forecasting method with the daily data clustering is significantly better than that of the direct hourly clustering. Also, the proposed daily data clustering has a faster processing time compared to the forecasting with the direct hourly clustering. Table 5 provides a comparative analysis of different latest forecasting models available in the literature with our proposed forecasting method. Utilizing the proposed method, the RMSE and MAE parameters obtained were notably smaller than the error parameters for the other methodologies, for different time horizons.

Figs. 5 and 6 show the GHI forecasts for two different weeks in winter (01/19/2009–01/25/2009) and summer (07/01/2009–07/07/2009) in Arizona State, respectively. These weeks contain different weather conditions including clear, cloudy and overcast sky. The figures demonstrate the efficiency of the proposed method to forecast the solar radiation in different weather conditions.

Fig. 7 shows the comparisons between the proposed solar forecasting method and six other well-established forecasting models including the Persistence method with clearness index [39,40], NN [41], NAR-NET (nonlinear autoregressive neural network) [16], ARIMA, RW (random-walk) and Hybrid method [9]. The solar radiation data in Washington USA [32] is used for the comparative

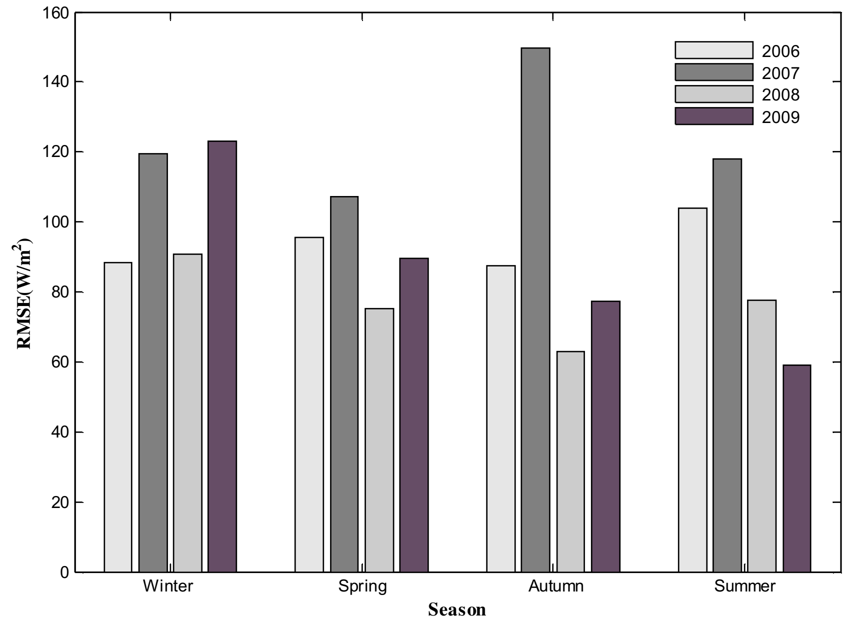

Fig. 4. RMSE (W/m^2) measures of the proposed forecasting method for different seasons between 2006 and 2009.

Table 3
Performance indicators for the forecast with different horizons.

