Predicting Next-Day Solar Power Generation Using Neural Networks and Support Vector Machines

Colton Morley

Department of Computer Science and Engineering, University of Nevada, Reno, NV

CS458

***Abstract*—This paper aims to create a 24 hour ahead solar energy generation forecast using data provided by the European Center for Medium-range Weather Forecasts. The approach taken begins with analysis of the data, variables, and correlations between different variables and power outputs. After these variables have been analyzed and preprocessed, the data is then used to train a Multi-Layer Perceptron Regressor. This regressor is then tested and the results are taken and compared with the expected values. Next the data is fed into a Support Vector Regressor. The results of the MLP and SVR are both compared both with each other as well as with the expected power generation outputs. The mean absolute error and root mean squared error are both used in this comparison as well. The proposed root of the issues in accuracy is deemed to be the regressors' reliance upon current power output to predict the next day. The regressors are retrained without the current power output variable and the results are compared.**

Keywords – Artificial Neural Network (ANN); Multi-layer Perceptron (MLP); Support Vector Regression (SVR);

I. INTRODUCTION

The problem of energy creation is a large issue in the world today. Research has concluded that if we do not decrease our reliance on fossil fuels soon that the global impact could be catastrophic. Solar, wind, and water power are potential alternatives, but as of now they cannot efficiently produce enough power to support the Earth. Solar power plants are efficient at creating energy, but a large issue arises in the inconsistency of the sun.

For that reason, this paper aims to help the unpredictability of the sun become more predictable, using the data provided by the European Center for Medium-range Weather Forecasts [1]. With the need for renewable resources increasing drastically, many studies have been done, and papers written regarding the prediction of renewable resource generation. With these references in hand, this paper aims to develop models and methods built off the solar power data given to predict future solar power generation at the three zones described by the data.

To accomplish this, this paper will begin by using a multi-layer perceptron, which is a class of feedforward artificial network. After preprocessing, this MLP regressor will be trained on a portion of the dataset, and then will be used to predict next day solar power production. After this, this paper will move to a support vector regressor. This support vector regressor will be fed the same preprocessed data, and then once trained, will also attempt to accurately predict the next day solar power generation.

The organization of the rest of the paper consists of Section II which speaks on background and other related works to this paper, Section III which consists of the methods used to carry out this forecast, Section IV which consists of the evaluation results of the models, and finally Section V which concludes the paper.

II. BACKGROUND AND RELATED WORK

The first relevant work that this paper will discuss is the "Ensemble Learning Approach for Probabilistic Forecasting of Solar Power Generation"[2], conducted by members of the Masdar Institute of Science and Technology. The authors of this paper use an ensemble of machine learning methods to create their forecast using seven individual machine learning-based regression models. The models used in this paper consist of Decision Trees, Gradient Boosting, K-Nearest Neighbors, Lasso, Random Forests, and Ridge Regression. Another paper that is relevant to this work is "A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting"[3], written by students of the University of Texas, San Antonio, and the Texas Sustainable Energy Institute. Like the previous paper, this paper also aims to accurately forecast future solar power generation using machine learning techniques. This paper uses Artificial Neural Networks and Support Vector Regression to accomplish this task.

Previous attempts to predict future solar power prediction have been fruitful. These previous papers have grouped data in various ways. Although this paper only groups the data by zone, the papers referenced previously used different methods to group solar power generation data. The ensemble learning approach [2] used 72 different groups of data, each zone having 24. They split the data based on factors such as power output and season along with splitting it by zone. Having many datasets with all the data being similar can help the regressors used in the ensemble approach to become more accurate due to the reduced outliers. The hierarchical approach [3] used different data, that consisted of inverters instead of zones. They used the inverters as well as power output to split the data before training the regressors.

The paper "Ensemble Learning Approach for Probabilistic Forecasting of Solar Power Generation" [2] tests different methods of splitting the data. The results conclude that splitting each dataset by the hour results in the lowest error scores and therefore produces the most accurate forecast.

