An Ensemble Learning Model Based on Bayesian Model Combination for Solar Energy Prediction

Na Dong, Jianfang Chang, Wai Hung Ip and Kai Leung Yung

Abstract—**Due to the widespread promotion of renewable energy, solar energy has become a hot issue. However, as an important part of solar power system, photovoltaic grid-connected system and solar thermal system, solar irradiance has the inherent characteristics of variability and uncertainty. Hence, resource planners must be adaptable to accommodate these uncertainties while conducting planning, which is of great significance for the design and management of solar power systems. To improve the reliability of solar irradiance prediction methods, an ensemble learning method based on the Bayesian model combination has been developed in this paper for solar utilization systems. Firstly, clustering and cross-validation are introduced in the data sampling process to ensure that the training subsets are differentiated and uniformly sampled. Secondly, an ensemble learning model with multiple base learners has been designed, each training subset is utilized to train the corresponding base learner. Then, a Bayesian model combination is applied to frame the combination strategy based on the accuracy of each base learner on the validation set. The prediction values of multiple learners are framed through the model combination strategy. Finally, experiments are carried out using an open data set and the method is compared with Artificial Neural Network (ANN), K-means ANN, Support Vector Machine (SVM), and Multi-Kernel SVM. The effectiveness as well as the reliability of the proposed method in solar energy prediction have been found to perform better and have verified our approach.**

Index Terms—**Solar system, Power system, Bayesian model combination, Ensemble learning, Solar energy prediction system**

I. Introduction

As a kind of clean, substantial and renewable energy, solar energy does not produce pollutants [1-2], which are vital features that other kinds of energies do not possess. Therefore, solar energy has become the main energy source for research and development.

Two issues that need to be addressed in the design and utilization of renewable energy (solar energy) system are the stability of individual energy producers and the creation of viable grid-connected systems which can reasonably manage and schedule individual energy producers. A particularly relevant aspect is the creation of a system that brings together many unstable individual producers to form a more stable energy network system [3-4]. Due to factors such as solar elevation angle, temperature, humidity, location, altitude and other climatic factors, the energy generated by individual producers is mostly unstable [5], and thus, applications of solar energy are often restricted. Despite the rapid development of smart grid systems, the stability of individual generator is critical to energy conservation and rational utilization, as well as the grid security [6]. Therefore, solar irradiance prediction is of vital importance to the stable operation of the entire grid system and the formulation of energy dispatching plans.

Short-term predictions of solar energy are extremely critical [2]. Support Vector Machine (SVM) [7-10] and Artificial Neural Network (ANN) [11-13] are mainly been applied to the prediction of solar power, which are prone to train the prediction model through accuracy [14]. However, the reliability of prediction results is more significant in applications [15]. Ensemble Learning [16] provides inspirations for improving the reliability of prediction results. Ensemble learning combines multiple base learners together to achieve better generalized performance and reliability than a single learner.

In this paper, a Bayesian model combination based ensemble learning (BMC-EL) prediction method has been proposed for solar energy prediction to improve the reliability of prediction methods. Firstly, clustering [17] and cross-validation [18] has been introduced in the data sampling process to generate multiple training subsets so as to ensure that the training subsets are differentiated and uniformly sampled. Secondly, an ensemble learning model with multiple base leaners is established, each training subset is utilized to train the corresponding base learner. Here Random Forest is applied as base learner for ensemble learning. Then, a Bayesian model combination [19] is applied to frame the combination strategy based on the accuracy of each base learner on the validation set. The prediction values of multiple learners are framed through the model combination strategy.

To verify the accuracy and reliability of the proposed method in solar energy prediction, the American Meteorological Society 2013-2014 Solar Energy Prediction Contest dataset [20] is introduced for experiments. Multikernel-SVM, K-means-RBF, as well as classical SVM and ANN are introduced to establish comparison tests. The experimental results verify the accuracy and reliability of the proposed

This work was supported by the Natural Science Foundation of China under 61773282 and the H-ZG3K project.

*Na Dong, is the corresponding author of this paper. She is an associate professor in School of Electrical and Information Engineering, Tianjin University, China. (e-mail: dongna@tju.edu.cn).

Jianfang Chang, is a Ph.D. candidate in School of Electrical and Information Engineering, Tianjin University, China. (e-mail: changjianfang@tju.edu.cn).

Wai Huang Ip is Principal Research Fellow of the Department of Industrial and Systems Engineering of Hong Kong Polytechnic University, and adjunct professor of University of Saskatchwan. (e-mail: wh.ip@polyu.edu.hk).

Kai Leung Yung is currently an associate head and chair professor of the Department of Industrial and Systems Engineering of Hong Kong Polytechnic University. (e-mail: kl.yung@polyu.edu.hk).

method in solar energy prediction.

II. Data Sampling

The ensemble learning model will have better performance when the diversity between base learners are more significant [19]. Therefore, the differences of training subsets should be increased to improve the diversity of the base learners.