Dataset		Forecast Horizon											
		1 h				24 h				48 h			
		RMSE (W/m^2)	nRMSE	MAE(W/m^2)	nMAE	RMSE(W/m^2)	nRMSE	MAE(W/m^2)	nMAE	RMSE(W/m^2)	nRMSE	MAE(W/m^2)	nMAE
AZ	GHI	102.989	0.097	60.720	0.057	110.351	0.104	43.754	0.041	120.247	0.113	54.962	0.052
	DNI	125.81	0.1218	74.57	0.0722	227.92	0.2206	136.44	0.1321	231.88	0.2245	139.33	0.1349
CA	GHI	67.933	0.066	38.111	0.037	111.315	0.108	44.814	0.044	122.122	0.119	49.407	0.048
	DNI	122.03	0.1271	71.93	0.0749	272.21	0.2836	165.36	0.1723	284.17	0.296	175.30	0.1826
CO	GHI	107.409	0.101	60.391	0.057	140.582	0.133	67.769	0.064	151.196	0.143	73.406	0.069
	DNI	160.86	0.1488	101.53	0.0939	276.06	0.2554	190.05	0.1758	272.84	0.2524	191.48	0.1771
IA	GHI	70.875	0.073	38.367	0.039	136.973	0.140	66.352	0.068	140.522	0.144	72.544	0.074
	DNI	107.67	0.1152	62.04	0.0664	249.26	0.2666	168.96	0.1807	254.10	0.2718	170.22	0.1821
KS	GHI	74.812	0.074	36.321	0.036	127.291	0.127	64.182	0.064	143.794	0.143	70.370	0.070
	DNI	111.63	0.1157	64.79	0.0671	230.77	0.2391	145.74	0.151	233.94	0.2424	147.23	0.1526
MI	GHI	73.997	0.077	38.264	0.040	129.961	0.135	63.798	0.066	136.752	0.142	69.940	0.073
	DNI	106.32	0.1135	61.72	0.0659	235.36	0.2512	153.89	0.1642	244.95	0.2614	165.76	0.1769
MT	GHI	82.029	0.084	42.350	0.043	123.687	0.126	66.609	0.068	131.150	0.134	72.053	0.074
	DNI	130.68	0.1298	79.16	0.0786	270.36	0.2685	195.04	0.1937	266.52	0.2647	180.85	0.1796
NV	GHI	93.617	0.088	48.963	0.046	112.275	0.106	52.493	0.049	119.254	0.112	56.163	0.053
	DNI	146.67	0.1386	89.89	0.085	268.28	0.2536	173.19	0.1637	274.55	0.2595	179.17	0.1693
NM	GHI	97.479	0.091	53.241	0.050	122.898	0.115	60.970	0.057	121.427	0.114	58.786	0.055
	DNI	147.39	0.1408	89.80	0.0858	337.71	0.3226	144.31	0.1378	242.79	0.2319	144.87	0.1384
NY	GHI	99.761	0.103	50.670	0.052	134.805	0.139	68.774	0.071	138.444	0.143	69.774	0.072
	DNI	118.76	0.1263	70.42	0.0749	239.81	0.2551	163.92	0.1744	257.82	0.2743	177.68	0.189
	GHI	79.020	0.083	42.868	0.045	116.204	0.121	56.542	0.059	122.186	0.128	63.289	0.066
	DNI	127.68	0.1325	82.71	0.0858	251.96	0.2614	172.58	0.179	254.29	0.2638	182.21	0.189
OK	GHI	74.770	0.075	37.357	0.037	139.870	0.139	67.753	0.068	148.293	0.148	75.994	0.076
	DNI	119.10	0.1251	69.83	0.0734	233.32	0.2451	142.45	0.1496	232.73	0.2445	151.48	0.1591
OR	GHI	75.185	0.075	41.439	0.041	103.845	0.103	47.573	0.047	111.349	0.111	52.373	0.052
	DNI	148.07	0.1475	89.30	0.0889	248.20	0.2472	169.40	0.1687	256.55	0.2555	173.97	0.1733
SD	GHI	73.588	0.076	39.441	0.041	126.503	0.130	62.777	0.065	139.296	0.143	70.884	0.073
	DNI	121.85	0.1255	69.73	0.0718	240.83	0.248	161.40	0.1662	247.78	0.2552	166.051	0.171
TX	GHI	99.971	0.098	57.689	0.056	129.488	0.127	66.611	0.065	155.955	0.152	82.504	0.081
	DNI	114.38	0.1182	63.24	0.0653	226.08	0.2336	132.43	0.1368	221.08	0.2284	137.33	0.1419
WA	GHI	68.986	0.072	36.011	0.038	97.979	0.103	47.674	0.050	109.349	0.115	55.845	0.059
	DNI	127.81	0.1353	81.55	0.0863	225.77	0.2389	154.21	0.1632	231.64	0.2451	157.28	0.1664

Table 4
Performance indicators and processing time for the GHI forecast (1-h ahead) with the daily data clustering and direct hourly clustering. (The best results are shown in bold)

