A Concept-Drift Based Predictive-Analytics Framework: Application for Real-Time Solar Irradiance Forecasting

Jessica Wojtkiewicz*, Satya Katragadda† and Raju Gottumukkala‡
*‡College of Engineering,†‡Informatics Research Institute
University of Louisiana at Lafayette
Lafayette, United States
Email: *jessicaw1, †satya, ‡raju@louisiana.edu

***Abstract*—Solar irradiance is the measurement of the amount of power from the Sun per unit area. Solar irradiance has a very high degree of variability, due to many environmental factors, including cloud cover, relative humidity, and air temperature. Predicting solar irradiance is very useful for measuring future solar energy production and power scheduling. Real-time solar irradiance can be forecasted using either maching learning or physics-based models, both having their own respective trade-offs. To overcome the limitations of both machine learning and physics-based simulations, we propose a novel framework in predictive analytics that combines the power of both types of models. In particular, we propose a concept-drift framework to develop a model that integrates satellite cloud imagery with real-time solar irradiance measurements to improve irradiance forecasting. We present preliminary results using real-world irradiance and cloud data to support the motivation for a concept-drift based approach. Based on our results, we found that one static model cannot fully capture the dynamics of solar irradiance on both days where cloud cover is present and days where there is no cloud cover.**

***Index Terms*—predictive analytics, concept-drift, solar irradiance, big data**

I. Introduction

Advances in data science and big data techniques to process high-volume, high-velocity data streams from multiple sources have enabled the development of predictive analytics in many applications including science, engineering, and business. By nature, machine learning techniques are purely data-driven, making them agnostic to real-world phenomena. Many scientific and engineering domains—particularly in the areas of climatic modeling, hydrology, material science, and environmental science—rely on well-established physics-based models to both understand and predict future behavior. While these models explain causality through closed-form equations or numerical simulations, the solutions are in some cases limited with respect to their computational tractability. Moreover, some real-world phenomena driven by complex relationships among variables cannot be fully explained with numerical models. To overcome the limitations of both machine learning and physics-based simulations, we propose a novel framework in predictive analytics that combines the power of both types of models. We specifically propose a concept-drift approach to create a hybrid model that integrates physics-based models with observed solar irradiance and cloud cover data to improve irradiance forecasting.

Properties of solar irradiance change depending on the time of year and the weather on any given day. On days with little to no cloud cover, solar irradiance measurements follow a similar and predictable pattern throughout the day, and many statistical and machine learning models can learn this daily and yearly pattern over time and forecast future behavior accurately. Contrarily, on days where cloud cover is present, solar irradiance measurements become more difficult to predict due to sudden changes in irradiance. Several forecasting techniques have been applied to solar irradiance over short and long time horizons, ranging from less than an hour to a couple of days. Physics-based models include Numerical Weather Prediction (NWP) modeling [1] and the use of satellite and ground-based sky image data to predict cloud formations [2]. In general, forecasting methods using satellite or cloud images perform best for very short time horizons, while NWP models outperform most physics-based models when predicting over a time horizon greater than six hours [3]. Statistical and machine learning models such as Regression [4], Autoregressive Integrated Moving Average (ARIMA) [5], and Artificial Neural Networks [6], primarily use historical data, exogenous meteorological data, or both to forecast solar irradiance along short time horizons. These data-driven models have the ability to capture the dynamics of solar irradiance on days with little to no cloud cover; however, when sudden changes in irradiance occur due to external factors, the same model fails to accurately forecast future behavior. In this case, a static model cannot generalize the stochastic nature of solar irradiance and accurately predict sudden short term changes. As a result, a class of hybrid or ensemble models have been developed to improve overall forecasting accuracy [7], [8]. These hybrid models combine both statistical and machine learning models and have shown improvements in forecasting solar irradiance compared to using individual models [4].

Traditionally when using statistical or machine learning models, parameters are adjusted by analyzing historical data

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

and using the same parameters to predict a certain target variable some time step into the future [9]. Applying this same methodology to forecast data streams in real-time presents a challenge as new data is constantly collected and the distribution of the data may change over time [10]. In order to combat this problem, concept-drift based approaches have been applied to many domains in order to identify when a target variables distribution changes over time and apply the correct model to produce accurate predictions. In this paper, we describe a concept-drift based methodology to forecast solar irradiance using historical time series along with satellite cloud cover data.

II. Methodology

In this section, we define the problem statement for predicting the solar irradiance values from historical data. We then introduce a concept drift sensitive forecasting framework to predict solar irradiance.