In order to alleviate the above problems, K means clustering and K fold cross validation have been applied at the same time to increase the diversity of training subsets, as shown in the Fig.1, which also ensures uniform sampling. (*In order to distinguish the subscripts of K means clustering, K fold cross validation is replaced by M fold cross validation.*)

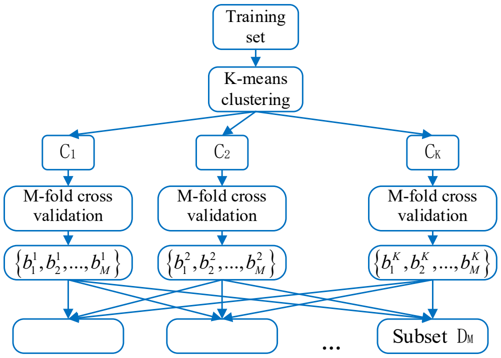


Fig.1. Schematic diagram of data sampling

Assuming that the training set is designed to be sampled as training subsets $\{D_1, D_2, ...D_M\}$. The training set is firstly divided into clusters $\{C_1, C_2, ..., C_k\}$ through K means clustering, where each cluster contains samples with similar (adjacent) weather conditions. After M fold cross validation, clusters C_1 can be randomly separated into packages $\{b_1^1, b_2^1, ..., b_M^1\}$. Import $\{b_2^1, b_3^1, ..., b_M^1\}$ into training subset D_1 and import $\{b_1^1, b_3^1, ..., b_M^1\}$ into training subset D_2, similarly, different M-1 packets are imported into the corresponding training subsets until $\{b_1^1, b_2^1, ..., b_{M-1}^1\}$ is imported into training subset D_M.

Since each cluster $\{C_1, C_2, ..., C_k\}$ contains samples with similar weather conditions. When each cluster is divided into different training subsets by M fold cross validation, each training subset can cover samples with different weather conditions, which alleviates the non-uniform sampling that traditional random sampling process may cause. Data sampling also increase the diversity of base learners.

III. Base Learner

Ensemble learning can reduce the prediction risk through the combination of multiple base learners. To meet the stringent requirements of applications, ensemble learning is utilized to improve the reliability of solar energy prediction. The pruning operation of Classification and Regression Tree (CART) in Random Forest can effectively alleviate the risk of over-fitting. Random Forest is simple, efficient and easy to implement, equipped with the characteristic of small computational cost and excellent generalization ability. So Random Forest has been defined as the basic learner of ensemble learning. The individual Random Forest is composed of multiple CART, and the divided training subsets are prepared for training individual Random Forests.

Supposing that the training set is $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. The node of the CART sets a segmentation point *s* for an attribute variable *j* of the sample x_i, samples with input variable greater than *s* are divided into one part $R_1(j,s) = \{x_i \mid x_i^{(j)} > s\}$, otherwise divided into other part $R_2(j,s) = \{x_i \mid x_i^{(j)} < s\}$. The partitioned parts are further divided by other different attribute variables, and the samples are divided into *m* parts according to the node segmentation points, which are respectively denoted as R_1, R_2, ..., R_m. The corresponding output for each part is defined as c_1, c_2, ..., c_m respectively. The CART could be described as formula (1):

$$f(x) = \sum_{m=1}^{m} c_m I(x \in R_m) \tag{1}$$

Where $I(x \in R_m) = \begin{cases} 1 & (x \in R_m) \\ 0 & (x \notin R_m) \end{cases}$. The square error of CART is as follows:

$$E = \sum_{x_i \in R_m} (y_i - f(x_i))^2 = \sum_{x_i \in R_m} \left(y_i - \sum_{m=1}^{m} c_m I(x \in R_m) \right)^2 \tag{2}$$

When c_m is equal to the average of samples output which belong to the R_m, the square error is optimal. So $c_m = ave(y_i \mid x_i \in R_m)$.

From the equation (2), the optimal output c_m of part R_m can minimize the square error. Traversing all the attribute variables *j* and the possible segmentation point *s* in the sample, the $R_m(j,s)$ whose excellent square error is the smallest are defined as part R_m. Similarly, the partitioned parts are further divided and the optimal segmentation variables and segmentation points ($R_m(j,s)$) could be obtained. The final CART model is $f(x) = \sum_{m=1}^{m} \hat{c}_m I(x \in R_m)$.

The average of multiple CART outputs is defined as the output of an individual Random Forest.

IV. The Combination Strategy

Ensemble learning can alleviate the poor generalization ability caused by the base learner error, which relies on ingenious model combination strategy. The Bayesian model combination defines the posterior probability of the model's prediction performance on the validation set as the weight of the model, assigns multiple Random Forest models with reasonable weights, and selects one best model combination strategy from a plurality of combination strategies. Here

Bayesian model combination has been utilized to generate Random Forest combination strategy.

In the Bayesian model averaging, assuming there are n samples in dataset D, and each sample d_i consists of attribute vector x_i and real value y_i. Hypothetical space H contains a finite number of individual hypotheses and h represents an individual hypothesis of hypothetical space. Under the preconditions of hypothetical space H and dataset D, the posterior distribution of y_i is as follows:

$$p\left(y_i\left|x_i,D,H\right.\right)=\sum_{h\in H}p\left(y_i\left|x_i,h\right.\right)p\left(h\left|D\right.\right) \qquad (3)$$

Where, $p\left(y_i\left|x_i,D,H\right.\right)$ is a weighted average of posterior distributions $p\left(y_i\left|x_i,h\right.\right)$ under all individual hypothesis, $p\left(y_i\left|x_i,h\right.\right)=\int p\left(y_i\left|\theta_k,h,D\right.\right)p\left(\theta_k\left|h,D\right.\right)d\theta_k$ is the posterior distribution of y_i under the individual hypothesis h, and θ_k is the parameter vector corresponding to the individual hypothesis h.