GHI.	Forecast with									
	Direct hourly clustering					Proposed daily data clustering				
	RMSE $\left(W/m^2\right)$	nRMSE	MAE $\left(W/m^2\right)$	nMAE	Processing Time	RMSE $\left(W/m^2\right)$	nRMSE	MAE $\left(W/m^2\right)$	nMAE	Processing Time
AZ	200.782	23.8742	116.7348	13.8805	12.6713	**49.941**	**5.9383**	**23.6117**	**2.8076**	**3.9115**
CA	171.7999	23.6313	86.6763	11.9225	14.033	**52.0548**	**7.1602**	**25.5079**	**3.5086**	**6.1507**
CO	178.5665	23.1304	97.6121	12.644	14.3672	**66.5337**	**8.6184**	**39.5335**	**5.1209**	**5.8368**
IA	130.5156	18.8334	59.5469	8.5926	12.9072	**52.4784**	**7.5726**	**28.0763**	**4.0514**	**6.6521**
KS	308.5785	40.4959	163.5829	21.4676	14.2213	**43.1674**	**5.665**	**21.6986**	**2.8476**	**6.0062**
MI	117.8845	20.0825	61.8544	10.5374	13.4598	**42.8883**	**7.3064**	**21.2479**	**3.6197**	**6.7977**
MT	117.6107	18.2909	56.4126	8.7733	12.5615	**46.4911**	**7.2303**	**23.8873**	**3.715**	**6.9413**
NV	185.8928	23.6806	97.2405	12.3873	16.8652	**70.314**	**8.9572**	**38.951**	**4.9619**	**6.8819**
NM	253.1594	29.5058	141.2716	16.4652	20.117	**47.712**	**5.5608**	**27.0436**	**3.1519**	**6.4584**
NY	115.3155	16.9084	56.7314	8.3184	14.364	**55.5169**	**8.1403**	**33.6076**	**4.9278**	**6.0845**
ND	113.5054	18.3666	49.5848	8.0234	18.0883	**71.5072**	**11.5707**	**37.0724**	**5.9988**	**6.0963**
OK	201.5174	25.5733	110.1208	13.9747	13.0365	**49.1464**	**6.2369**	**22.803**	**2.8938**	**6.0525**
OR	143.2477	21.8033	73.5785	11.1992	14.2129	**51.0227**	**7.766**	**27.5721**	**4.1967**	**6.4683**
SD	115.9315	18.82	55.5836	9.0233	19.3708	**46.7538**	**7.5899**	**23.2042**	**3.7669**	**6.8596**
TX	224.4828	26.6607	122.8069	14.5851	12.4474	**46.3416**	**5.5038**	**23.0707**	**2.74**	**6.5654**
WA	107.3748	18.0766	54.8879	9.2404	20.9984	**46.7262**	**7.8664**	**24.4495**	**4.1161**	**8.6253**

Table 5
A summary of the forecast results for several forecasting methodologies.

Author(s)	Model(s)	Forecast Horizon (Hour)	Performance Measures		Location of data used
			RMSE(W/m2)	MAE(W/m2)	
Gala et al. [35]	ECMWF: The European Center for Medium–Range Weather Forecast	1	–	30.22	Galicia in Spain's north western coast
		3	–	65.92	
		24	–	243.09	
	D-SVR: Post-processing ECMWF with SVR at daily resolution added to 1-h resolution using CSM (Climate system model)	1	–	28.00	
		3	–	59.21	
		24	–	203.89	
	3H-SVR: Post-processing ECMWF with SVR at 3-h resolution added to 1-h resolution using CSM	1	–	28.00	
		3	–	58.76	
		24	–	218.05	
Perez et al. [36]	An empirical correlation between the sky cover data and measured GHI (Global horizontal irradiance)	1	80	–	Desert Rock SURFRAD station
		2	88	–	
		3	96	–	
		4	104	–	
		5	116	–	
		6	142	–	
		24	125	–	
		48	139	–	
Cornaro et al. [37]	ECMWF-NWP: A global forecast model used for NWPs	24(Overcast)	127	–	The ESTER Laboratory of the University of Rome
	STNN: Statistical model based on ANN		233	–	
	ECMWF-MOSNN: Developed model based on ANN and ECMWF NWP data		138	–	
	ECMWF – NWP	24 (Clear sky)	103	–	
	STNN		93	–	
	ECMWF-MOSNN		57	–	
Manjili and Niknamfar [38]	NREL (National Renewable Energy Laboratory) Golden CO	24	149.29	98.09	Golden CO, USA
	BLUE	24	99	61	Austrian dataset
	ECMWF-OL (Oldenburg)	24	101	65	
	The traditional synoptic model of the meteorologists of Blue Sky	24	112	70	
Proposed Method	A combination of a novel clustering technique, a new hourly time-series classification, a new cluster selection algorithm and MLPNN	1 (Clear sky)	57.556	23.118	NREL
		1 (Cloudy sky)	64.856	31.369	
		1 (Overcast)	69.451	33.72	
		24 (Clear sky)	71.181	39.656	
		24 (Cloudy sky)	97.126	55.502	
		24 (Overcast)	106.15	59.082	
		48 (Clear sky)	127.615	68.313	
		48 (Cloudy sky)	131.96	75.197	
		48 (Overcast)	139.28	77.32	

