Evaluate Machine Learning Models Used for Upscaling Surface Ocean CO_2 Measurements

Jiye ZENG[1] and Zheng-Hong TAN[2]

[1]National Institute for Environmental Studies, Japan. <zeng@nies.go.jp>
[2]Department of Environmental Science, Hainan University, China. <tanzh@xtbg.ac.cn>

Abstract

Upscaling measurements from ground-based sites or underway monitoring to a regional or global scale provides important information to policy makers for environmental management and to researchers looking for a better understanding of relevant issues. We used the reconstruction of global surface ocean CO_2 as an example to evaluate the performance of four machine learning models: Random Forest (RF), Support Vector Machine (SVM), Feedforward Neural Network (FNN), and Self-Organization Map (SOM). The results show the performance from high to low as RF, SVM, FNN, and SOM. However, the overall differences of modelled CO_2 among the four models are insignificant. Considering the discrete characteristics of RF, it is recommended to use SVM or FNN when the number of data point is not large and continuous estimations are expected. RF has an advantage particularly when the number of data points is very large and the data include categorial variables.

Key words: Global Surface Ocean CO_2, Machine Learning Models, Comparisons.

Introduction

Machine intelligence is starting to transform everything from daily life to scientific research[1] . In the broad spectrum of machine learning models, Random Forest (RF)[2] , Support Vector Machine (SVM)[3] , Feedforward Neural Network (FNN)[4] , and Self-Organization Map (SOM)[5] have been frequently used to upscale measurements from ground-based sites or underway monitoring to a regional or global scale. In the field of constructing surface ocean CO_2 concentration, a SOM was first applied to the Atlantic subpolar gyre[6] ; later a FNN was used to construct the monthly global CO_2 maps[7] ; and recently a RF and SVM came into the practice of CO_2 mapping [8] [9] . A SOM was also used to categorize the Atlantic surface water into different domains to improve the prediction of a FNN[10] .

Zeng et al.[11] compared the performances of a SVM, FNN and SOM with the track-gridded database of the Surface Ocean CO_2 Atlas (SOCAT) version 3.0. Since then, the number of data in the 1x1 degree mesh database has grown from 158,052 for the 1990-2014 period to 212,455 for the 1990-2017 period. Training SVM with half of the data points becomes difficult with an ordinary personal computer. In this study, we included a RF in the comparison and used a bootstrapping method to circumvent the data size problem. The method also reduces the possibility of overfitting.

Method

Models

A RF comprises uncorrelated random decision trees that are constructed by a top-down method to splits nodes until certain criteria are met. The terminal nodes of binary splitting provide prediction values for classification or regression. A RF model ensembles prediction values from multiple trees to gain better results. The first RF algorithm was created by Ho[12] and later latter extended by Breiman[13] . We used the ranger package[1] of Wright and Ziegler[2] in this study.

A SVM is a supervised learning model that was conceptualized in the in 1960s for classifications problems and later extended to regression analysis[14] . We implemented a least-square support vector machine for regression[15] that seeks to minimize the error between observations and the model outputs $\boldsymbol{y}$, which is expressed as a kernel function of $\boldsymbol{x}$:

$$\mathbf{y} = \boldsymbol{c}^T \boldsymbol{\Phi}(\boldsymbol{x}) + \boldsymbol{b}, \tag{1}$$

$$\Phi_{ij} = exp\left(-\frac{|x_i - x_j|^2}{-2\sigma}\right). \tag{2}$$

where σ is a constant that determined the nonlinearity of the kernel function, and the coefficients c and b are obtained by solving a set of linear equations

$$\begin{bmatrix} 0 & \boldsymbol{u}^T \\ \boldsymbol{u} & \boldsymbol{\Phi} + \gamma^{-1}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{c} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix}, \tag{3}$$

where $\boldsymbol{u}$ is a vector with all components being 1 and γ a constant. The performance of the SVM is largely determined by σ and γ whose optimal values depend on the input data of an application and are usually obtained by greedy search.