A. Problem Statement

Solar irradiance data can be represented as a data stream $d = \{ir_1, ir_2, ir_3, \ldots, ir_n, \ldots\}$, where ir_t is the irradiance value at time t. The granularity of the irradiance values is set to one hour. From these data streams, a sample of data is extracted based on a sliding window of size s. Based on the batch of data in each sliding window, a feature set is extracted. Each feature set consists of a number of elements and corresponds to a feature from the feature space. This set of features and the irradiance value from the next time step can be represented as a learning pair $\{x, y\}$, where x is the feature set extracted from the batch of data at time t, and y is the irradiance value at time $t + 1$. In a typical scenario, data arrives continuously in the form of a data stream D. The solar irradiance prediction algorithm is presented with a sequence of historical examples $p_t = \{x_t, y_t\}$ for $t = 1, 2, \ldots, T$. At each time step t, the prediction algorithm analyzes the appropriate historical instances $\{p_1, p_2, \ldots, p_t\}$ and the incoming distance d_{t+1}, which is treated as a training example. The goal of the irradiance forecasting approach is to predict the future irradiance value.

B. Solar Irradiance Forecasting Approach

The solar irradiance forecasting methodology is presented in Fig. 1. At a given time t, a set of previous irradiance values $s = \{ir_{t-i}, ir_{t-i+1}, \ldots, ir_t\}$ are extracted from the data stream, where i is the size of the sliding window. A group of features are extracted from this set s to identify the trend, seasonality, and the pattern of the irradiance function. In the current context, we extract features like the most recent irradiance value, irradiance value at time $t - 24$, difference between irradiance values ir_t and ir_{t-1}, as well as the difference between irradiance values at ir_{t-24} and ir_{t-23}. These features can also incorporate additional external features like cloud cover, relative humidity, air temperature, etc.

The repository of trained models based on historical data is used to build an ensemble predictor to identify the next irradiance value. Each individual model and a combination of models are used to identify the most appropriate set of models that can be used to predict irradiance. The models are selected based on the similarity between the test instance and the trained instances on which the models are based. A bagging approach is used to combine the outputs of various trained models. The output from this approach is used to predict the irradiance value at time $t + 1$.

The most recent pair d_t is used to incrementally update the existing training models.

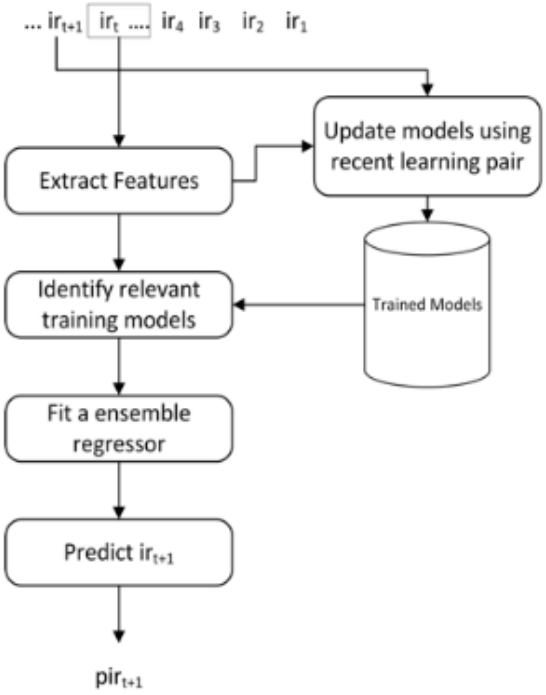

Fig. 1. Proposed solar irradiance forecasting framework.

III. Results

We present preliminary results using real-world solar irradiance and cloud cover data in order to support the motivation for a concept-drift based approach for real-time solar irradiance forecasting.

A. Data Description

In order to forecast solar irradiance, real-world data sets containing solar irradiance and cloud cover are selected. The National Solar Radiation Database, maintained by the National Renewable Energy Laboratory (NREL) provides free access to solar irradiance and meteorological data from around the world at hourly time intervals [11]. In our case, hourly global horizontal irradiance for Phoenix, Arizona, from January 1, 2004, to December 31, 2014, is used as the target variable. The satellite cloud data is downloaded from NOAAs International Satellite Cloud Climatology Project (ISCCP) [12]. The data set ISCCP HXG contains cloud data at three-hour intervals spanning latitude and longitude values at one-tenth of a degree.

Cloud amount is recorded as a binary variable, using ones and zeroes to depict whether cloud cover is present or not, respectively. The same date interval of data is chosen in the cloud data set, and the three-hour intervals are repeated for each hour to match the hourly solar irradiance data set. The data is subset to include only hours from 9:30 to 14:30 as forecasting accuracy during sunrise and sunset is much higher, and there are limitations given that the cloud data is recorded in three-hour intervals. A combination of training and testing is used to measure forecasting accuracy for each set of data. Depending on the size of each data set, it is divided into 80 percent training and 20 percent testing.