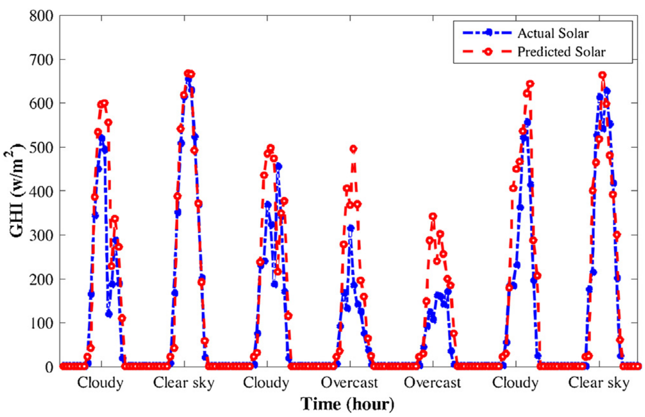

Fig. 5. Forecast results for the winter week (01/19/2009–01/25/2009).

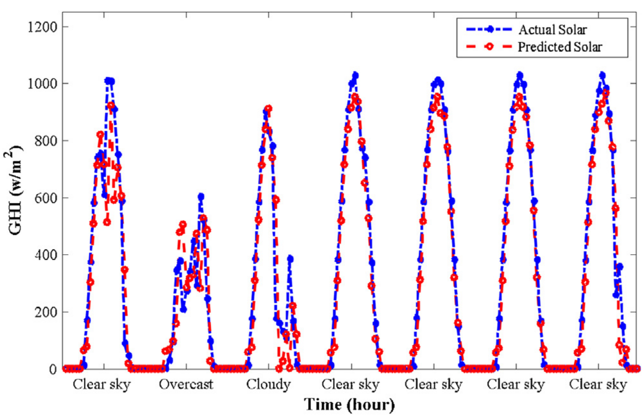

Fig. 6. Forecast results for the summer week (07/01/2009–07/07/2009).

analysis in Fig. 7. The results are the average RMSE (W/m^2) for year 2014. The comparison demonstrates that the proposed method provides more accurate forecasting results than the other forecasting models.

4. Conclusions

A hybrid forecasting method is developed in this paper that combines a novel clustering technique, a new hourly time-series classification method, an innovative cluster selection algorithm and MLPNNs to predict solar radiations. The silhouette algorithm is used to estimate the number of clusters. The combination of the proposed T.S.C K-means clustering and the new hourly time-series classification method provides clusters that give higher resolution in preparation for input into a MLPNN. The cluster whose data has the highest correlation with the testing data is determined by the cluster selection method to provide the training data for the solar radiation forecasting of each individual hour. The appropriate sub-train of solar radiation data as well as the temperature, wind speed, wind direction and extraterrestrial radiation data are trained by the MLPNN to provide the solar radiation forecasting result. The performance of the proposed T.S.C K-means is evaluated using several different types of datasets and compared with different variants of K-means clustering as well as K-medoids, SOM and FCM clustering algorithms. The comparison demonstrates the improved quality of the clustering for the proposed T.S.C K-means. Solar datasets with different solar characteristics and variations are used to determine the accuracy of the proposed forecasting method with T.S.C K-means. The proposed hybrid method is also compared with the existing state-of-the-art techniques for their forecast accuracies. The results show a significant accuracy improvement for the proposed forecasting model.

