Energy 78 (2014) 247–256

Contents lists available at ScienceDirect

Energy

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

Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 2013–14 Solar Energy Prediction Contest

S.K. Aggarwal [a,*], L.M. Saini [b,1]

[a] Electrical Engineering Department, M. M. Engineering College, Mullana, Ambala, Haryana, 133203, India
[b] Electrical Engineering Department, National Institute of Technology, Kurukshetra, Haryana, 136119, India

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form
23 September 2014
Accepted 2 October 2014
Available online 31 October 2014

Keywords:
Artificial neural network
Ensemble learning
Least square regression
Regularization
Solar energy forecasting
Variable segmentation

ABSTRACT

In 2013, American Meteorological Society Committees on AI (artificial intelligence) Applications organized a short-term solar energy prediction competition aiming at predicting total daily solar energy received at 98 solar farms based on the outputs of various weather patterns of a numerical weather prediction model. In this paper, a methodology to solve this problem has been explained and the performance of ordinary LSR (least-square regression), regularized LSR and ANN (artificial neural network) models has been compared. In order to improve the generalization capability of the models, more experiments like variable segmentation, subspace feature sampling and ensembling of models have been conducted. It is observed that model accuracy can be improved by proper selection of input data segments. Further improvements can be obtained by ensemble of forecasts of different models. It is observed that the performance of an ensemble of ANN and LSR models is the best among all the proposed models in this work. As far as the competition is concerned, Gradient Boosting Regression Tree has turned out to be the best algorithm. The proposed ensemble of ANN and LSR model is able to show a relative improvement of 7.63% and 39.99% as compared to benchmark Spline Interpolation and Gaussian Mixture Model respectively.

1. Introduction

The contribution of solar energy production from solar PV (photovoltaic) and solar concentrator installations is growing rapidly due to the dramatic cost reductions in solar technology. As of 2012, more than 100 GW of solar had been installed worldwide [1]. With about 37,007 MW of solar PV power installed in 2013, world solar PV power capacity increased about 35%–136,697 MW [2]. Because of the exponential rate of growth of solar installations, there is an increasing need for precise short term forecasting (12–24 h forecasting horizon) of amount of solar irradiance or solar energy received by a particular region. This is important for planning of the operations of solar based power plants, optimizing energy storage systems capacity ratings, system reserves and managing the energy market activities [3,4].

Solar power forecasts typically are derived from NWP (numerical weather prediction) models; but, statistical and machine learning techniques are increasingly being used in conjunction with the NWP models to produce more accurate forecasts since NWP forecasts are available at a few distinct spatial points only. Many AI (artificial intelligence) and statistical techniques have been proposed for forecasting of solar radiation such as ANN (artificial neural network) , FL (fuzzy logic) , GA (genetic algorithm) , and hybrid systems etc. [5–8]. Depending on the explanatory variables used the statistical models can be divided into three major categories: (i) structural models which are based on other meteorological and geographical parameters [7–9]; (ii) time-series models which only consider the historically observed data of solar irradiance as input features [10]; and (ii) hybrid models which consider both i.e. solar irradiance and other variables as exogenous variables [11–23]. While this is the categorization based on input features, prominent function approximation techniques used for the task are: ANN, FL, ANFIS (artificial network based fuzzy inference system) , SVM (support vector machine) , ELM (extreme learning machine) and conventional methods such as LSR (linear least square regression) [11–23].

* Corresponding author. Tel.: +91 8059931082 (mobile).
E-mail addresses: vasusanjeev@yahoo.co.in (S.K. Aggarwal), lmsaini@rediffmail.com (L.M. Saini).
1 Tel.: +91 941613773 (mobile).

http://dx.doi.org/10.1016/j.energy.2014.10.012
0360-5442/

Organizing forecasting competitions for the problems related to electric power system is a new development [24,25] and this is the first time that a solar energy prediction contest has been organized. In the year 2013 (July, 8 to November, 15), American Meteorological Society (AMS) Committees on AI Applications to Environmental Science, Probability and Statistics, and Earth and Energy in association with EarthRisk Technologies Inc. (http://www.earthrisktech.com/) organized AMS 2013–2014 Solar Energy Prediction Contest.The contest was administered by a team of experts from the University of Oklahoma. The competition aimed to predict solar energy output forecasts by employing machine learning techniques considering various forecast weather patterns generated from a NWP forecasting system as input features. The main objective was to predict the total daily incoming solar energy at 98 Oklahoma Mesonet sites, which served as "solar farms" for the contest. The goal of the contest was to discover better statistical and machine learning techniques for the best short term predictions of solar energy production. The Kaggle platform (https://www.kaggle.com) was provided to the participants to conduct the contest.