Our FNN comprises three layers[7] : an input layer, a hidden layer, and an output layer. The logistic kernel function is used by a neuron to transform its input vector $\boldsymbol{x}_a$ to an output y_a

$$y_a = \frac{1}{1+\exp(-(b_0 + \boldsymbol{w}^T \boldsymbol{x}_a))}. \tag{4}$$

A training updates the offset b_0 and weight $\boldsymbol{w}$ by minimizing the cost function

[1] github.com/imbs-hl/ranger

ISBN: 978-1-5386-5609-9

$$f(\boldsymbol{w}') = \frac{1}{2}\boldsymbol{e}^T\boldsymbol{e} = \frac{1}{2}\|\boldsymbol{y}_f - \boldsymbol{y}_m\|^2 , \quad (5)$$

where **w**' is the vector including b and $\boldsymbol{w}$, $\boldsymbol{y}_f$ the vector of model outputs, and $\boldsymbol{y}_m$ the vector of observations. We implemented the Levenberg–Marquardt algorithm[16] that updates parameters by

$$\boldsymbol{w}'(\mathrm{t}) = \boldsymbol{w}'(\mathrm{t}-1) - \alpha\boldsymbol{g} , \quad (6)$$

$$\boldsymbol{g} = (\mathbf{J}^T\mathbf{J} + \gamma\boldsymbol{I})^{-1}\mathbf{J}^T\boldsymbol{e} , \quad (7)$$

$$\nabla f(\boldsymbol{w}') = \mathbf{J}^T\boldsymbol{e} , \quad (8)$$

where α denotes the learning rate.

A SOM is a type of artificial neural network that is trained using unsupervised learning[17] . We implemented the batch learning algorithm[18] that makes the training results independent of the sequential order of training samples. In a training step, each sample is associated with a neuron cell to which the difference between the cell's parameters $\boldsymbol{p}$ and the sample data is smaller than to other neuron cells:

$$d = \|\boldsymbol{p} - \boldsymbol{x}\|. \quad (9)$$

An associated neuron cell for a sample is called the best matching cell (BMC). After the BMCs for all samples are found, the cell parameters are updated by

$$\boldsymbol{p}_i = \frac{\sum_k h_{ik}\boldsymbol{x}_k}{\sum_k h_{ik}}, \quad (10)$$

where i and k denote the indexes of neuron cells and training samples respectively. The neighborhood function that determines the weight factor h is defined as

$$h_{ik} = \exp(-\frac{r_{ik}}{q}), \quad (11)$$

Where r_{ik} is the distance between the ith neuron cell and the BMC of the kth training sample on a two-dimensional plane, and q is a factor that decreases linearly with the iteration loop of training.

Our implementations of SVM, FNN, and SOM can be downloaded from the China Southern Forest Ecosystem web site[2]. The programs scale sample data internally according to the discussions of [11] .

Data

The machine learning models assume the dependence of the monthly mean CO_2 fugacity (fCO2) in the surface seawater on latitude (LAT), sea surface temperature (SST), sea surface salinity (SSS), surface chlorophyll concentration (CHL), and mixed layer depth (MLD):

$$\mathrm{f}CO2 = f(LAT, SST, \mathrm{dSST}, SSS, CHL, MLD). \quad (12)$$

The linear trend of CO_2 was estimated by the method of [7] and the CO_2 data were normalized to the reference year of 2005. The derived monthly fluctuation of SST about the global annual mean SST was included to express the dependence of CO_2 on season.

The SOCAT version 6.0 product[3][19] were used in this study. A total of 212,455 CO_2 data points was extracted for the 1990-2017 period using the criteria set by [11] . The monthly means of SST were extracted from the Optimum Interpolation V2 product[4][20] , SSS from the World Ocean Atlas 2013 product[5][21] , CHL from the SNPP VIIRS climatology[6] of NASA[22] , and MLD from the Monthly Isopycnal and Mixed-layer Ocean Climatology[7][23] .

