Solar Energy Prediction using ANN

Ankit Choudhary (1301CS53)
Vishal Chauhan (1301CS46)

1. ABSTRACT

The Solar Energy production is on the rise recently due to the considerable cost reduction in solar technology. As the solar installations are growing at a rapid rate there is a need for short term forecasting of amount of solar energy that will fall in a particular region. In this light for the first time American Meteorological Society (AMS) organized a solar energy prediction contest on kaggle. The main objective was to predict the total daily incoming solar energy at 98 Oklahoma Mesonet sites. These sites had the solar farms where solar energy was needed to be predicted.

We have tackled the problem using ANN technique. This was a computationally expensive technique as it involved some big data analysis, feature extraction from multi-dimensional data set and high model development and execution time.

2. RESOURCES

Data Resource: https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/data

Ideas: https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest/discussion

Libraries used: netCDF4, keras with Theano backend, numpy, matplotlib, sklearn

3. WORK DONE

3.1 INPUT DATASET

The following data was provided by the organizers to the competitors:

A) Training Data: This was provided by National Oceanic and Atmospheric Administration's (NOAA) Global Ensemble Forecast System (GEFS). This has 5 daily predictions for the given 15 variables for each day from a period of 1994 to 2007 for each of the 144 GEFS locations (on the 16 x 9 grid of latitudes and longitudes):
   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 kg-1
   7) Total cloud cover over the entire depth of the atmosphere %
   8) Total column-integrated condensate over the entire atmosphere kg m-2
   9) Maximum Temperature over the past 3 hours at 2 m above the ground K
   10) Minimum Temperature over the past 3 hours at 2 m above the ground 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) Target Vector: This had the total daily incoming solar energy at the 98 Oklahoma Mesonet sites from period 1994 to 2007 for which the total solar energy needs to be predicted.

C) Testing Data: This is same as the training data but for a different period of time i.e. from 2008 to 2012 for each of the 15 variables for each of the 144 GEFS locations.

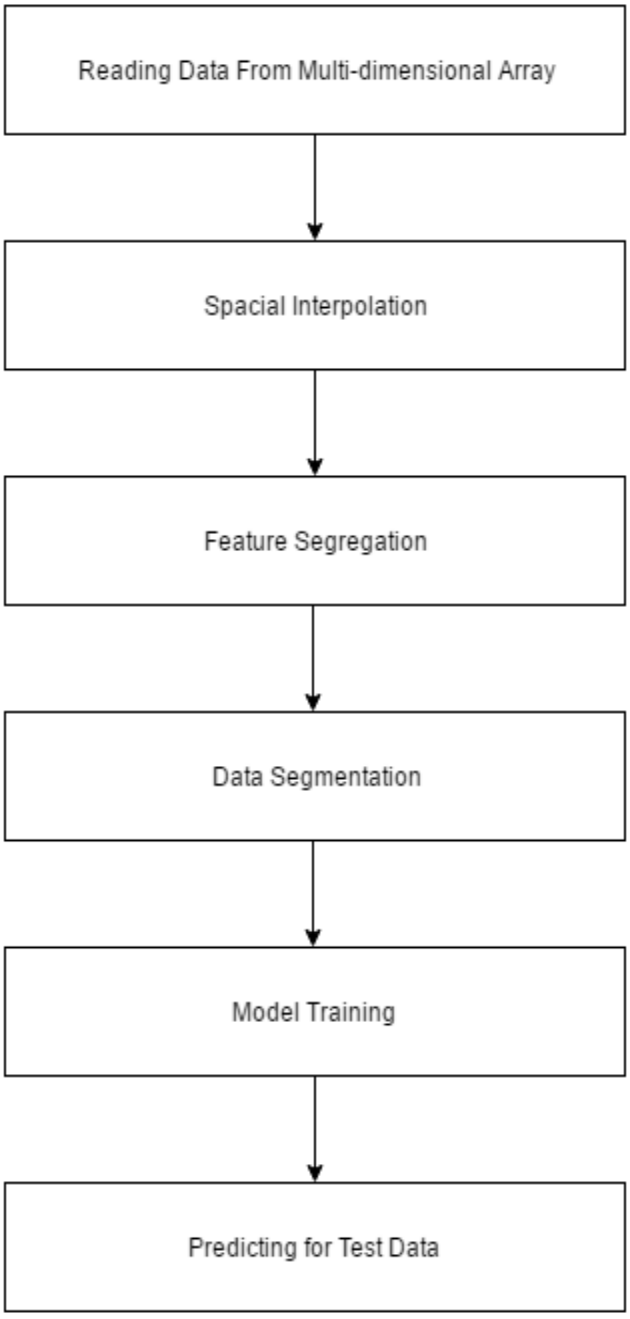
GEFS is a weather forecast model made of 11 ensemble members. The GEFS has eleven ensemble models with different starting conditions. Now each of the ensemble models makes forecasts based on their mathematical model. Each will make different predictions as they different initial conditions. These models make predictions 5 times per day for each hours 12:00, 15:00, 18:00, 21:00 and 24:00.

