Retrieval of oceanic primary production using support vector machines

Shilin Tang, Chuqun Chen, Haigang Zhan

LED, South China Sea Institute of Oceanology, Chinese Academy of Sciences, sltang@scsio.ac.cn

ABSTRACT: One of the most important tasks of ocean color observations is to determine the distribution of phytoplankton primary production. A variety of bio-optical algorithms have been developed estimate primary production from these parameters. In this communication, we investigated the possibility of using a novel universal approximator-support vector machines (SVMs)-as the nonlinear transfer function between oceanic primary production and the information that can be directly retrieved from satellite data. The VGPM (Vertically Generalized Production Model) dataset was used to evaluate the proposed approach. The PPARR2 (Primary Production Algorithm Round Robin 2) dataset was used to further compare the precision between the VGPM model and the SVM model. Using this SVM model to calculate the global ocean primary production, the result is 45.5 PgC yr^{-1}, which is a little higher than the VGPM result.

KEY WORDS: Ocean primary production; Ocean color remote sensing; Support vector machine (SVM)

INTRODUCTION

Retrieval of ocean primary production by remote sensing has become a topical research issue in the oceanographic community. It is now possible to use satellite remote sensing to estimate primary production on global scales (Platt & Sathyendranath, 1988; Morel, 1991; Behrenfeld & Falkowski, 1997; Behrenfeld et al. 2002; Smyth, et al., 2005; Behrenfeld, et al., 2005). Among these models, the VGPM model developed by Behrenfeld and Falkowski (1997) being used the most popular, because all of the parameters in VGPM can be easily obtained from satellite data.

In the VGPM and most other models, primary production is regarded as a function of surface chlorophyll concentration, photosynthetically available radiation (PAR), euphotic depth, and maximum C fixation rate etc. Statistical models have also been developed in order to obtain phytoplankton primary productivity estimates from other variables that are easier to measure. However, the statistical regressions have limitations because of the nonlinear relationship between primary production and other variables. The uncertainty of the statistical models necessitates careful calibration.

Support vector machines (SVMs) may be able to solve these problems. SVMs do not refer to hardware; rather it is a software-based scheme. SVMs have been developed by Vapnik (Cortes and Vapnik, 1995; Vapnik, 1999; Vapnik, 2000) within the area of statistical learning theory and structural risk minimization (SRM). Today SVMs show better results than (or comparable outcomes to) neural networks (NNs) and other statistical models on important benchmark problems. SVMs are advantageous, because they were developed especially for sample limited datasets. The goal of the training is to obtain the optimal function other than the infinite sample result. So it can avoid the phenomenon that the learning process converges but the forecast performance is poor. SVM training leads to a convex quadratic programming (QP), obtaining the optimal result, and avoiding the problem of local extremum in traditional NNs. The SVM maps the original problem to a high feature space using nonlinear transformation. In addition, SVMs construct linear discriminants in high feature space instead of nonlinear discriminants in the original feature space, allowing for high generalization of the model. It solved the dimension disaster subtly, and let the complexity of the algorithm is independent of the dimension of samples. In this article, we have developed a primary production model using SVMs., and we used it to estimate global ocean primary production.

The initial work on SVM focused on optical character recognition (OCR). Within a short period of time, SV classifiers became competitive with the best available systems for both OCR and object recognition tasks. A comprehensive tutorial on SV classifiers has been published by Burges (1998). But also in regression and time series prediction applications, excellent performances were soon obtained (Stitson et al., 1999;

Mattera and Haykin). In this section, we will briefly introduce the basic ideas of SVMs for regression. For details see some books (Vapnik, 2000; Cristianini & Shawe-Taylor, 2000). The training software used in our experiments is LIBSVM (Chang and Lin, 2001; Fan et al., 2005).

RETRIEVAL EXPERIMENT

Data description and preprocessing

To carry out the retrieval of primary production using SVM, we first obtained an in situ dataset that archived by the VGPM project; these data were measured throughout the world's oceans from 80°N to 80°S, including both eutrophic regimes and oligoptrophic regimes from all major ocean basins. The dataset includes almost 3000 profiles of primary production measured by ^{14}C in situ and simulated in situ incubations. It also includes a series of other parameters related to primary production: measurement time, locations, C fixation rate, chlorophyll concentration (Chl), sea-surface temperature (SST), daily surface irradiance (Eo) and euphotic depth (Zeu), etc.