The posterior probability $p\left(h\left|D\right.\right)$ of individual hypothesis h under the condition of dataset D can be calculated by equation $p\left(h\left|D\right.\right)=\dfrac{p\left(D\left|h\right.\right)p\left(h\right)}{\sum_{h\in H}p\left(D\left|h\right.\right)p\left(h\right)}$. Here $\sum_{h\in H}p\left(D\left|h\right.\right)p\left(h\right)$ is a constant, so $p\left(h\left|D\right.\right)\propto p\left(D\left|h\right.\right)p\left(h\right)$. $p\left(D\left|h\right.\right)=\int p\left(D\left|\theta_k,h\right.\right)p\left(\theta_k\left|h\right.\right)d\theta_k$ is the integral likelihood estimate of the individual hypothesis h, $p\left(\theta_k\left|h\right.\right)$ is the prior distribution of the vector parameter θ_k corresponding to individual hypothesis h, and $p\left(D\left|\theta_k,h\right.\right)$ is the likelihood estimation. $p\left(h\right)$ is the priori probability of individual hypothesis h.

In order to ensure that all base learners have higher predictive performance, there is no difference in base learner parameter initialization, so prior knowledge probability $p\left(h\right)$ does not need to be "biased" on any individual hypothesis, so $p\left(h\right)=\dfrac{1}{k}$ (k is the number of individual hypotheses in the hypothetical space).

In the Bayesian model averaging, the calculation of the integral likelihood estimate sets a very high weight on the hypotheses that makes accuracy slightly increased [21], the Bayesian model averaging is easier to over fitting than stacking [22].

In order to alleviate the above condition, the Bayesian model averaging can be modified into the Bayesian model combination (BMC) method. Equation (3) is modified to equation (4):

$$p\left(y_i\left|x_i,D,H,E\right.\right)=\sum_{h\in E}p\left(y_i\left|x_i,H,e\right.\right)p\left(e\left|D\right.\right) \qquad (4)$$

Where e is the individual hypothesis model in the combined model space E. Bayesian model averaging and Bayesian model combination are shown in Fig.2 and Fig.3 below respectively:

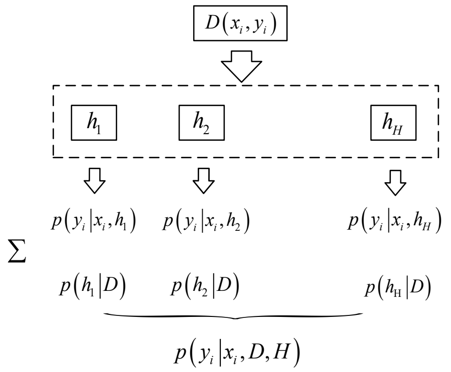


Fig. 2. Bayesian model averaging

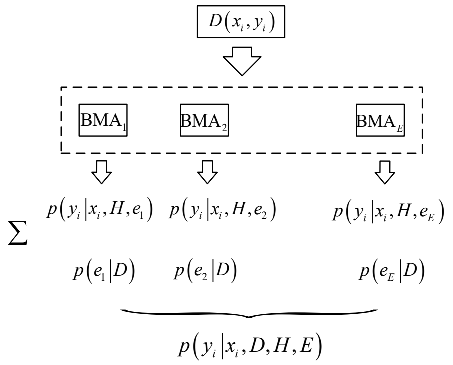


Fig. 3. Bayesian model combination

V. Ensemble Learning Based on Bayesian Model Combination

The flow chart of the ensemble learning prediction method based on Bayesian model combination is as follows:

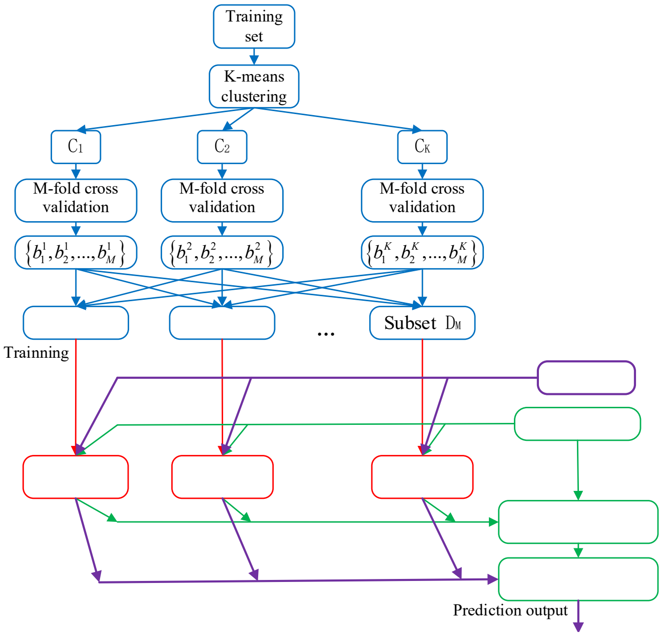


Fig. 4. Ensemble Learning Based on Bayesian Model Combination

Implementation steps of ensemble learning prediction method based on Bayesian model combination are as follows:

1. The original data are normalized by the formula $x^* = \frac{x-\min(x)}{\max(x)-\min(x)}$, and cluster $\{C_1, C_2, ..., C_k\}$ is generated by K-means clustering, M-fold cross validation is performed on each cluster, and training subsets $\{D_1, D_2, ...D_k\}$ are sequentially generated.