B. Model Performance

A regression in logs model is used to test the performance of forecasting models against solar irradiance data. Each variable is transformed by a natural logarithm, and zero values are transformed by adding a small constant. The model uses the previous time step, previous-day same-time step, and the difference between the current time step and previous time step in order to forecast one-time step into the future. The regression in logs model is as follows:

$$\begin{aligned} \ln X_{t+1} &= \alpha_0 + \alpha_1 \ln X_{t-1} + \alpha_2 \ln X_{t-6} \\ &\quad + \alpha_3 \ln(X_t - X_{t-1}) + \epsilon_t \end{aligned}$$

where t represents time, α represents the coefficients, and X represents solar irradiance. The performance of this forecasting model is evaluated based on two evaluation metrics: Mean Absolute Percentage Error (MAPE) and Multiple R-Squared. Table I shows the results for predicting solar irradiance using each of the four data samples. Overall, the models performance on the full data set from 9:30 to 14:30 is better than the days where cloud cover is present. On clear days where no cloud cover is present, the performance of the model is almost 75 percent better than on days in which cloud cover is present. In the case of data streams, these results show that one model does not accurately capture the dynamics of solar irradiance on both days where cloud cover is present and clear days. The proposed framework which categorizes the days where cloud cover is present or not is an ideal choice given the wide range of accuracy in these data sets.

TABLE I
PRELIMINARY RESULTS

Data Set	***MAPE***	***Multiple*** R^2
Clear days	5.0046	0.8602
Clouds present 9:30 - 11:30	18.3747	0.7917
Clouds present 12:30 - 14:30	19.9239	0.7601
Full data 9:30 - 14:30	10.7225	0.8331

IV. CONCLUSION AND FUTURE WORK

Given the stochastic nature of solar irradiance, a single forecasting model does not have the ability to fully capture the dynamics of solar irradiance. On days in which cloud cover is not present, a log-regression model performs relatively well against a real-world solar irradiance data set; contrarily, the same model is unable predict sudden changes in irradiance with the same accuracy due to cloud cover. To combat this, we propose a concept-drift based framework to extract the features of the current day, apply an appropriate training model on this window, and build an ensemble regressor to predict the next irradiance value.

In order to improve the efficiency and effectiveness of the proposed concept-drift based framework, we plan to optimize the framework and implement this approach using real-time solar irradiance data. Firstly, we want to identify and determine the optimal number of trained models to maintain in the repository for accurately predicting solar irradiance. To do this, we must continue to evaluate various statistical and deep learning models that perform best for predicting solar irradiance. Secondly, we plan to study and evaluate the best way to incrementally update the existing models using the most recent learning pair. In the case of data streams, updating the forecasting models is an important issue as new data is continuously collected over time. With this approach, we are able to produce forecasts based on changes in the environment and fully capture the dynamics of solar irradiance, leading to a higher overall forecasting accuracy.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation under grant numbers CNS-1429526 and CNS-1650551.

REFERENCES

[1] R. Perez, M. Beauharnois, K. Hemker, S. Kivalov, E. Lorenz, S. Pelland, J. Schlemmer, and G. Van Knowe, "Evaluation of numerical weather prediction solar irradiance forecasts in the us," in *Proc. solar*, 2011.

[2] P. Mathiesen, C. Collier, and J. Kleissl, "A high-resolution, cloud-assimilating numerical weather prediction model for solar irradiance forecasting," *Solar Energy*, vol. 92, pp. 47–61, 2013.

[3] R. Perez, S. Kivalov, J. Schlemmer, K. Hemker Jr, D. Renné, and T. E. Hoff, "Validation of short and medium term operational solar radiation forecasts in the us," *Solar Energy*, vol. 84, no. 12, pp. 2161–2172, 2010.

[4] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, 2009.

[5] D. Yang, P. Jirutitijaroen, and W. M. Walsh, "Hourly solar irradiance time series forecasting using cloud cover index," *Solar Energy*, vol. 86, no. 12, pp. 3531–3543, 2012.

[6] F. N. Melzi, T. Touati, A. Same, and L. Oukhellou, "Hourly solar irradiance forecasting based on machine learning models," in *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pp. 441–446, IEEE, 2016.

[7] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, 2013.

[8] S. Cao and J. Cao, "Forecast of solar irradiance using recurrent neural networks combined with wavelet analysis," *Applied Thermal Engineering*, vol. 25, no. 2-3, pp. 161–172, 2005.

[9] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, p. 44, 2014.

[10] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016.

[11] "National solar radiation database (nsrdb)." https://nsrdb.nrel.gov/. Accessed: 2018-10-31.

[12] "International satellite cloud climatology project (isccp)." https://www.ncdc.noaa.gov/isccp. Accessed: 2018-10-31.