Each of input training and test data was provided in the form of 15 multi-dimensional netcdf4 files. NetCDF4 is python interface to the netCDF C library. NetCDF is a set of software libraries and machine-independent data formats that helps in handling array-oriented scientific data.15 files were given for each of the variables. Each file had 5 dimensional matrix shape. The dimensions were:

1) Number of the day
2) Ensemble member of the GEFS
3) Time-steps of the weather prediction values
4) Geographical longitude for the GEFS location.
5) Geographical latitude for the GEFS location.

The spatial locations of the 98 mesonet sites on a 2 dimensional grid along with their elevations were also given.

3.2 OUR APPROACH



3.2.1 Feature Segregation

As mentioned before each file for a variable had 5 dimensional matrix shape. The dimensions being the number of the day, ensemble member of GEFS, time-steps of the day, latitude and longitude of the GEFS site. All the 15 variables are strongly co related with the solar energy being received at a particular site.

Now to obtain a relation between these weather variables and target variable, we adopt two ways to get features:

1) Average of all 11 ensemble members over all 5 time steps has been taken giving 15 x 1 input feature matrix for each day. (15 Features)
2) Average of all 11 ensemble members at each time step was considered separately giving 15 x 5 input feature matrix for each day. (75 Features)

Then normalization of the input matrix was performed to scale the data set in the range of [-1, 1].

3.2.2 Spatial Interpolation

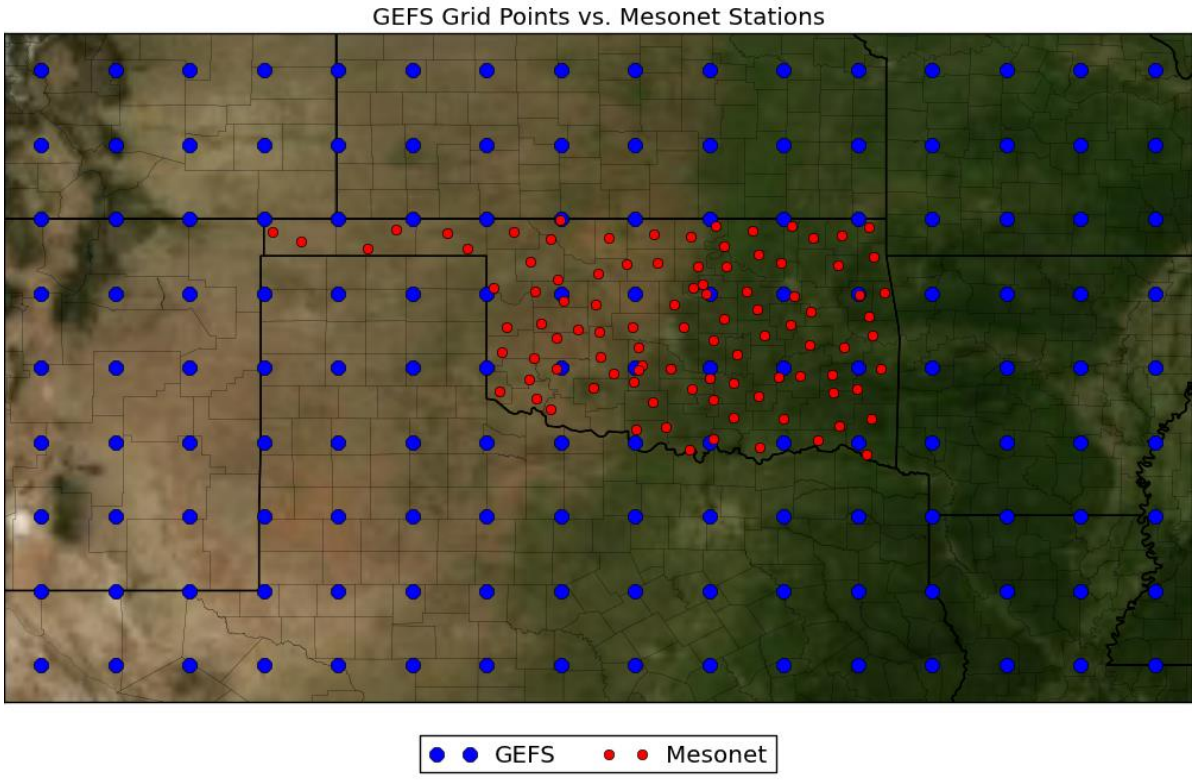


Figure 1 : mesonet vs gefs points (source: kaggle)

As we can see that the GEFS locations are apart 1 degree latitude and longitude marked in blue containing a total of 144 locations. The 98 mesonet sites belong to the Oklahoma state and are marked in red.