During this competition, as many as 160 participants took part and produced results with techniques like GBRT (Gradient Boosted Regression Tree) and SVM etc. [26–28]. The top participants in the competition used GBRT in their models; however, the authors took part in this competition based on ANN technique and were able to outperform all the three benchmark techniques provided by the organizers. This was a computationally intensive exercise, which required big data analysis, feature extraction from multi-dimensional datasets and a long model development and execution time. However, after closing of the competition, having found the problem interesting, it was decided to further explore the characteristics of data and make a comparison of various machine learning techniques [29–32]. The main focus and contribution of this work is to explain the application of various linear and non-linear machine-learning algorithms on such an involving function approximation problem and improving their generalization capability using data segmentation and ensemble of different models. In this paper, the authors have presented their approach and have discussed more on this problem. An important conclusion drawn from our experiments is that given similar input features, non-linear learning algorithms like ANN can perform better than the linear regularized LSR techniques. However, learning of ANN based models takes a lot of time as compared to LSR technique. Moreover, overall performance can be further improved by using an ensemble of ANN and LSR models.

This paper is organized as follows: In Section 2, task description, research motivation and data analysis has been presented. In Section 3, the methodology for preparing the training and test data has been explained. Forecasting techniques used in this work have been explained in Section 4. Section 5 consists of the forecasting results and discussion. Conclusion has been presented in Section 6.

2. Problem description

In this section, the data set, its analysis and task given to the participants has been explained.

2.1. Competition task description

The organizers of the contest provided the following data to the competitors:

(A) Training data (Input Matrix): This data has been taken from the National Oceanic and Atmospheric Administration's NOAA/ESRL Global Ensemble Forecast System (GEFS) Reforecast Version 2 [33]. It consists of 5 daily predictions for the next day for each of the following 15 variables from period 1994 to 2007 across each of 144 GEFS locations on a 16 × 9 grid. (5113 days)

  (1) 3-Hour accumulated precipitation at the surface 'kg/m^2'
  (2) Downward long-wave radiative flux average at the surface 'W/m^2'
  (3) Downward short-wave radiative flux average at the surface 'W/m^2'
  (4) Air pressure at mean sea level 'Pa'
  (5) Precipitable water over the entire depth of the atmosphere 'kg/m^2'
  (6) Specific Humidity at 2 m above ground 'kg'
  (7) Total cloud cover over the entire depth of the atmosphere 'kg^{-1}'
  (8) Total column-integrated condensate over the entire atmosphere 'kg/m^2'
  (9) Maximum Temperature over the past 3 h at 2 m above the ground 'K'
  (10) Minimum Temperature over the past 3 h at 2 m above the 'K'
  (11) Current temperature at 2 m above the ground 'K'
  (12) Temperature of the surface 'K'
  (13) Upward long-wave radiation at the surface W/m^2
  (14) Upward long-wave radiation at the top of the atmosphere W/m^{-2}
  (15) Upward short-wave radiation at the surface W/m^{-2}

(B) Training data (Target vector): Total daily incoming solar energy at 98 Oklahoma Mesonet sites from period 1994 to 2007. The incoming downward solar radiation is directly recorded by a Li-Cor Pyranometer every five minutes, and the result summed over the entire day. Solar irradiance is presented as a time series, where each point of the series represents the sum of the solar flux received over a day.

(C) Testing data: This is similar to training input data and consists of daily predictions of the above 15 variables from period 01-01-2008 to 30-11-2012 across each of 144 GEFS locations on a 16 × 9 grid. (1796 days)

The complete daily solar energy data for training as well as testing period was provided by the Oklahoma Mesonet (https://www.mesonet.org/). The Oklahoma Mesonet is a network of environmental monitoring stations designed to measure the environment at the size and duration of mesoscale weather events [34]. The task was to make predictions of the total solar daily incoming solar radiation at 98 Oklahoma Mesonet sites for each day of test period from 01 to 01-2008 to 30-11-2012.

The MAE (Mean Absolute Error) was the metric being used for this competition. It is commonly used in regression problems and by the renewable energy industry to compare forecast performance. The formula is given by:

$$\mathrm{MAE} = \frac{1}{n}\sum_{1}^{n}|(\text{Actual value} - \text{Forecast value})|$$

The contestants were allowed to submit 2 submission files for the complete test period. During this period, they could see the performance of their models on a validation set (40% of the complete dataset known as public data). However, 60% of the data, known as private data, was reserved for final evaluation of the models at the end of the competition.

The GEFS is a weather model that predicts weather variables at various locations, and the training and test data is those predictions. The GEFS is the US-run global NWP model and an ensemble of NCEP Global Forecast System (GFS) [33]. The GEFS has 11 ensemble members with perturbed initial conditions. Each of