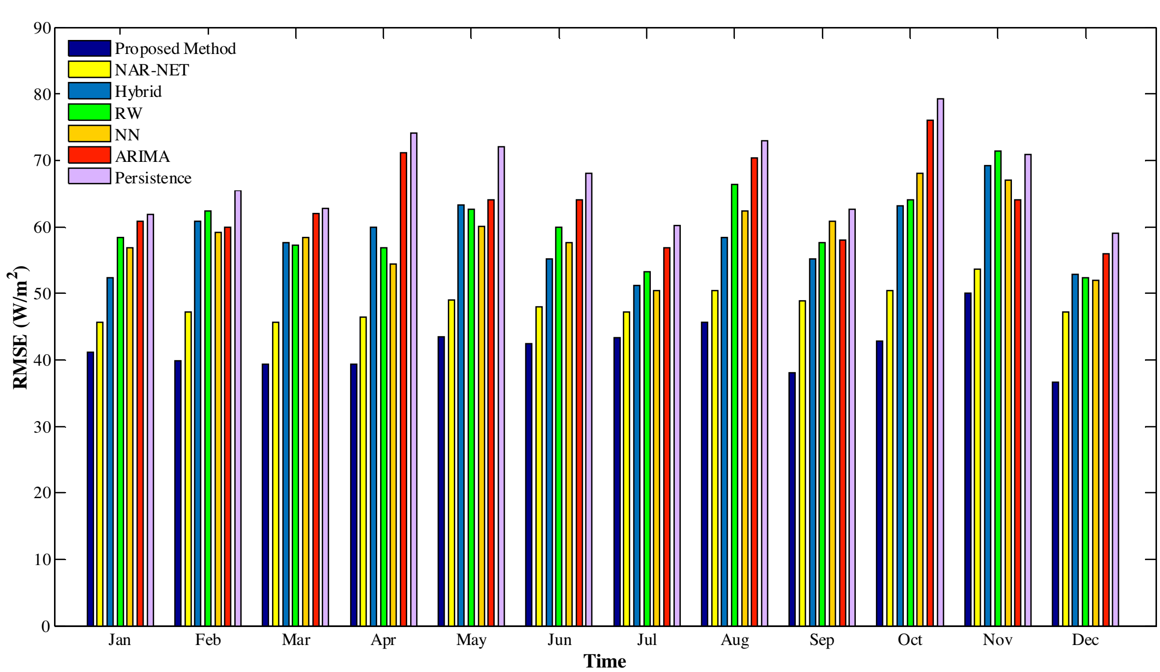

Fig. 7. Accuracy results of several forecasting methods for the solar radiation data in Washington (2014).