2. Training subsets are used to train base learners (Random Forest) of ensemble learning.

3. Import validation sets into random forests and output predicted values $(y_1, y_2, ..., y_k)$. Assuming that the real output of the verification set is y, a matrix $(y, y_1, y_2, ..., y_k)$ is constructed and imported into a Bayesian model combination. Bayesian model combinations can make the optimal model combination strategies $p(y_i|x_i, D, H, E) = \sum_{h\in E} p(y_i|x_i, H, e)p(e|D)$.

4. Import test sets into random forests and output predicted values $(\overline{y}_1, \overline{y}_2, ..., \overline{y}_k)$. The final prediction output of the integrated learning algorithm is $p(Y|\overline{y}_k, D, H, E) = \sum_{h\in E} p(y_i|x_i, H, e)p(e|D)$.

VI. Simulation Study

To test the performance of the proposed method in solar irradiance prediction applications, the American Meteorological Society 2013-2014 Solar Energy Prediction Contest dataset [17] is introduced to establish prediction experiments.

A. Performance Indicators

The average error rate (*AER*) and the rate of success (*RS*) are introduced here, as described in formula5-7:

$$Er = \frac{|Y_{pre} - Y_{real}|}{Y_{pre}} \tag{5}$$

$$AER = \frac{\sum_{i=1}^{Num} Er(i)}{Num} \tag{6}$$

$$RS(0.1) = \frac{num(Er < 0.1)}{Num} \tag{7}$$

Where Y_{pre} is the prediction output, Y_{real} is the real data, E_r is the error rate for each sample, and *AER* is average error rate. *Num* is the total number of samples in the test set, and *num* is the number of samples whose error rate is less than 0.1. The *AER* reflects the average value of prediction error. *RS* reflects the reliability of the prediction method.

B. Diversity of Training Subset

Normalize the raw meteorological data to [-1,1], and then perform a 10 means clustering operation on the training set to divide the training set into 10 clusters $\{C_1, C_1, ..., C_{10}\}$. The dswrf_sfc, dswrf_sfc, and tmp_sfc of the samples are applied to establish a three-dimensional coordinate, the samples in the training set are distributed in the coordinate as shown in Fig.5.

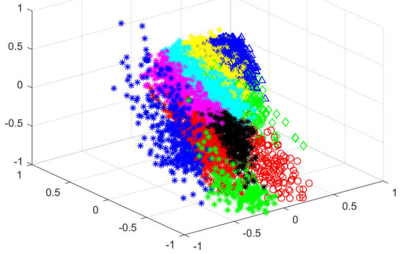


Fig.5. Distribution of training set

The divided clusters are respectively subjected to 10 fold cross-validation, and 9 different packages are respectively introduced into each training subset. The distribution of samples under different weather conditions in the training subset is shown in Fig. 6.

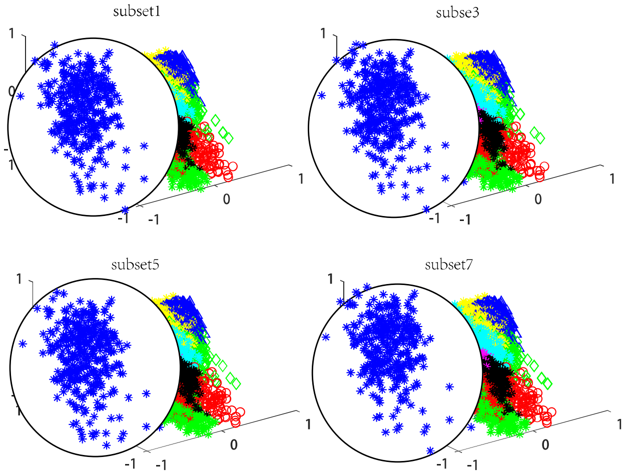


Fig.6. Distribution of training subsets

(*4 typical training subsets are given to illustrate the distribution*)

Since clustering is followed by 10 fold cross-validation, the size of each training subset is 90% of the training set. According to distribution of samples in training subsets, the sampling process does not affect the sample distribution of different weather conditions. On the other hand, the cross-validation operation of the clusters also increases the diversity of training subsets (as samples of the marked area), and correspondingly increases the sample perturbation of the base learners.

C. Model Error Estimation and Performance Testing

Random forests estimate model errors by out of bag (OOB). Due to the attribute disturbance, the Random Forest will converge to a lower generalization error only when there is a certain amount of individual CART. Importing training set, the relationship between OOB error and the number of CART is shown in Fig. 7.

As can be seen from Fig. 7, when the number of CART reaches 200, the OOB error of Random Forests tends to decrease slowly. In order to make the random forest has a lower error and save the computational cost, the number of CART is set to 200.

