Impact of Correlation-based Feature Selection on Photovoltaic Power Prediction

Jung-Hyok Kwon
Smart Computing Laboratory
Hallym University
Chuncheon, South Korea
jhkwon@hallym.ac.kr

Sang-Woo Lee, Sol-Bee Lee, Eui-Jik Kim*
School of Software
Hallym University
Chuncheon, South Korea
{Sam_2011, thfqla3535, *ejkim32}@hallym.ac.kr

***Abstract*—This paper empirically presents the impact of the correlation-based feature selection on the accuracy of the photovoltaic (PV) power prediction, and then selects the weather variables that maximize prediction accuracy. To this end, the experiments are conducted using the weather dataset consisting of eighteen weather variables (i.e., features). For experiments, we first calculate a correlation coefficient of each weather variable by analyzing the correlation between PV power and each weather variable. Then, we create the subsets of weather variables considering the absolute value of correlation coefficient and generate the multiple prediction models using the created subsets. Finally, the accuracy of the generated prediction models is compared with each other to find the most accurate prediction model. The experiment results provide a reference guideline for selecting the weather variables that maximize the accuracy of PV power prediction.**

***Index Terms*—Correlation coefficient, feature selection, machine learning, photovoltaics power prediction, weather variables**

I. Introduction

As solar energy has been considered as one of the most promising renewable resources, the photovoltaic (PV) system, also known as the solar power system, has received considerable attention from academia and industry [1, 2]. The PV system aims to generate electricity from solar energy and uses PV cells to convert solar energy into electric power [3]. The output of the PV system (i.e., PV power) dynamically varies according to the environmental conditions [4]. This variability and uncertainty of PV power may adversely affect the PV system operation with respect to reliability and stability [5–7]. Therefore, an accurate prediction of PV power is regarded as one of the key challenges for the PV system.

In general, a prediction model for PV power is generated based on weather data since the PV power is greatly affected by weather conditions [8]. A variety of weather variables (i.e., features), such as solar radiation, temperature, humidity, wind speed, etc., can be used for generating the prediction model. However, some weather variables have a low correlation with PV power, which may degrade the accuracy of the prediction model. Therefore, for accurate prediction, it is necessary to select weather variables that are highly related to PV power before generating the prediction model.

In this paper, we analyze the impact of the correlation-based feature selection on the accuracy of the PV power prediction, with the aim of selecting the weather variables that maximize the prediction accuracy. To this end, we conduct experiments using R version 3.4.3 and use the weather dataset consisting of eighteen weather variables (i.e., features) and the PV power dataset measured every fifteen minutes. To investigate the impact of the correlation-based feature selection on the prediction accuracy, we first analyze the correlation between PV power and each weather variable and obtain a correlation coefficient of each weather variable. Then, the subsets of weather variables are created considering the absolute value of correlation coefficient and are used to generate the multiple prediction models. Finally, the accuracy of each prediction model is calculated through root mean square error (RMSE) and is compared with each other to find the most accurate prediction model. The experiment results show that correlation-based feature selection improves the accuracy of PV power prediction.

The rest of this paper is organized as follows. In Sect. 2, the correlation-based feature selection is presented in detail. In Sect. 3, the experiment results are presented. Finally, Sect. 4 concludes this paper.

II. Correlation-based feature selection

In this section, we describe the correlation-based feature selection with the aim of creating the subsets of weather variables. In this paper, the correlation-based feature selection is conducted based on the filter method, which acts as preprocessor by selecting the weather variables that are highly related to PV power.

To this end, we calculate the correlation coefficient of each weather variable, which is a numerical measure of the relationship between two variables [9]. It is represented in the range of –1.0 to 1.0. The values of –1.0 and 1.0 indicate strong negative and strong positive relationships between two variables, respectively. On the other hand, the value of zero indicates no relationship between two variables. To calculate the correlation coefficient, we use Pearson correlation coefficient (PCC) that measures the linear correlation between two variables [10]. For a pair of variables (X, Y), the correlation coefficient (ρ) can be calculated by:

978-1-7281-3755-1/19/$31.00 ©2019 IEEE

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

where $\mathrm{cov}(X,Y)$ is covariance of variables (X,Y), and σ_X and σ_Y are the standard deviation of X and Y, respectively.