References

[1] R.H. Inman, H.T.C. Pedro, C.F.M. Coimbra, Solar forecasting methods for renewable energy integration, Energy Combust. Sci. vol. 39 (2013) 535–576.
[2] M.A. Sanz-Bobi, Use Operation and Maintenance of Renewable Energy Systems Experiences and Future Approaches, Springer, 2014.
[3] J. Soto, P. Melin, O. Castillo, Optimization of interval type-2 and type-1 fuzzy integrators in ensembles of ANFIS models with genetic algorithms, 2013 Mexican International Conference on Computer Science (2013) 99–104.
[4] F. Gaxiola, P. Melin, F. Valdez, O. Castillo, Interval type-2 fuzzy weight adjustment for backpropagation neural networks with application in time series prediction, Inf. Sci. vol. 260 (2014) 1–14.
[5] J. Peralta Donate, P. Cortez, Evolutionary optimization of sparsely connected and time-lagged neural networks for time series forecasting, Appl. Soft Comput. vol. 23 (2014) 432–443.
[6] B.-C. Zhang, X.-X. Han, Z.-J. Zhou, L. Zhang, X.-J. Yin, Y.-W. Chen, Construction of a new BRB based model for time series forecasting, Appl. Soft Comput. vol. 13 (2013) 4548–4556.
[7] H.-W. Peng, S.-F. Wu, C.-C. Wei, S.-J. Lee, Time series forecasting with a neuro-fuzzy modeling scheme, Appl. Soft Comput. vol. 32 (2015) 481–493.
[8] C.N. Babu, B.E. Reddy, A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data, Appl. Soft Comput. vol. 23 (2014) 27–38.
[9] Z. Dong, D. Yang, T. Reindl, W.M. Walsh, A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance, Energy 82 (2015) 570–577.
[10] R.S. Boata, P. Gravila, Functional fuzzy approach for forecasting daily global solar irradiation, Atmos. Res. vol. 112 (2012) 79–88.
[11] S.X. Chen, H.B. Gooi, M.Q. Wang, Solar radiation forecast based on fuzzy logic and neural networks, Renew. Energy vol. 60 (2013) 195–201.
[12] N. Sengupta, S. Aloka, B. Narayanaswamy, H. Ismail, S. Mathew, Time series data mining for demand side decision support, Innov. Smart Grid Technol. – Asia (ISGT Asia (2013).
[13] H. Mori, M. Takahashi, Development of GRBFN with Global Structure for PV Generation Output Forecasting, Power and Energy Society General Meeting, IEEE, 2012.
[14] Y. Ying, D. Lei, Short-Term PV generation system direct power prediction model on wavelet neural network and weather type clustering, Intell. Hum.-Mach. Syst. Cybern. (IHMSC (2013).
[15] R. Li, H. Wang, Y. Cui, X. Huang, Solar flare forecasting using learning vector quantity and unsupervised clustering techniques, Sci. China Phys. Mech. Astron. vol. 54 (no. 8) (2011) 1546–1552.
[16] K. Benmouiza, A. Cheknane, Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models, Energy Convers. Manage. vol. 75 (2013) 561–569.
[17] A. Di Piazza, M. Di Piazza, A. Ragusa, G. Vitale, Environmental data processing by clustering methods for energy forecast and planning, Renew. Energy vol. 36 (no. 3) (2011) 1063–1074.
[18] A. Zagouras, H.T.C. Pedro, C.F.M. Coimbra, Clustering the solar resource for grid management in island mode, Sol. Energy vol. 110 (2014) 507–518.
[19] P. Jiménez-Pérez, L. Mora-López, in: H. Blockeel, M. van Leeuwen, V. Vinciotti (Eds.), Modeling Daily Profiles of Solar Global Radiation Using Statistical and Data Mining Techniques, Advances in Intelligent Data Analysis XIII, vol. 8819, Springer International Publishing, 2014, pp. 155–166.
[20] J.A. Lozano, J.M. Pena, P. Larranaga, An empirical comparison of four initialization methods for the K-Means algorithm, Pattern Recognit. Lett. vol. 20 (no. 10) (1999) 1027–1040.
[21] A. Likas, N. Vlassis, J.J. Verbeek, The global k-means clustering algorithm, Pattern Recognit. vol. 36 (no. 2) (2003) 451–461.
[22] D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, Louisiana, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
[23] M. Malinen, R. Mariescu-Istodor, P. Fränti, K-means*: clustering by gradual data transformation, Pattern Recognit. vol. 47 (no. 10) (2014) 3376–3386.
[24] E. Rubio, O. Castillo, A new proposal for a granular fuzzy C-Means algorithm, in: P. Melin, O. Castillo, J. Kacprzyk (Eds.), Design of Intelligent Systems Based on Fuzzy Logic, Neural Networks and Nature-Inspired Optimization, Springer International Publishing, Cham, 2015, pp. 47–57.