Another factor that can affect the accuracy of the results is the data that is used to train these regressor. This paper uses only the previous day's data to train the regressor. The students at the University of Texas, San Antonio used not only the prior day's data, but the data of the entire week before the date being forecasted. This method creates more variables for the regressors to use for training which can help to improve accuracy. These students also used a "one-time-step-ahead" method in which they not only predicted the next day's power generation, but also the power generation 15 minutes from the current measurements, and one hour from the current measurements.

When comparing results these students found that larger times between measurement and prediction, with their 15 minute ahead forecast being more accurate than one hour ahead, and one hour ahead being accurate than 24 hour ahead predictions. This paper also concludes that the results of the ANN and SVR are fairly even in all these cases.

The paper "Ensemble Learning Approach for Probabilistic Forecasting of Solar Power Generation" [2] concludes that in terms of single machine learning regressors, Gradient Boosting was able to produce the most accurate results with a Mean Absolute Error of 0.037 and a Root Mean Squared Error of 0.083. All the other regressors tested resulted in Mean Absolute Errors above .04 although some others such as Random Forest and Ridge Regression produced comparable Root Mean Squared Errors.

III. METHODS

The data being used in this paper consists of 12 weather variables along with the resulting solar power output for the dates ranging from April 1st, 2012 to July 1st, 2014.

3.1 Preprocessing

To begin the analysis and preprocessing of the data, a correlation chart of all the relevant variables was analyzed using a colormap. This colormap uses red to represent a positive correlation and blue to represent a negative correlation, with the opacity of the color representing the strength of the correlation.

Fig. 1. Correlation Colormap of all variables included in dataset

Of all the variables included in the dataset, one can observe that there isn't a variable that strongly correlates to the next day's power generation besides the previous days power generation. For this reason, it was concluded that, at least to begin, all variables should be included in the training of the regressors, except for time stamp and zone ID.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The regressors used in this paper work best with scaled data. For this reason, the data was scaled using a min-max scaler which performs the operations shown in equation 1 to scale the data.

3.2 Hyperparameters

This paper uses both an MLP and an SVR to forecast power predictions. The hyperparameters of both can have a monumental impact in their results and accuracy. For this reason, hyperparameter tuning is key. An attempt was made to tune these hyperparameters using Grid Search, but there were issues with time complexity and result analyzation. Instead, a group of nested loops was used to test different combinations of hyperparameters and their errors. This method should in theory produce similar if not the same results as a successful Grid Search.

3.3 Multi-Layer Perceptron

The first regression method used in this prediction is a Multi-Layer Perceptron Regressor (MLP). This is a class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of node. These layers consist of an input layer, a hidden layer, and an output layer. This regressor was selected because ANNs are good at handling input with large dimension.

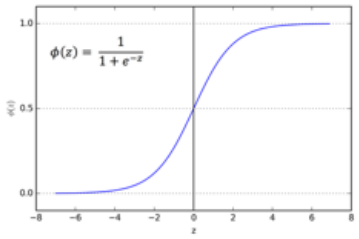

Fig. 2. Sigmoid Activation Function

$$w := w - \eta \nabla Q(w) = w - \frac{\eta}{n} \sum_{i=1}^{n} \nabla Q_i(w), \quad (2)$$

The hyperparameter tuning method from section 3.2 was used on the multi-layer perceptron with the training data. The activation function of a neural network decides whether a neuron should be activated or not, to introduce non-linearity to the output of a neuron. The activation function for this multi-layer perceptron was selected to be the Sigmoid activation function, which can be seen in Fig. 2.

The solver used to solve for weight optimization used is Stochastic Gradient Descent, and the hidden layer sizes are 100.

3.3 Support Vector Regression

The second and final regression method used in this paper is SVR, which was selected because like the MLP, it works well on input with high dimensionality. The SVR is a powerful regressor and is unique from a lot of other regression strategies because it can handle the user defining how much error is acceptable in the model. The SVR works by transforming the data into a high-dimension space, and predictions are taken from the training data and support vectors are used to to approximate the outputs.

The hyperparameters of this model were tuned using the same method as the MLP, and it was found that using linear kernel produced the best results.

3.4 Performance Evaluation

The results of these regressors will be evaluated by two different error calculations. The first being mean absolute error (MAE) and the second being root mean squared error (RMSE).