To create subsets of weather variables, we sort the weather variables in order of the absolute value of the correlation coefficient. Then, the weather variables are selected whose absolute value of the correlation coefficient is greater than the predefined threshold.

III. Experiments

To investigate the impact of the correlation-based feature selection on the accuracy of the PV power prediction, the experiments are conducted using R version 3.4.3. We use PV power dataset measured every fifteen minutes at latitude 37.490026 and longitude 126.977463 (i.e., Seoul, South Korea). The weather dataset is collected through an open application programming interface (API) provided by the Korea meteorological administration (KMA) [11]. The collected weather datasets consist of eighteen weather variables that are listed in Table I.

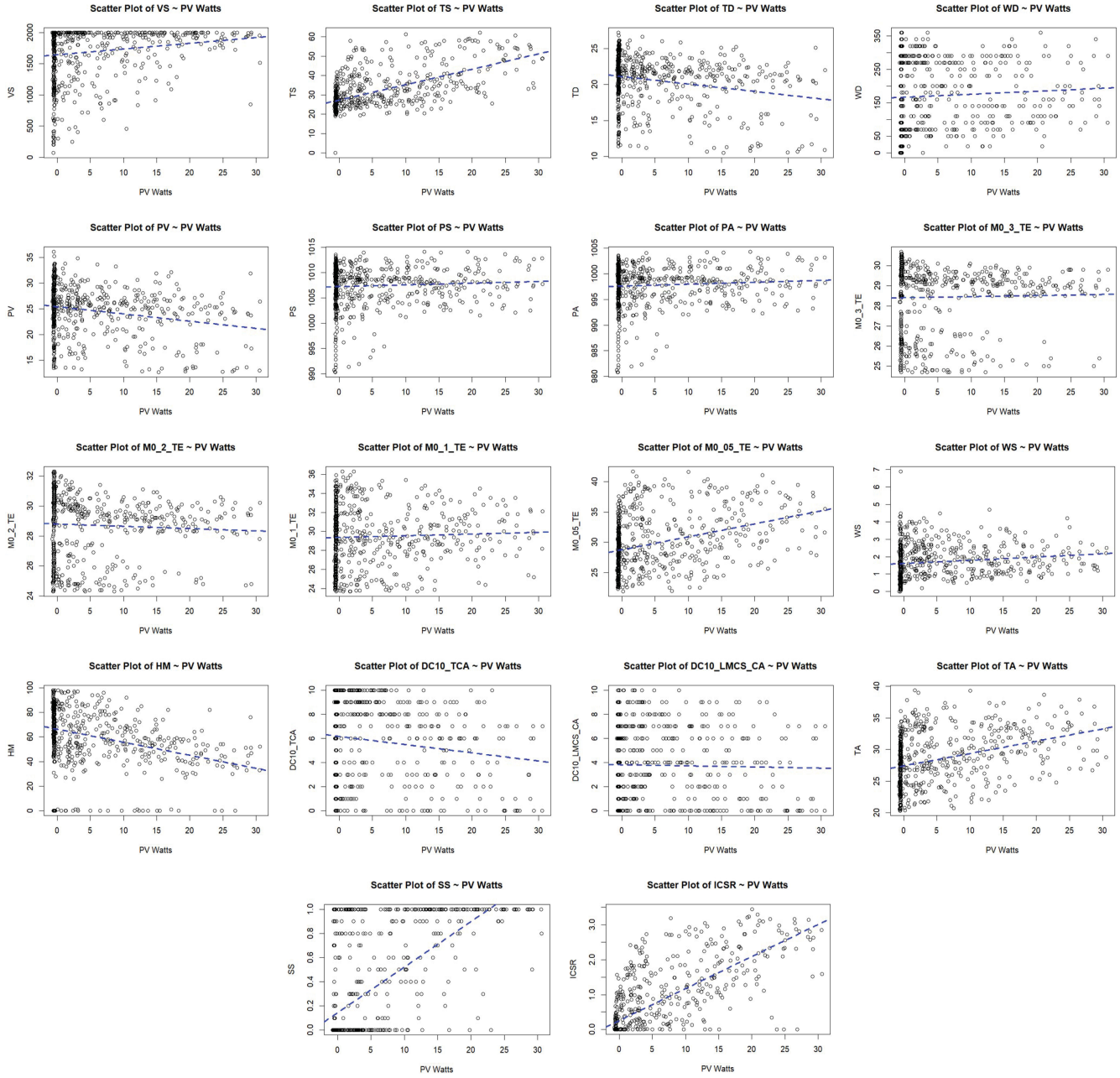

Fig. 1. Correlation between PV power and weather variables.

TABLE I. WEATHER VARIABLES

Weather Variable	Symbol
Surface temperature	TS
Visibility	VS
Dew point temperature	TD
Wind direction	WD
Vapor pressure	PV
Sea level pressure	PS
Local air pressure	PA
30cm earth temperature	M0_3_TE
20cm earth temperature	M0_2_TE
10cm earth temperature	M0_1_TE
5cm earth temperature	M0_05_TE
Wind speed	WS
Humidity	HM
Total cloud amount	DC10_TCA
Medium-level cloud amount	DC10_LMCS_CA
Temperature	TA
Sunshine	SS
Solar radiation	ICSR

Figure 1 shows the scatter plot of correlation between PV power and each weather variable, which is obtained through the modern applied statistics with S (MASS) package in R. In the figure, the linear correlation between PV power and each weather variable is depicted as a blue line. The scatter plots show that the solar radiation has the highest correlation with PV power. The correlation coefficient of each weather variable is given in Table II.