Bootstrapping

The SVM is notorious for having memory shortage problem with a large number of training samples[11] . Equation (3) indicates that the matrix dimension of the linear systems equals the sample size. If half of the samples were used for training, the matrix alone will take at least 90 GB memory and solving the equation could not be completed in a few days even by a high rank personal computer. Training FNN and SOM will also take a long time with such a large sample size.

We circumvented the problem by bootstrapping. By sampling randomly with replacement, ten sub-sample pairs were prepared for training and validation, resulting in 10 sub-models for each of the four machine learning models. In each sub-sample pair, the training used 10% of the whole data and the validation used the rest. The prediction for a given input was the average of the ten sub-models. Although a large sample size does not affect the training time of the RF as much as others and bootstrapping is part of its internal algorithm, we used the RF in the same way as using others for comparability.

Results

Overfitting could be a problem of machine learning. In principle, the RF and SOM make predictions by nearest neighborhood interpretation; therefore, they are less likely to overfit than the SVM and FNN. We tested the overfit of SVM and FNN using 100 randomly generated data points for 5 variables. With one variable as the response, both the SVM and FNN can fit the data perfectly well (Fig.1c and Fig.1d). Note that the fitting of FNN degrade with fewer hidden neurons (Fig.1a and Fig.1b). The capability of the SVM and FNN to establish a good relation for any uncorrelated variable seems to be a serious problem, but in practice a validation step is used to guard against such an overfitting. For example, the trained model above failed to make acceptable predictions for another independent random data set. Overfitting could also occur when the models are used to fill large gaps of independent variables. Discussion of this issue is beyond the scope of this study.

The statistics of training and validation are listed in Table 1. The correlation coefficient (R), bias, and the standard error (SE) between model outputs and observations were calculated

[2] united-csfe.com/fcew/ann.zip

[3] www.socat.info/

[4] www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html

[5] www.nodc.noaa.gov/OC5/woa13/

[6] oceancolor.gsfc.nasa.gov/cgi/l3

[7] www.pmel.noaa.gov/mimoc/

ISBN: 978-1-5386-5609-9

for each training and validation. The mean and standard deviation of R, bias, and SE of 10 trainings and validations were then summarized.

The statistics show that all the four models produced unbiased estimate for CO_2. The biases (model minus observations) are less than 1 μatm, which is small comparing with the variability of the gridded CO_2 data[7] . The SEs are in the magnitude of the variations of multiple observations in same grids and month[7] . The small standard deviations of among sub-models indicate that using 10% of the whole samples are sufficient for making reliable predictions for the rest 90%, which in turn indicates the reliability of constructing CO_2 distribution for largely unmeasured global oceans.

The correlation reveals that the RF performed the best, followed by the SVM, FNN, and SOM. The ratio of R of validation to training is a indicator for degradation of prediction and is expected to decrease with the ratio of the sample size of prediction to that of training. Our experiments yielded a R ratio of 0.76 for RF, 0.85 for SVM, 0.94 for FNN, and 0.57 for SOM. This indicates that the FNN's prediction is more consistent than others if only a small portion of the whole dataset is used for training.

Over all the differences of predicted CO_2 are small among RF, SVM, and FNN (Table 2). SOM sometimes showed much smaller or larger predicted CO_2 than others. Here we only presented the statistics in January and July 2005.

Fig. 2 shows modeled CO_2 distributions in the same months. In general, all the models captured the main distribution patterns in the observations very well, nevertheless the bias could be large regionally. For example, in the equator area off the Africa continent, the models, especially the SOM, produced much higher CO_2 concentrations than the observed values. However, the large discrepancy does not indicate bad model outputs as the number of measurements is small in the area and the CO_2 variation near the equator was large[7] .

Conclusion

Our investigation show that the RF is an excellent machine learning model when the data size is so large as to make using the SVM or FNN unfeasible. However, the discrete characteristics of the RF may not be desirable in some applications. Although the SVM performs better than the FNN, the later seems to be more flexible in controlling overfitting. In the sense of neural network, the SVM can be considered as a single layer neural network whose neuron cells equal the sample size or the number of supporting vectors; therefore, it does not have flexibility of choosing the number of neuron cells.