Behrenfeld and Falkowski set up a VGPM model based on this dataset in 1997:

$$PP_{eu} = 0.66125 \times P_{opt}^{B} \times [E_0/(E_0+4.1)] \times Z_{eu} \times Chl_{opt} \times D_{irr} \quad (7)$$

Behrenfeld and Falkowski (1997) examined the relationship between Chl_{opt} and Chl_{surf} (chlorophyll concentration at the surface), it shows a high correlation (r^2 =0.94) and Behrenfeld and Falkowski (1997) concluded that Chl_{opt} can be replaced by Chl_{surf} in the VGPM. In Behrenfeld and Falkowski (1997), a formula for P_{opt}^{b} as a function of SST was proposed:

All inputs and the output were $\log_{10}$-transfomed. The advantage of these transformations is that the distribution of transformed data will become more symmetrical and closer to normal. The input variables and output variable are the same as in VGPM. The dataset contains more than 3000 groups, but not all of them contain all the variables we would like to use. We choose 2412 groups that contain all the variables and in which primary production is not equal to zero. Half of them (1206) were picked as the training set and the other half as the validation set. We adjust the parameters to search the best result. Finally we set C=16, ε=0.08, σ=0.5, because these values perform the best. After we fixed these parameters, the SVM automatically determines the number and the locations of the RBF centers during its training.

Results

The performance of the models was evaluated using root-mean-square error (RMSE) and coefficient of determination (R^2) (Fig.1). The correlation between the retrieved and the in situ primary production for the training set is 0.9153 and the RMSE is 0.1491. The RMSE for the validation set is 0.1703, and its R^2 is 0.8878.

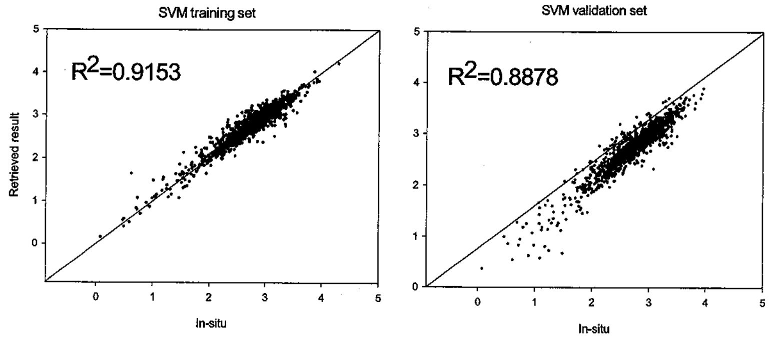

Fig.1. Comparison of the SVM-derived versus measured daily primary production on training and validation dataset. All data were $\log_{10}$-transformed.

The VGPM algorithm was applied to the same training and validation sets, and the comparison with observed values is shown on Figure 2. The correlation between the retrieved and the in situ primary production for the training set was 0.8605 and the RMSE for the training set was 0.2125. The RMSE for the validation set was 0.2178, and its R^2 was 0.8498.

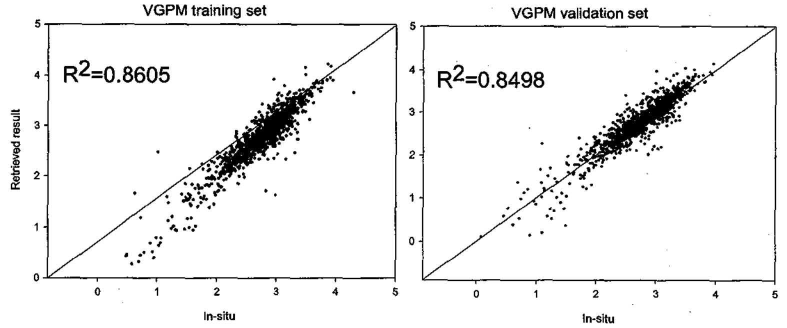

Fig.2. Comparison of measured primary production and VGPM modeled primary production. All data are $\log_{10}$-transformed.

Comparison of SVM model and VGPM model

To further compare the SVM model and the VGPM model, we chose another dataset from a project of PPARR2 to review our model. The PPARR2 data set has a data set of complete independence and large geographical distribution. It contains 89 groups of data. Compbell et al. (2002) used it to check up several primary production models. The SVM performed better result than the VGPM (Fig.3). The coefficient of determination is higher, 0.51 versus 0.47, and the RMSE decreases from 0.31 to 0.26. The result for this dataset has a lower degree of accuracy compared with the result in 3.2, because P_{opt}^{b} was modeled in that case through the 7th order polynomial function of SST used in the VGPM.