- Mean absolute error (MAE)

$$MAE = \frac{1}{number\ of\ points} \sum_{t} |P_t - \hat{P}_t| \quad (3)$$

- Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{number\ of\ points} \sum_{t} |P_t - \hat{P}_t|^2} \quad (4)$$

Absolute error represents the total error between the prediction and the actual. This is calculated by taking the difference between the prediction and actual. The MAE represents the average of this measurement throughout the data and can be seen in equation 3.

A residual is a measure of how far the data points are from the regression line. RMSE represents the standard deviation of

these residuals, or prediction errors. The equation used to calculate RMSE can be seen in equation 4.

IV. Results

To verify that the tests work, the regressors were run multiple times and the results were compared with each other. To run these regressions, the Sci-Kit Learn [5] library was used. The approach was tested by analyzing the MAE and RMSE of the results to see the accuracy of these regressors.

Table 1. MAE and RMSE for MLP

ZONEID	1	2	3	Overall
MAE	0.113	0.116	0.118	0.116
RMSE	0.025	0.024	0.025	0.025

Table 2. MAE and RMSE for SVR

ZONEID	1	2	3	Overall
MAE	0.094	0.093	0.095	0.094
RMSE	0.018	0.019	0.019	0.018

As shown in Table 1, the Multi-Layer Perceptron model resulted in an average MAE of 0.116 and an average RMSE of 0.025. Table 2 shows that the average MAE of the SVR is 0.094 and RMSE 0.018.

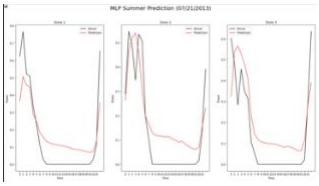

Fig 3. MLP Summer Day prediction vs actual for all three zones

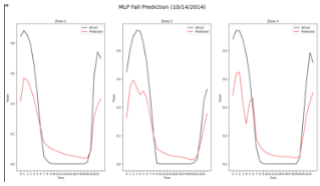

Fig 4. MLP Fall Day prediction vs actual for all three zones

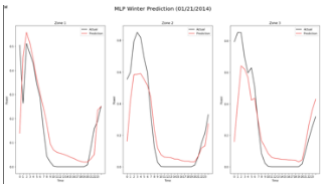

Fig 5. MLP Winter Day prediction vs actual for all three zones

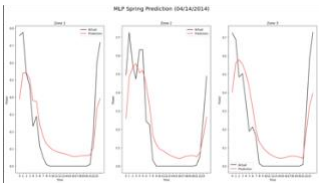

Fig 6. MLP Spring Day prediction vs actual for all three zones

Figures 3-6 demonstrate the actual power output, shown in black, compared against the predicted output, shown in red, for four days contained in the data, with one coming from each season, while using the MLP regressor. The three graphs in each figure represent Zones 1, 2, and 3 respectively. Figure 3 contains the results for a summer day. By analyzing the shown charts, we can see that the predictions struggle getting low enough during the nighttime. It is also shown that it has trouble predicting rapid fluctuations in power output. Figure 4 shows the same regressor's results on a fall day. As shown in the figure, in the fall there is an improvement getting closer to 0 at night, but a new problem arises in its tendency to predict lower power outputs at the peak of the day. In Figure 5, which represents the winter, it can be seen that the predictions do a better job of estimating the power production both at night and at the peaks of the day, however it still tends to underestimate the output at the peaks. Finally in figure 6, which represents the Spring it can be seen that the data still has trouble when there are sharp changes in the graph, which is consistent with the other three seasons.