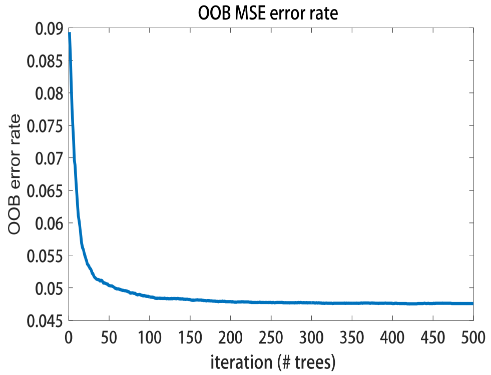
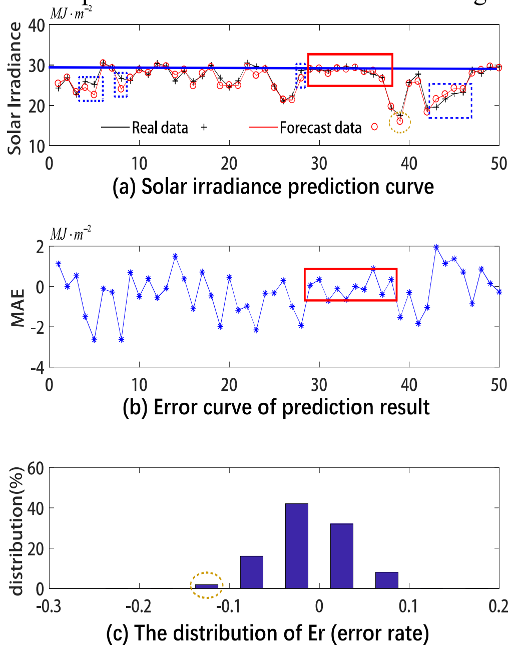


Fig. 7. Error estimation of random forest

The solar irradiance of the HOBA Mesoscale station and the meteorological data of the surrounding GEFS stations from 1994 to 2007 (5113 samples) are introduced to test proposed method. In chronological order, the first 4000 samples are used as the training set, the next 500 samples are defined as the verification set, and the samples 4501-4550 are defined as the test set. The experimental results are as shown in Fig.8.



(a) Solar irradiance prediction curve

(b) Error curve of prediction result

(c) The distribution of Er (error rate)

Fig. 8 Prediction performance of proposed method

Fig 8.(a) depicts the real values and prediction values of 50 samples, while Fig 8.(b) illustrates the MAE of 50 samples predictions, of which $MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_{pre}^{(i)} - y_{real}^{(i)}\right|, (m=1)$. Fig 8.(c) shows the distribution of prediction error rate (*Er*).

In Fig 8.(a), the long blue line indicates the average of solar irradiance under fine weather conditions. The samples near the blue line have better prediction performance, indicating that the proposed method has capability for solar irradiance prediction under fine weather conditions.

The samples in the blue dotted box of Fig.8.(a) show slightly larger errors. These samples are far away from the blue line, indicating that these samples have low solar irradiance and complex meteorological conditions which causes big interference. Although the sample in the third dotted box is closer to the blue line, the first two samples of this sample appear to fluctuate greatly, indicating that fluctuating meteorological conditions have a significant impact on the prediction of solar energy. Continuously fine weather conditions are of great benefit for accurate prediction, such as the samples framed by the red box, where their MAEs are kept within $1 MJ \times m^{-2}$.

The sample in the yellow dotted ellipse of Fig.8.(a) has very low solar irradiance, suggesting bad weather conditions. However, the proposed solar prediction method still guarantees stable prediction performance. Although its prediction error is less than $2 MJ \times m^{-2}$, the evaluation of this sample is not dominant when calculating the error rate due to its low solar irradiance (as the sample in the dotted ellipse of Fig.8.(c)).

It can be drawn from the distribution graph that the error rate(Er) of more than 70% samples are less than or equal to ±2.5%, and the error rate of other samples is less than or equal to ±7.5% except for extreme weather conditions.

D. *Establishment of Experiments*

The comparison experiment settings are shown in Tab.1. EL utilizes the average of multiple base learners as the final output, and other settings are consistent with BMC-EL.

Tab.1 Parameter settings of solar energy prediction comparison experiment

Method	Parameter settings
ANN [11]	$N15\text{-}24\text{-}24\text{-}1\quad trainParam.epochs = 1$ $trainParam.goal = 0.00001\quad a = 0.1$
K-means_ANN [13]	$density\ coefficient\ \psi = 0\quad overlap\ coefficient\ \varepsilon = 1$ $cluster\ radius\ a = 1\quad N_{15\text{-}24\text{-}24\text{-}1}\quad trainParam.epochs = 1$ $trainParam.goal = 0.00001\quad a = 0.1$
SVM [7]	$cost = 1\quad gama = 1\quad model = epsilon\text{-}SVR$ $epsilon = 0.01\quad kernel = RBF$
MultiKernel_SVM [9]	$cost = 1\quad gama = 1\quad model = epsilon\text{-}SVR$ $epsilon = 0.01$ $k(x,y) = 0.15 * [x^T y + c] + 0.15 * [(ax^T y + c)^d]$ $+ 0.5 * e^{-g\|x-y\|^2} + 0.2 * e^{-\frac{\|x-y\|^2}{2s^2}}$

E. *Experimental Results and Observations*

The solar irradiance of the HOBA Mesoscale station and the meteorological data of the surrounding GEFS stations are defined as dataset. The samples from January 1, 1994 to December 31, 2004 are defined as training set. The samples from January 1, 2005 to December 31, 2006 are defined as a validation set. The samples of 2007 are utilized as test set for solar energy prediction experiments.

In order to analyze the prediction performance of different methods under different meteorological conditions, four representative months (February, May, August, and November) are selected to demonstrate the prediction output as well as prediction error, as shown in Fig.9-Fig.12.