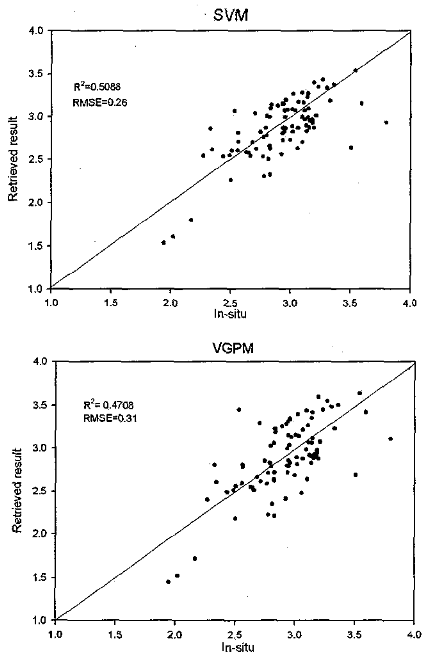

Fig.3. Comparison of SVM and VGPM model using PPARR2 dataset

GLOBAL PRIMARY PRODUCTION ESTIMATE USING SVM

We used our SVM model to estimate global primary production. The sea surface chlorophyll concentration, temperature and PAR data were downloaded from NASA's monthly standard products. Euphotic depths were calculated according to Morel and Berthon (1989). P_{opt}^{b} was modeled with a 7-order polynomial of SST developed by Beherenfeld and Falkowski (1997). D_{irr} was calculated using the program developed by the VGPM research team (http://marine.rutgers.edu/opp/). Thus the input data are the same as for the VGPM, i.e., P_{opt}^{b}, PAR, Z_{eu}, Chl_{surf} and D_{irr}. We used global monthly data covering 2003 to estimate primary production for every month and then integrated it to a year. For comparison, we also calculated the primary production using VGPM. The global primary production calculated using SVM is 4.55×10^{16} gC yr^{-1}, and the result of the VGPM is 4.3×10^{16} gC yr^{-1}.

CONCLUSION

In this paper, we introduced the SVMs for the first time in estimating ocean primary production. A great number of experiments and exhaustive comparison have been conducted. Analysis of VGPM and PPARR2 datasets shows that the performance of SVMs is more accurate than the VGPM model. The good results obtained by SVMs suggest that they constitute effective tools that could constitute effective tools that could lead to improvements in other biophysical parameter problems. It indicates that the SVMs are a feasible and universal method for modeling ocean primary production. Global ocean primary production was estimated using SVM on the basis of satellite remote sensing data.

LITERATURE CITED

Behrenfeld MJ, Falkowski PG(1997) Photosynthetic rates derived from satellite-based chlorophyll concentration. Limnol Oceanogr. 42: 1–20.

Behrenfeld MJ, Boss E, Siegel DA, Shea DM (2005) Carbon-based ocean productivity and phytoplankton physiology from space, Global Biogeochem. Cycles, 19, GB1006, doi:10.1029/2004GB002299.

Behrenfeld MJ, Maranon E, Siegel DA, et al. 2002. Photoacclimation and nutrient-based model of light-saturated photosynthesis for quantifying oceanic primary production. Mar Ecol Prog Ser 228: 103-117.

Burges, CJC (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2: 121 – 167.

Campbell J, Antoine D, Armstrong R, Arrigo K, Balch W, Barber R, Behrenfeld M, Bidigare R, et al. 2002. Comparison of algorithms for estimating ocean primary production from surface chlorophyll, temperature, and irradiance. Global Biogeochem Cycle. 16(3): 1035. GB1006[doi:10.1029/2004GB002299]

Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cortes C, Vapnik VN (1995) Support-vector networks, Machine Learning. Vol. 20, Issue 3: 273-297.

Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines And Other Kernel-Based Learning Methods. Cambridge University Press. PP: 136-138.

Mattera D Haykin S (1999) Support vector machines for dynamic reconstruction of a chaotic system. In Scholkopf B, Burges CJC, Smola AJ, Advances in Kernel Methods | Support Vector Learning. MIT Press. PP: 211-242.

Morel A, Berthon JF(1989). Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated in view of remote sensing applications. Limnol. Oceanogr. 34: 1545-1562.

Morel A (1991) Light and marine photosynthesis: a spectral model with geochemical and climatological implications. Prog. Oceanog. 26: 263-306.

Stitson M, Gammerman A, Vapnik V, Vovk V, Watkins C, Weston J (1999) Support vector regression with ANOVA decomposition kernels. In Scholkopf B, Burges CJC, and Smola AJ, Advances in Kernel Methods | Support Vector Learning. PP: 285-292. MIT Press.

Platt T, Sathyendranath S (1988) Oceanic Primary Production: Estimation by Remote Sensing at Local and Regional Scales. Science. 24: 1613-1620.

Smyth TJ, Tilstone JH, Groom SB (2005) Integration of radiative transfer into satellite models of ocean primary production. J. Geophys. Res., 110, C10014. [doi: 10.1029/2004JC002784]

Fan RE, Chen PH, Lin CJ (2005) Working set selection using the second order information for training SVM. J. Machine Learning Res. 6: 1889-1918.

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans. Neural Networks. 10: 988-1000.

Vapnik VN (2000) The Nature of Statistical Learning Theory, 2nd ed. New York: Springer-Verlag.

Acknowledgements

We thank the VGPM research group who shared the online primary production data and many software with us (http://marine.rutgers.edu/opp/); thanks to Chih-Chung Chang and Chih-Jen Lin for the SVM software package LIBSVM; we also thank the NASA Ocean Color Research Group for the standard ocean satellite data.