TABLE II. CORRELATION COEFFICIENT

Weather Variable	correlation coefficient	Weather Variable	correlation coefficient
TS	0.63	M0_1_TE	0.04
VS	0.15	M0_05_TE	0.35
TD	–0.22	WS	0.15
WD	0.06	HM	–0.34
PV	–0.22	DC10_TCA	–0.14
PS	0.06	DC10_LMCS_CA	–0.02
PA	0.07	TA	0.36
M0_3_TE	0.02	SS	0.67
M0_2_TE	–0.05	ICSR	0.75

To create the subsets of the weather variables, we set five thresholds from 0 to 0.4, and select weather variables whose absolute value of the correlation coefficient is greater than the specified threshold. The subsets of the weather variables (i.e., the results of the correlation-based feature selection) are shown in Table III.

TABLE III. RESULTS OF CORRELATION-BASED FEATURE SELECTION

Threshold	Subset of weather variables
0	TS, VS, TD, WD, PV, PS, PA, M0_3_TE, M0_2_TE, M0_1_TE, M0_05_TE, WS, HM, DC10_TCA, DC10_LMCS_CA, TA, SS, ICSR
0.1	TS, VS, TD, PV, M0_05_TE, WS, HM, DC10_TCA, TA, SS, ICSR
0.2	TS, TD, PV, M0_05_TE, HM, TA, SS, ICSR
0.3	TS, M0_05_TE, HM, TA, SS, ICSR
0.4	TS, SS, ICSR

The prediction model for PV power is generated through the support vector regression (SVR) that uses the same principle as a support vector machine (SVM), a classification algorithm [12]. Unlike SVM, SVR predicts real values rather than a class by performing regression. In the experiment, we generate five different prediction models using the results of the correlation-based feature selection. For this, we use e1071 package in R.

To evaluate the accuracy of the generated prediction models, RMSE is calculated. The RMSE is obtained by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \qquad (2)$$

where n is the number of samples, y_i is i-th actual value, and $\hat{y}_i$ is i-th predicted value.