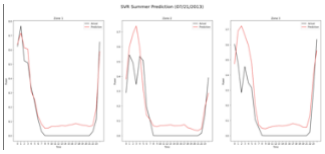

Fig 7. SVR Summer Day prediction vs actual for all three zones

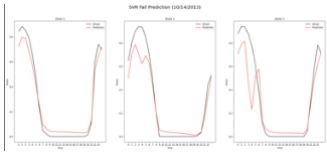

Fig 8. SVR Fall Day prediction vs actual for all three zones

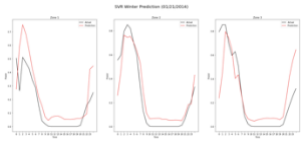

Fig 9. SVR Winter Day prediction vs actual for all three zones

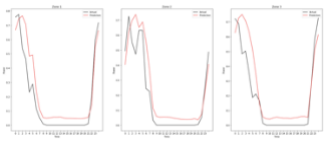

Fig 10. SVR Spring Day prediction vs actual for all three zones

When looking at the same summer day's results with the SVR, Figure 7 demonstrates that the SVR does a better job to estimate the lows of the power output at night. In the Summer it tends to sometimes overestimate the power generation. Figure 8 shows SVRs prediction on a fall day, and it shows that the results are fairly accurate, especially when looking at Zone 1 (left). Figure 9 shows the winter day results and again the predicted power output follows closely to the actual output, although there is again some potential overestimation at the peaks of the day. Finally figure 10 demonstrates the similar results to the previous, in the it often correctly estimates the power when there are not sharp or sudden changes in the actual power output.

The results of these regressors are accurate in most cases, but they struggle to adapt to the sudden changes in real power output. As mentioned in Section III, the current power being produced, and the power that will be produced 24 hours from the measurements have a very strong correlation. The proposed reason for the inability to adapt to sudden changes in power output is that these regressors became reliant on the previous day's power output while being trained. If this was the case then the current days power output would not be able to help the regressors make an accurate decision about the next day's power output if there was to be a drastic change in power output 24 hours later.

For this reason, the current power output was removed from the dataset that was used to train both regressors. At this point the hyperparameters were optimized once again for this trimmed dataset and the regressions were run again. This was done in hopes to remove the regressor's reliance on the previous days power and help it to accurately predict days in which there is sudden change in solar power generation.