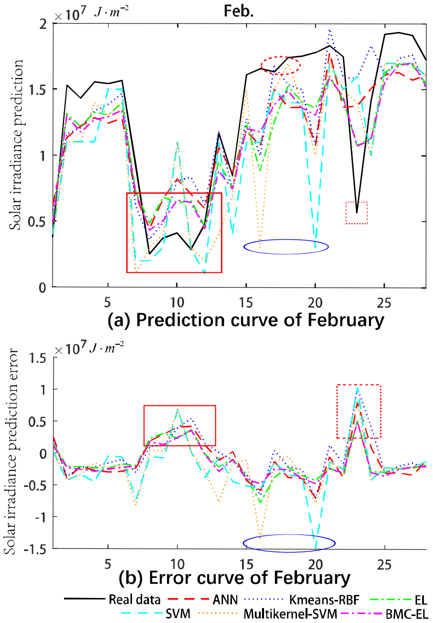


Fig.9. Prediction performance of different methods in February

As shown in Fig.9.(a), the solar irradiance of the samples in the red box is low, indicating complex meteorological conditions which may causes interference to solar irradiance prediction. Different prediction methods can basically track the trend of solar irradiance on these harsh samples, and the prediction curves of SVM and MultiKernel-SVM show large spikes (errors). Although SVM and MultiKernel-SVM have better prediction accuracy on two samples in the red dotted ellipse in Fig 9. (a), other tow spikes occur on the blue ellipse-framed samples. The prediction curves of SVM and MultiKernel-SVM show large spikes respectively, although we have maximally alleviated over-fitting and under-fitting. The solar irradiance of the sample in red dotted box of Fig.9.(a) returns to normal after a sharp decay, where the prediction outputs of different methods show a large deviation. The error directions of the different prediction methods are consistent in the red dotted box of Fig.9.(b), and the peak amplitude of the BMC-EL method is the smallest. The output curve of the BMC-EL on the harsh samples are closest to the real curve.

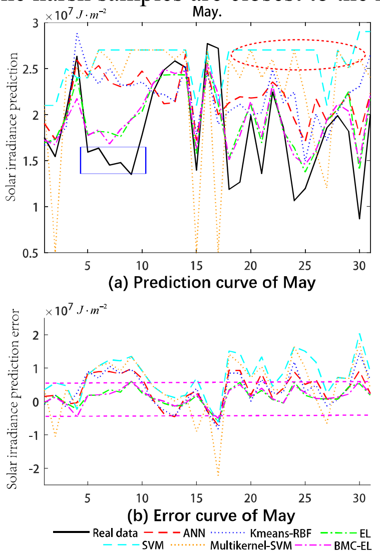


Fig.10. Prediction performance of different methods in May

As shown in Fig.10.(a), solar irradiance fluctuated sharply in May, which caused obstacles to solar prediction. Although numerous trials have been carried out to make the experiment results fitting, SVM and MultiKernel-SVM deviate significantly from the real value, as shown in the red ellipse. In the blue box-framed samples, the output curves of EL and BMC-EL are closer to the real curve, which shows that ensemble learning is more advantageous on harsh samples and BMC-EL has better performance. With the error curve shown in Fig.10.(b), the error curve boundary of the BMC-EL in the third graph is also leaner.

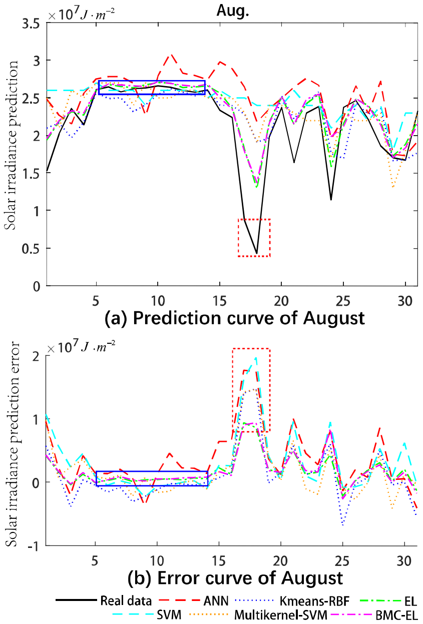


Fig.11. Prediction performance of different methods in August

Solar irradiance is abundant in August. As shown in Fig.11, continuously fine weather conditions are of great benefit for accurate prediction, such as samples in the blue box of Fig.11, which is consistent with the analysis of Part E. Consistent with the previous, such as samples in red dotted box of Fig.11.(b), BMC-EL maintains stable prediction performance in samples with large fluctuations.

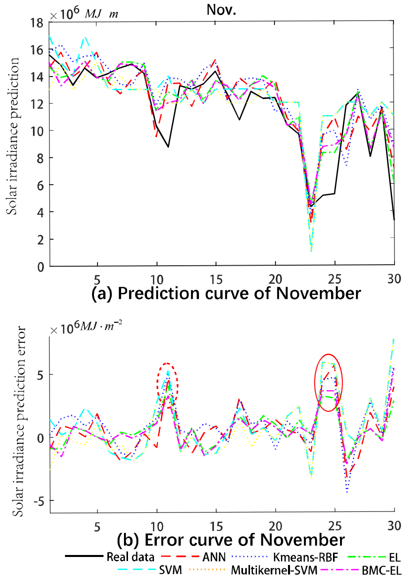


Fig.12. Prediction performance of different methods in November

As shown in Fig.12, since Bayesian model combination chooses the best model combination strategy from hypothetical space, BMC-EL can guarantee the stable prediction precision in some harsh prediction samples, and thus, BMC greatly improves the reliability of the prediction.