[25] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, Berkeley, Calif University of California Press, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
[26] K.A. Abdul Nazeer, M.P. Sebastian, Improving the accuracy and efficiency of the k-means clustering algorithm, Proceedings of the World Congress on Engineering (WCE) vol. 1 (July 1–3) (2009).
[27] M.E. Celebi, H.A. Kingrav, P.A. Vela, A comparative study of efficient initialization methods for the K-means clustering algorithm, Expert Syst. Appl. 40 (1) (2013) 200–210.
[28] P. Berek, M. Prilepok, J. Platos, V. Snasel, Classification of EEG signals using vector quantization, in: L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada (Eds.), Artificial Intelligence and Soft Computing: 13th International Conference, ICAISC 2014, Zakopane, Poland, June 1–5, 2014, Proceedings, Part II, Springer International Publishing, Cham, 2014, pp. 107–118.
[29] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. vol. 20 (1987) 53–65.
[30] I.T. Jolliffe, Principal Component Analysis, 2nd edn, Springer, New York, 2002.
[31] A.C. Rencher, Methods of Multivariate Analysis, 2nd edn., Wiley, New York, 2002.
[32] [Online]. Available: http://www.nrel.gov/.
[33] T. Kohonen, The self-organizing map, Proc. IEEE vol. 78 (no. 9) (1990) 1464–1480.
[34] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. vol. 3 (no. 3) (1973) 32–57.
[35] Y. Gala, Y.Á Fernández, J. Díaz, J. Dorronsoro, in: J. Pan, M. Polycarpou, M. Woźniak, M.A. de Carvalho, H. Quintián, E. Corchado (Eds.), Support Vector Forecasting of Solar Radiation Values. Hybrid Artificial Intelligent Systems, vol. 8073, Springer, Berlin Heidelberg, 2013, pp. 51–60.
[36] R. Perez, S. Kivalov, J. Schlemmer Jr., K. Hemker, D. Renné, T.E. Hoff, Validation of short and medium term operational solar radiation forecasts in the US, Sol. Energy vol. 84 (no. 12) (2010) 2161–2172.
[37] E. Lorenz, J. Remund, S.C. Mller, W. Traunmller, D.G. Steinmaurer, J.A. Ruiz-Arias, V.L. Fanego, L. Ramirez, M.G. Romeo, C. Kurz, L.M. Pomares, C.G. Guerrero, Benchmarking of different approaches to forecast solar irradiance, 24th European Photovoltaic Solar Energy Conference (2009).
[38] Y. Manjili, M. Niknamfar, in: A. El-Osery, J. Prevost (Eds.), Big Data Analytic: Cases for Communications Systems Modeling and Renewable Energy Forecast. Control and Systems Engineering, vol. 27, Springer International Publishing, 2015, pp. 109–134.
[39] R. Marquez, C.F.M. Coimbra, Proposed metric for evaluation of solar forecasting models, J. Sol. Energy Eng. vol. 135 (no. 1) (2012).
[40] H.G. Beyer, J. Polo, M. Suri, J.L. Torres, E. Lorenz, S. Muller, C. Hoyer-Klick, and P. Ineichen, Report on Benchmarking of Radiation Products, Technical report, Management and Exploitation of Solar Resource Knowledge, 2009.
[41] A.K. Yadav, S.S. Chandel, Solar radiation prediction using artificial neural network techniques: a review, Renew. Sustain. Energy Rev. vol. 33 (2014) 772–781.

M. Ghofrani received his B.Sc. degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran in 2005, the M.Sc. degree from University of Tehran, Tehran, Iran, in 2008, and the Ph.D. degree from the University of Nevada, Reno, in 2014. He is currently an Assistant Professor at the School of Science, Technology, Engineering and Mathematics, University of Washington, Bothell. His research interests include power systems operation and planning, renewable energy systems, smart grids, electric vehicles and electricity markets.

M. Ghayekhloo received her B.Sc. degree in Computer Engineering from Mazandaran University of Science and Technology, Babol, Iran, and the M.Sc. degree from Science and Research Branch, Islamic Azad University, Qazvin, Iran in 2011 and 2014, respectively. Her research interests include optimization algorithms, artificial neural networks, computational intelligence and their applications in power systems.

R. Azimi received his B.Sc. degree in Software Engineering from Mehrastan University, Guilan, Iran, and the M.Sc. degree from Science and Research Branch, Islamic Azad University, Qazvin, Iran in 2011 and 2014 respectively. His research interests include data clustering, artificial Intelligence and their applications in power systems.