As it can be seen that the mesonet sites containing the solar farms are distributed unevenly and do not coincide with the GEFS locations. Now as the GEFS data locations and the Mesonet site locations are “shifted” spatially we have done spatial interpolation using cubic splines to adjust the GEFS values to the mesonet locations. It took nearly 7h to get the interpolated data for the mesonet sites.

3.2.3 Data Segmentation

Here we try to take advantage of the fact that total solar energy received exhibits a seasonal pattern. For example, more solar energy is received in winters as compared to summers. So the data was separated out season wise, which may reduce the error value by training different neural networks for different seasons.

Data for different seasons may belong to different distributions. The success of any machine learning algorithm depends upon the data set which should belong to similar distribution. So instead of using a single data segment, we segment our input data into 4 segments, 1 for each season i.e. 90 days interval. We compared this method with the baseline method of using only a single data segment (365 days).

3.3 Neural Network Architecture

We designed the neural network having two hidden layers. The first layer consisted 15 nodes and the second layer had 11 nodes. We used **tanh** transfer function at the first and second layers and linear transfer function was used at the output layer. We used **sgd** as the training algorithm.

Mean absolute error (MAE) was considered as the cost function or error criterion which is mostly used for regression problems. Maximum epochs were set at 2000.

We used PCA (principal component analysis) to reduce features from 15 to 7 and 75 to 10.

3.4 RESULTS

MODEL	TRAIN MAE	TEST MAE
FFNN(15 features, single data segment)	6324693	6963159
FFNN(75 features, single data segment)	6085464	6654795
FFNN(15 features, four data segments)	4130633	4601898

As can be seen, there was an increase in performance using data segmentation. The following 3 figures relate the MAE with the Epochs.

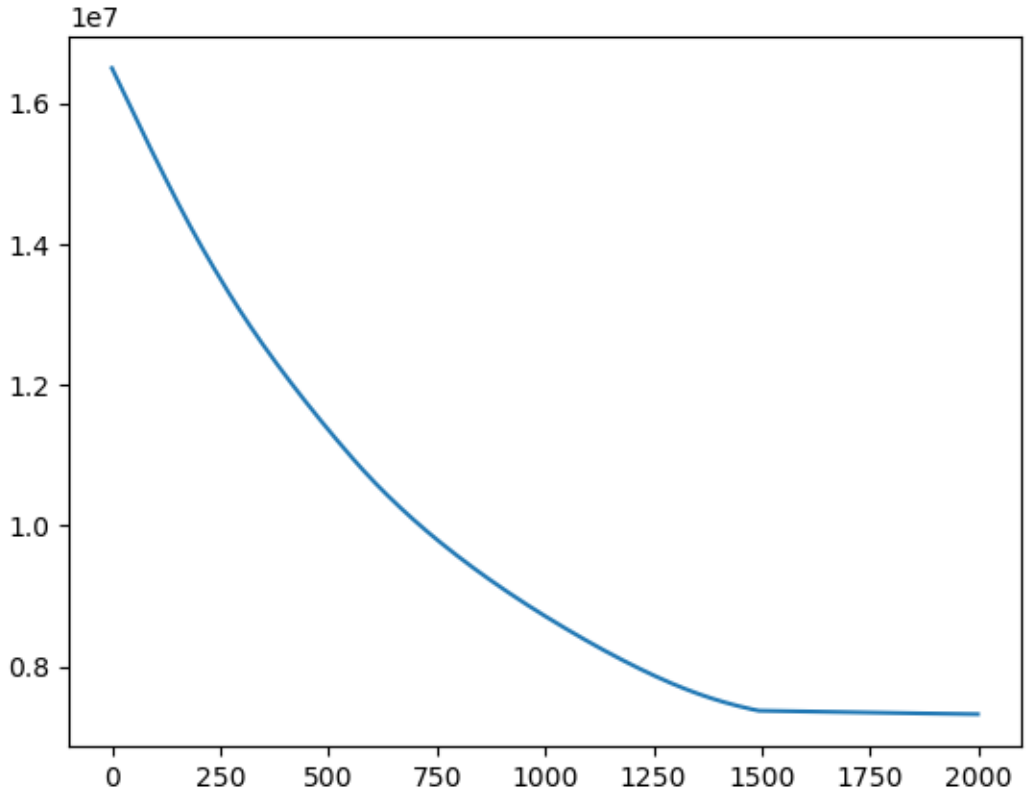


Figure 2 : 15 features (single segment)