TABLE IV. ACCURACY OF PV POWER PREDICTION

Threshold	0	0.1	0.2	0.3	0.4
RMSE	1.048	0.784	0.921	1.271	2.597

Table IV shows the accuracy of PV power prediction in detail. In the table, the RMSE is the lowest when the threshold is 0.1. This is because, the selected weather variables have a strong relationship between PV power, and the number of weather variables is enough to predict PV power accurately. Consequently, TS, VS, TD, PV, M0_05_TE, WS, HM, DC10_TCA, TA, SS, ICSR are selected as the subset of weather variables. Compared to the case that the feature selection is not performed (i.e., Threshold = 0), the case that weather variables whose correlation coefficient is greater than 0.1 are selected (i.e., Threshold = 0.1) obtains 33.7% lower RMSE.

IV. CONCLUSION

This paper investigates the impact of the correlation-based feature selection on the accuracy of the PV power prediction. To this end, we conduct experiments using the weather dataset consisting of eighteen weather variables (i.e., features). The results show that the prediction accuracy is improved when the correlation-based feature selection is performed. Specifically, when the weather variables whose correlation coefficient is greater than 0.1 are selected, RMSE of the prediction model is reduced by 33.7%. We expect that our study provides a

reference guideline to improve the accuracy of the PV power prediction.

ACKNOWLEDGMENT

This research was supported by the Ministry of Trade, Industry & Energy (MOTIE), Korea Institute for Advancement of Technology (KIAT) through the Encouragement Program for The Industries of Economic Cooperation Region (P0008630). This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03031055).

REFERENCES

[1] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE J. Power Energy Syst.*, vol. 1, no. 4, pp. 38−46, Dec. 2015.

[2] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. M. Pison, and F. A. Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78−111, Oct. 2016.

[3] S. Sobri, S. K. Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Convers. Manage.*, vol. 156, pp. 459−497, Jan. 2018.

[4] M. Pierro, F. Bucci, M. D. Felice, E. Maggioni, D. Moser, A. Perotto, F. Spada, and C. Cornaro, "Multi-model ensemble for day ahead prediction of photovoltaic power generation," *Solar Energy*, vol. 134, pp. 132−146, Sep. 2016.

[5] S. Alessandrini, L. D. Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Appl. Energy*, vol. 157, pp. 95−110, Nov. 2015.

[6] R. Bessa, A. Trindade, C. Silva, and V. Miranda, "Probabilistic solar power forecasting in smart grids using distributed information," *Int. J. Elect. Power Energy Syst.*, vol. 72, pp. 16−23, Nov. 2015.

[7] A. Tascikaraoglu, B. M. Sanandaji, G. Chicco, V. Cocina, F. Spertino, O. Erdinc, N. G. Paterakis, and J. P. S. Catalão, "Compressive spatio-temporal forecasting of meteorological quantities and photovoltaic power," *IEEE Trans. Sustain. Energy*, vol. 7, no. 3, pp. 1295−1305, Jul. 2016.

[8] Z. Ziadi, M. Oshiro, T. Senjyu, A. Yona, N. Urasaki, T. Funabashi, and C.-H. Kim, "Optimal voltage control using inverters interfaced with PV systems considering forecast error in a distribution system," *IEEE Trans. Sustain. Energy*, vol. 5, no. 2, pp. 682−690, Apr. 2014.

[9] S. K. Tyagi, "Correlation coefficient of dual hesitant fuzzy sets and its applications," *Appl. Math. Model.*, vol. 38, no. 22, pp. 659−666, Nov. 2014.

[10] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A new sampling method in particle filter based on Pearson correlation coefficient," *Neurocomputing*, vol. 216, pp. 208–215, Dec. 2016.

[11] [online] Available: https://data.kma.go.kr/api/selectApiList.do

[12] H. Xue, S. Chen, and Q. Yang, "Structural regularized support vector machine: A framework for structural large margin classifier," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 573−587, Apr. 2011.