Although the machine learning models have the advantage of not requiring an explicit formula for the response variable, a user still needs to know their working mechanism to obtain a good result. For example, the models are not good at simulating trend when the signal is week and a iteration method should be used[7] , a circular variable must be transformed[7] or replaced by a proxy variable, and data should be scale properly if the software does not do it internally[11] .

Acknowledgements

The Surface Ocean CO_2 Atlas (SOCAT) is an international effort, endorsed by the International Ocean Carbon Coordination Project (IOCCP), the Surface Ocean Lower Atmosphere Study (SOLAS), and the Integrated Marine Biogeochemistry and Ecosystem Research program (IMBER), to deliver a uniformly quality-controlled surface ocean CO_2 database. The many researchers and funding agencies responsible for the collection of data and quality control are thanked for their contributions to SOCAT.

Table 1. Statistics of the correlation coefficient (R), bias (model minus observation), and standard error (SE) between model outputs and observations of 10 trainings. Each training used 10% of the randomly sampled data and validation used the rest.

	R^2	Bias (μatm)	SE (μatm)
	Training		
RF	0.948±0.001	-0.02±0.01	8.03±0.10
SVM	0.822±0.005	-4E-5±3E-5	13.88±0.20
FNN	0.707±0.005	0.02±0.00	17.77±0.21
SOM	0.662±0.005	-3E6±4E6	19.10±0.19
	Validation		
RF	0.723±0.002	0.11±0.09	17.20±0.05
SVM	0.698±0.002	0.18±0.11	17.97±0.07
FNN	0.662±0.002	0.23±0.12	19.01±0.06
SOM	0.378±0.004	-0.06±0.17	26.75±0.15

Table 2. Mean differences (μatm) among modeled monthly CO_2 distribution in 2005, which is the reference year for normalization. The number of data points is 31,892 in January and 30,076 in July.

	SVM	FNN	SOM
	January		
RF	0.04±9.21	-0.49±10.27	-2.73±17.12
SVM		0.54±8.52	-2.78±15.86
FNN			-2.24±15.64
	July		
RF	0.23±9.32	-0.01±10.69	-0.21±18.89
SVM		0.24±8.04	-0.44±17.38
FNN			-0.44±17.38

References

[1] Y. LeCun et al., Nature, 521, pp. 436-444, 2015.
[2] M. N. Wright and A. Ziegler, Journal of Statistical Software, 77, pp.1-17, 2017.
[3] C.-C. Chang and C.-J. Lin, Neural Computation, 14(8), pp. 1959-1977, 2002.
[4] B. M. Wilamowski and H. Yu, IEEE T. Neural Network, 21, pp. 930-937, 2010.
[5] T. Abe et al., Genom. Inform., 13, pp. 12-20, 2002.
[6] N. Lefèvre et al., Tellus B, 57, pp. 375-384, 2005.
[7] J. Zeng et al., J. Atmos. Ocean Tech., 31, pp. 1838-1849, 2014.
[8] E. Jang et al., Remote Sens. 9(821), pp. 1-23, 2017.
[9] L. Gregor et al., Biogeosciences, 14, pp. 5551-5569, 2017.
[10] P. Landschützer et al., Biogeosciences, 10, pp. 7793–7815, 2013.
[11] Zeng et al., Ocean Sci., 13, pp. 303-313, 2017.
[12] T. K. Ho, Proceedings of the 3rd International Conference on Document Analysis and Recognition, pp. 278-282, 1995.
[13] L. Breiman, Machine Learning, 45 (1), pp. 5-32, 2001.
[14] D. Basak, et al., Neu. Inf. Pro.-Letters and Reviews, 11, pp. 203-224, 2007.
[15] K. Pelckmans et al., presented at Neural Information Processing Systems, 2002.
[16] B. M. Wilamowski and H. Yu, IEEE T. Neural Network, 21, pp. 930–937, 2010.
[17] T. Kohonen, Springer, Berlin, 1984.