A cartesian coordinate is established, the abscissa indicates the measured value of solar irradiance (real value), and the vertical axis indicates the prediction value. The line y=x, defined as the baseline, indicating that the predicted value is equal to the real value. The prediction results of the remaining eight months are plotted in the cartesian coordinate, as shown in Fig. 13, the closer the point is to the baseline, the more accurate the prediction result is. The red dotted line divides the scatter plot into three parts, which respectively indicate that solar irradiance is scarce, fairish and rich.

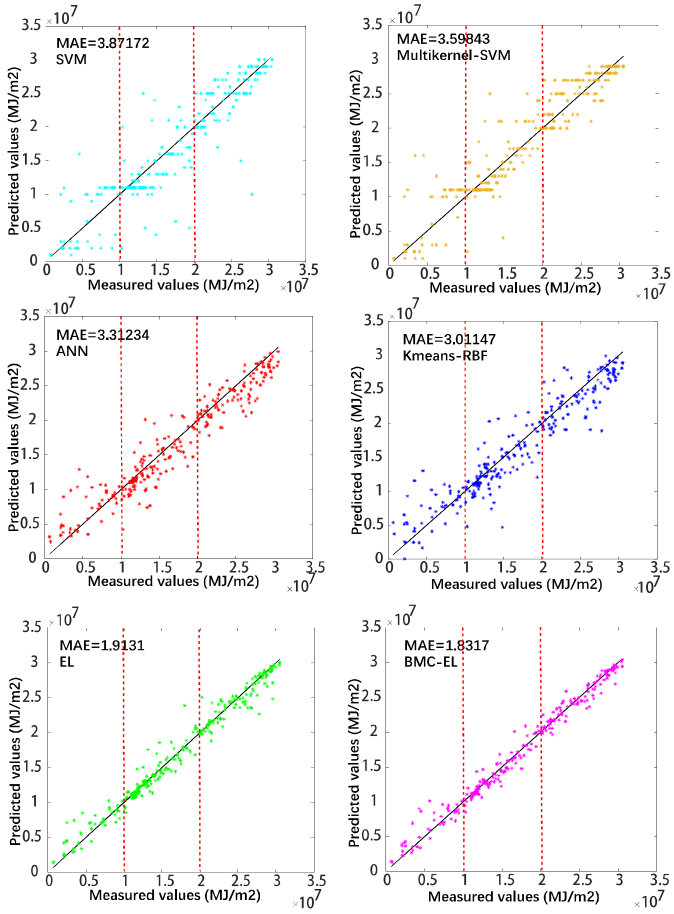


Fig.13 Scatter plot of solar irradiance prediction for the remaining eight months

In the solar irradiance prediction simulation experiment, the samples corresponding to different prediction methods fall near the baseline. We have done abundant tries to adjust the experimental parameters as much as possible and tried several times to ensure the optimal performance of the different prediction methods.

In the solar-rich part, meteorological conditions receive less interference, so samples of different prediction methods are centralized around the baseline. In the solar-scarce area, the meteorological environment is complex and the interference is large, and the samples of different prediction methods have large deviations. However, the scatter plot distribution of EL and BMC-EL methods is the most harmonious. Since Bayesian model combination chooses the best model combination strategy from hypothetical space, the scatter plot of the BMC-EL method is the most centralized.

The performance indicators of different prediction methods throughout the year are shown in Tab. 2, of which the proposed BMC-EL has the best performance.

Tab. 2 The performance indicators of different prediction methods

	MSE	MAE($MJ{\cdot}m^{-2}$)	RS	AER
BMC-EL	**6.79E+12**	**1.8628**	**53.87%**	**0.2085**
EL	7.53E+12	1.9002	51.42%	0.21
ANN	1.95E+13	3.2448	36.10%	0.31796
Kmeans-RBF	1.78E+13	3.01484	40.08%	0.2966
Multikernel-SVM	2.95E+13	3.62957	44.30%	0.3468
SVM	3.27E+13	3.88253	40.18%	0.3738

In summary, the random forest prediction method based on Bayesian model combination has good prediction performance and reliability in solar irradiance prediction, which can achieve accurate and stable prediction under different weather conditions.

VII. Conclusion

With the development of smart grids, the interconnections of renewable energy have become a hot issue in power system. Solar energy is rich but volatility, which makes stable solar prediction essential for interconnections and dispatch of power system.

An ensemble learning method based on Bayesian model combination is proposed for solar irradiance prediction, which aims to improve the reliability of solar power application. Firstly, the data sampling process not only ensures uniform sampling, but also improves the diversity of the training subset, which is able to improve the diversity of the base learner. Secondly, multiple training subsets are utilized to train the individual Random Forests in the ensemble learning. So far, the diversity of ensemble learning comes from the diversity of training subsets. After that, the Bayesian model combination is used to construct a combination for base learners. It formulates a combination strategy based on the performance of each base learning on the verification set by choosing the best model combination strategy from hypothetical space, and thus effectively improve the performance of the model. In solar energy prediction experiments, BMC-EL significantly reduces the uncertainty of a single learner by adding Bayesian model combination and increases the reliability of result. The experimental results show that the proposed prediction method has better prediction accuracy and reliability, which can be utilized to estimate dynamic fluctuating solar power. Therefore, the proposed method provides a basis for accurate estimation of solar power, which can promote further development of the whole power system.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of China under 61773282 and was partially supported by a grant of the Hong Kong Polytechnic university (H-ZG3K).

REFERENCES

[1] Wu Y, Wang J. A novel hybrid model based on artificial neural networks for solar radiation prediction[J]. Renewable Energy, 2016, 89:268-284.