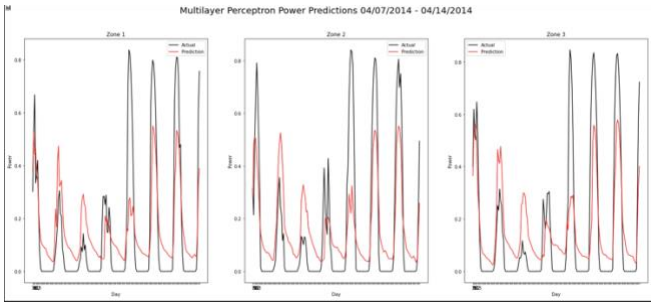

Fig. 11 MLP One week predictions vs actual with previous power used to train

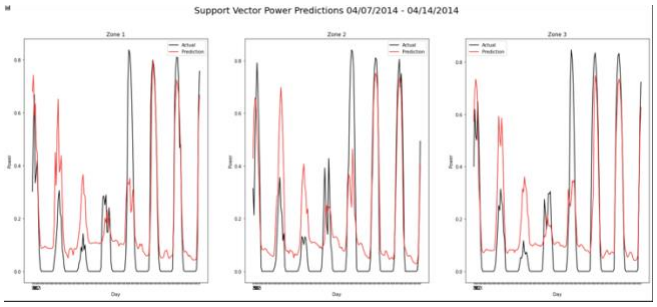

Fig 12. SVR One week predictions vs actual, with previous power used to train

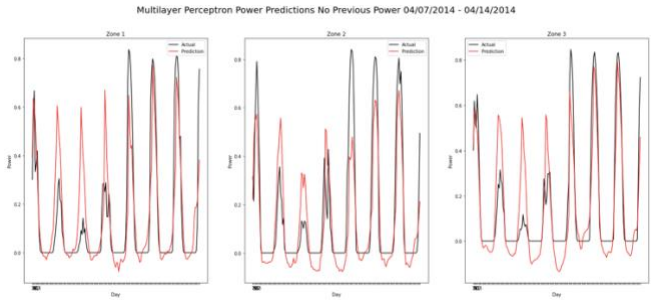

Fig. 13 MLP One week predictions vs actual without using previous power to train

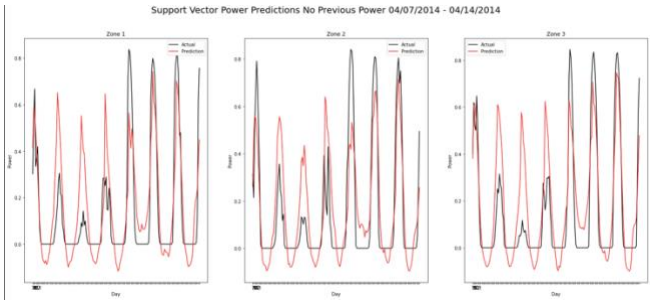

Fig 14. SVR One week predictions vs actual without using previous power to train

When retraining the MLP regressor, the goal was to increase its ability to adapt to days that have a sudden change in power. When comparing the results shown in Figure 11 and 13, we can see that this was not accomplished. The days that are relatively low in power output are overestimated even more now than they were in previous. The issue of the MLP overestimating power output at night, however, has been resolved using this strategy, and it now actually dips into the negatives on most nights. The MLP regressor became more accurate after training it without the current power as an input variable, as is shown in table 3.

Table 3. MAE and RMSE for MLP

ZONEID	1	2	3	Overall
MAE	0.082	0.088	0.093	0.088
RMSE	0.017	0.017	0.019	0.018

Table 4. MAE and RMSE for SVR

ZONEID	1	2	3	Overall
MAE	0.095	0.095	0.100	0.097
RMSE	0.018	0.018	0.019	0.019

Previously, the SVR results had been more accurate than the results of the MLP. After removing this current power variable, we can see that the MLP becomes more accurate than both the previous SVR, and the new SVR that is trained without using the previous power. Although the inability to adjust for days with drastic changes in power has gotten worse instead of better, the overall error has been improved due to the regressors ability to better estimate the highs and lows of each day's solar power generation.

The results of the SVR before and after this data trim can be seen in seen in Figures 12 and 14. Similar to the MLP, the ability to predict days with sudden changes in peak power output has also decreased. It also has gained a better ability to predict the highs and lows of the day's solar generation output. The SVR, however, did not see the same accuracy improvement that the MLP did. The error remained relatively similar, and only had a slight increase. This is believed to be because the SVR was already more accurate in estimating the highs and lows of the output, and the improvement in these areas is what caused the overall error of the MLP regressor to decrease.

V. Conclusion

In the paper, both MLP and SVR regressors are created. These regressors are used to predict the 24 hour ahead solar power generation of three zones. The dataset is split into the three zones, and these zones are analyzed invidivually. These regressors use 12 weather variables along with the current power to predict the power that will be generated at the same time the next day. This paper concludes that with the current power being used as a training variable, the SVR is able to produce more accurate results, but when it is removed the MLP becomes the more accurate method to forecast 24-hour ahead power generation. The MLP, when not using current power in training, can predict solar power generation 24 hours ahead with a mean average error of 0.082 and a root mean squared error of 0.017

Acknowledgment

The author would like to acknowledge the European Centre for Medium-Range Weather Forecasts for providing the data used in this paper. The author would also like to acknowledge

Feng Yan and Lei Yang of the University of Nevada, Reno who's lectures on machine learning techniques were used frequently to accomplish the task of solar power generation forecasting

REFERENCES

[1] ECMWF Solar Power Generation Dataset. 2014. https://www.ecmwf.int/en/forecasts/datasets

[2] Ahmed Mohammed, A.; Aung, Z. Ensemble Learning Approach for Probabilistic Forecasting of Solar Power Generation. *Energies* **2016**, *9*, 1017. https://doi.org/10.3390/en9121017

[3] Li, Z.; Rahman, S.M.; Vega, R.; Dong, B. A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting. *Energies* **2016**, *9*, 55. https://doi.org/10.3390/en9010055

[4] Yunchan Liu, Amir Ghasemkhani, Lei Yang, Jun Zhao, Junshan Zhang, Vijay Vittal; Seasonal Self-evolving Neural Netwroks Based Short-term Wind Farm Generation Forecast

[5] API design for machine learning software: experiences from the scikit-learn project, Buitinck *et al.*, 2013.