ISBN: 978-1-5386-5609-9

[18] T. Abe et al., Genom. Inform., 13, pp. 12-20, 2002.
[19] D. C. E. Bakker et al. Earth System Science Data, 8, pp. 383-413, 2016.
[20] R. W. Reynolds et al., J. Climate, 15, pp. 1609-1625, 2002.
[21] T. P. Boyer et al., NOAA Atlas NESDIS 72, pp. 1-209, 2013.
[22] J. E. Reilly et al., NASA Tech. Memo. 2000-206892, 11, 49 pp., 2000.
[23] S. Schmidtko et al., J. Geophys. Res., 118, pp. 1658-1672, 2013.

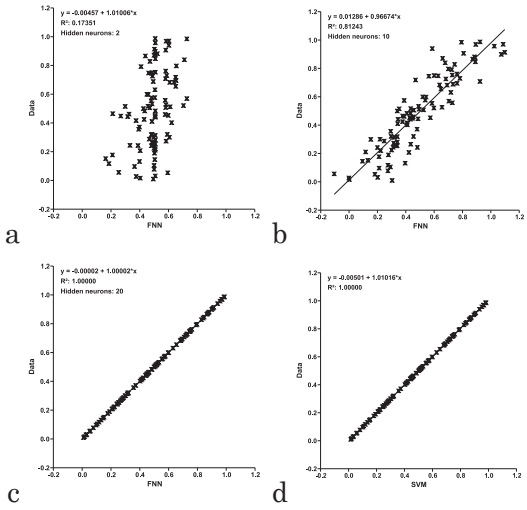

Figure 1. Overfit test for FNN with hidden neurons of 2 (a), 10 (b), and 20 (c) and for SVM (d).

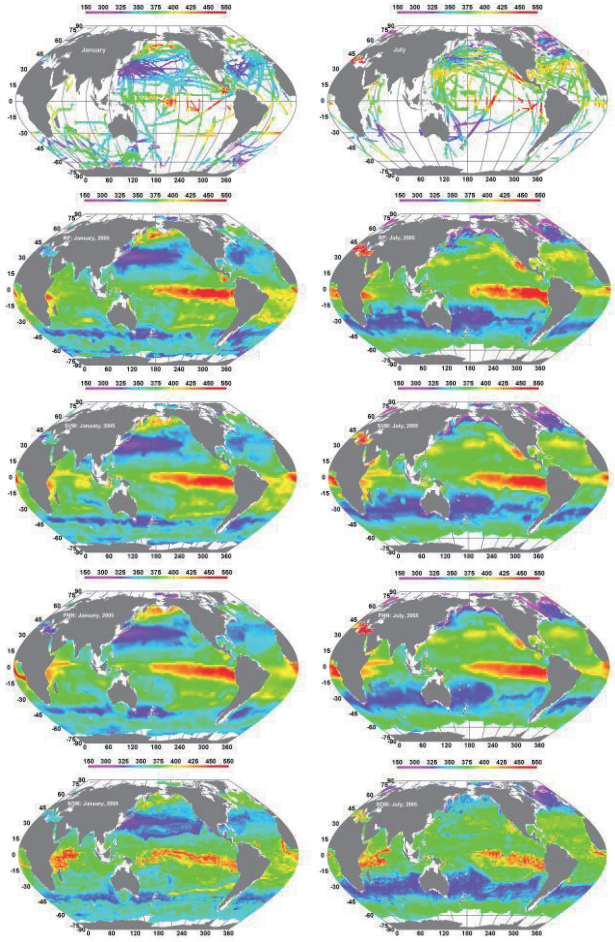

Figure 2. Modeled monthly CO_2 (μatm) distributions in January (left) and July (right), 2005. The figures on the top are the composite maps of normalized CO_2 in 1990-2017. Following them are model outputs of RF, SVM, FNN, and SOM respectively.

ISBN: 978-1-5386-5609-9