[2] Mellit A, Pavan A M. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy[J]. Solar Energy, 2010, 84(5):807-821.

[3] Maity I , Rao S . Simulation and Pricing Mechanism Analysis of a Solar-Powered Electrical Microgrid[J]. IEEE Systems Journal, 2010, 4(3):275-284.

[4] Das M K , Jana K C , Sinha A . Performance evaluation of an asymmetrical reduced switched multi-level inverter for a grid-connected PV system[J]. IET Renewable Power Generation, 2018, 12(2):252-263.

[5] Lin J . Potential Impact of Solar Energy Penetration on PJM Electricity Market.[J]. IEEE Systems Journal, 2012, 6(2):205-212.

[6] Akram U , Khalid M , Shafiq S . Optimal sizing of a wind/solar/battery hybrid grid-connected microgrid system[J]. IET Renewable Power Generation, 2018, 12(1):72-80.

[7] Yang X, Jiang F, Liu H. Short-term solar radiation prediction based on SVM with similar data[C]// Renewable Power Generation Conference. IET, 2014:1.11-1.11.

[8] Guo W, Mingjia L I, Tao L I, et al. Parameter identification of Hammerstein ARMAX model based on APSO-WLSSVM algorithm[J]. China Sciencepaper, 2018.

[9] Alam S, Kang M, Pyun J Y, et al. Performance of classification based on PCA, linear SVM, and Multi-kernel SVM[C]// Eighth International Conference on Ubiquitous and Future Networks. IEEE, 2016:987-989.

[10] Zhou Y, Cui X, Hu Q, et al. Improved multi-kernel SVM for multi-modal and imbalanced dialogue act classification[C]// International Joint Conference on Neural Networks. IEEE, 2015:1-8.

[11] Rabbi K M, Nandi I, Saleh A S, et al. Prediction of solar irradiation in Bangladesh using artificial neural network (ANN) and data mapping using GIS technology[C]// Development in the in Renewable Energy Technology. IEEE, 2016:1-6.

[12] Anamika, Kumar N, Akella A K. Prediction and efficiency evaluation of solar energy resources by using mixed ANN and DEA approaches[C]// Pes General Meeting | Conference & Exposition. IEEE, 2014:1-5.

[13] Yadav A K, Malik H, Chandel S S. ANN based prediction of daily global solar radiation for photovoltaics applications[C]// India Conference. IEEE, 2016:1-5.

[14] Chen L G, Chiang H D, Dong N, et al. Group-based chaos genetic algorithm and non-linear ensemble of neural networks for short-term load forecasting[J]. Iet Generation Transmission & Distribution, 2016, 10(6):1440-1447.

[15] Baili H, Li Y F. Online reliability prediction of energy systems with wind generation[C]// IEEE, International Midwest Symposium on Circuits and Systems. IEEE, 2016:1-4.

[16] Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning[C]// International Conference on Neural Information Processing Systems. MIT Press, 1994:231-238.

[17] Daniel Aloise, Amit Deshpande, Pierre Hansen, et al. NP-hardness of Euclidean sum-of-squares clustering[J]. Machine Learning, 2009, 75(2):245-248.

[18] Wan J, Canedo A, Faruque M A A. Functional Model-Based Design Methodology for Automotive Cyber-Physical Systems[J]. IEEE Systems Journal, 2017, 11(99):1-12.

[19] Monteith K, Carroll J L, Seppi K, et al. Turning Bayesian model averaging into Bayesian model combination[C]// International Joint Conference on Neural Networks. IEEE, 2011:2657-2663.

[20] AMS 2013-2014 Solar Energy Prediction Contest, Forecast Daily Solar Energy with An Ensemble of Weather Models, https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest.

[21] Monteith K, Carroll J L, Seppi K, et al. Turning Bayesian model averaging into Bayesian model combination[C]// International Joint Conference on Neural Networks. IEEE, 2011:2657-2663.

[22] Clarke B, STAT. UBC. C A. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored[J]. Journal of Machine Learning Research, 2003, 4(4):683-712.

Na Dong received her Ph.D. degree in control theory and control application at Nankai University in 2011. She is currently an associate professor in School of Electrical and Information Engineering, Tianjin University, China.

Her current research areas encompass intelligent control algorithms, heuristic optimization algorithm, neural networks, data-driven control, deep learning and image processing.

JianFang Chang was born in June 1993. He is currently working toward the Ph.D. degree in the School of Electrical and Information Engineering, Tianjin University, China.

His current research areas encompass neural networks, deep learning, machine Learning, heuristic optimization algorithm, and image processing.

Wai Huang Ip received his Ph.D. degree from Loughborough University in the U.K., MBA from Brunel University, M.Sc. in Industrial Engineering from Cranfield University, and LLB (Hons) from the University of Wolverhampton. He is Professor Emeritus and adjunct professor of Mechanical Engineering, University of Saskatchewan and principal research fellow of the Hong Kong Polytechnic University. He is a chartered engineer and senior member of IEEE. His research interests are Space Systems, AI and deep learning.

Kai Leung Yung received his BSc in Electronic Engineering at Brighton University in 1975, MSc, DIC in Automatic Control Systems at Imperial College of Science, Technology & Medicine, University of London in 1976, and Ph.D. in Microprocessor Applications in Process Control at Plymouth University in 1985 in the United Kingdom and became a Chartered Engineer (C..Eng.,MIEE) in 1982. His research interests are Space Systems, Mechatronics and AI applications.