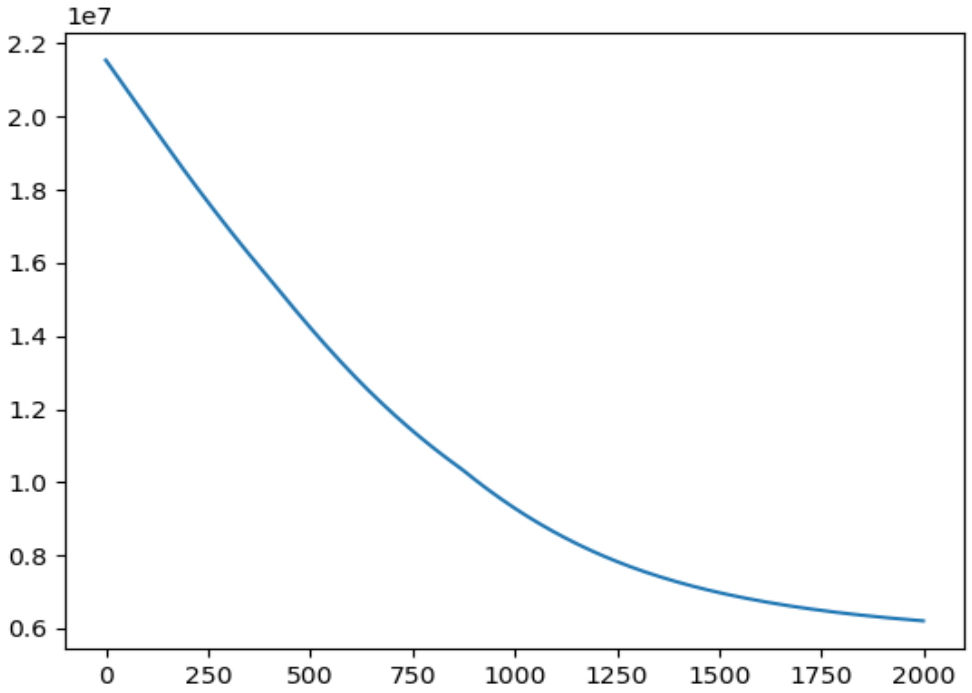


Figure 3 : 75 features (single segment)

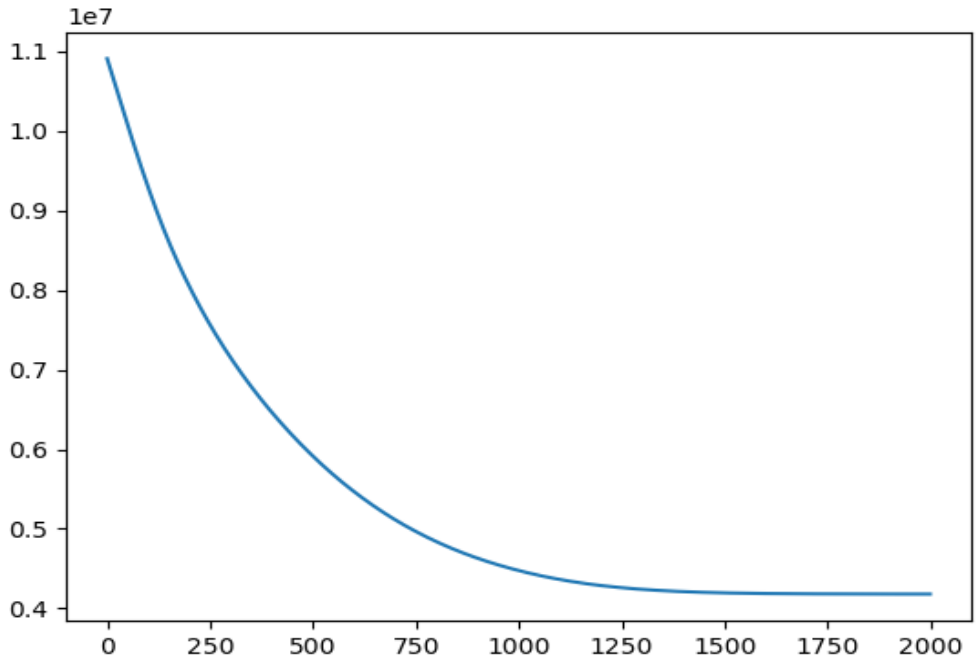


Figure 4: 15 features (4 segments)

3.5 GitHub Repository

https://github.com/ankitcs53/Solar-Energy-Prediction-Using-ANN

4. Future Work

There is a possibility of reducing the training time by exploring different training algorithms. Data can be further segmented to monthly data, which may reduce the mean absolute error further. Sites specific data segmentation can be done to get better results.