Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Automatic Red Tide Detection from MODIS Satellite Images

Weijian Cheng, Lawrence O. Hall, Dmitry B. Goldgof
Department of Computer Science and Engineering
University of South Florida
Tampa, USA
{wcheng2, hall, goldgof}@cse.usf.edu

Inia M. Soto, Chuanmin Hu
College of Marine Science
University of South Florida
St. Petersburg, USA
{isoto, hu}@marine.usf.edu

***Abstract*—Red tides pose a significant environmental and economic threat in the Gulf of Mexico. Timely detection of red tides is important for understanding this phenomenon. In this paper, learning approaches based on k-nearest neighbors, random forests and support vector machines have been evaluated for red tide detection from MODIS satellite images. Detection results from our algorithms were compared with ground truth red tide data collected in situ. Our results show that red tide identification methods based on machine learning approaches outperform baseline algorithms based on bio-optical characterization.**

***Keywords*—remote sensing, Florida's red tides, k-nearest neighbors, random forests, support vector machines**

I. Introduction

Toxic *Karenia brevis* blooms (commonly known as Florida's red tides) represent a serious problem for local fisheries and the tourism economy in Florida. Red tides occur frequently along the west Florida shelf, typically in late summer and fall.

With a series of polar orbiting ocean color satellite sensors, red tides can be potentially monitored and studied in near real-time every day over the entire eastern Gulf of Mexico, given cloud-free conditions. However, before satellite data can be utilized towards an automated system to provide rapid detection and early warning to the public, reliable algorithms must be developed to differentiate red tides from non-toxic blooms and other water disturbances in satellite imagery.

Pattern recognition has been used in the past to help interpret remote sensing imagery. The remotely sensed data have been classified using feed-forward neural networks [1], decision trees [2], expert systems [3] and rule-based systems [4] by a number of researchers. A computer expert system was developed to classify multi-spectral remote sensing imagery for red tide recognition [6]. Briefly, based on the spectral reflectance, an initial segmentation was performed using a fuzzy clustering algorithm (FCM) [7]. The algorithm assumes that the number of classes c is known, in addition, a partition distance metric, a fuzziness measure, and a stopping criterion are supplied. The FCM algorithm partitions the data set X into c classes including the red-tide class. However, this algorithm depends heavily on pre-defined parameters and might suffer from the problem of under-clustering (different water types classified as the same).

The Moderate Resolution Imaging Spectroradiometer instrument (MODIS) onboard the EOS Aqua satellite provides 1-km resolution ocean color products (which we refer to as bands) including chlorophyll-a (CHL), fluorescence line height (FLH), and spectral normalized water-leaving radiance (NLW). Recently, two thresholding approaches using these satellite bands have been proposed for red tide detection: the CHL anomaly threshold [12] and the backscattering threshold [18]. However, these methods can be problematic due to uncertainties in atmospheric correction, ocean bottom reflection, interference with other colored compounds in the ocean like colored dissolved organic matter (CDOM) and false alarms due to non-toxic phytoplankton blooms [10]. It is therefore desirable to combine these bio-optical approaches with the previously established pattern recognition frameworks for better detection.

In this paper, we evaluated three approaches based on k-nearest neighbors [16], random forests [5] and support vector machines [10] for automatic red tide identification using MODIS data and compared the results with the CHL anomaly and backscattering detection methods. In each approach, every image pixel covering 1 km^2 of seawater is classified as red tide or non red tide water using up to seven bands: CHL, FLH, particulate backscattering (BBP) at 551 nm and NLW at 412 nm, 551 nm, 678 nm and 869 nm.

II. Proposed Methods

Multi-spectral satellite data, namely the spectral NLW data, contain information about water constituents such as chlorophyll, suspended sediments, and colored dissolved organic matter. Using only one band of satellite data is prone to problems with artifacts caused by sediment and bottom reflection. Considering the heterogeneous nature of our problem, machine learning algorithms with good training noise tolerance can be used to automatically separate various types of environmental conditions and detect red tides. We propose machine learning approaches based on k-nearest neighbors (KNN), random forests (RF) and support vector machines (SVM) for red tide detection. We also propose hybrid

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

approaches combining machine learning models and thresholding approaches introduced in [12, 18].

A. Machine learning approaches

1) K-nearest neighbors

The k-nearest neighbors algorithm assigns an object to the class most common among its k nearest neighbors [16]. That is, for an input vector v, its distance d_i to every item in training data set t_i is computed as $d_i = \|v - t_i\|$. Let $d_{i_1}, d_{i_2}, ...d_{i_k}$ be the k shortest distances. v will be classified as the majority class among the class labels for $t_{i_1}, t_{i_2}, ...t_{i_k}$. We set k=3 in our experiments and used the Euclidean distance.

2) Random forests

A random forest (RF) [5] is an ensemble of decision trees. Each tree is constructed by randomly selecting N features at each internal node from the training set. Each node of the tree contains the test that creates the best split based on those N features. The classification result will be obtained from allowing all of the decision trees to vote.

We built 1000 trees for the random forests and set the split criteria to the C4.5 style. Random forests experiments were done using the OpenDT [9] system.

3) Support vector machines

Support vector machines (SVM) [10] first map the data into a higher dimension, then use a hyperplane in that feature space to separate the data into two classes. In the feature mapping stage, a kernel function is used to avoid explicit inner product calculation. We applied SVM as C-SVM [10] through the LIBSVM [11] system, using a radial basis function kernel.

B. Hybrid approaches

We were able to tune both unweighted and weighted voting to combine classification results produced by individual algorithms.

For unweighed voting with threshold N, a pixel will be classified as red tide if not less than N algorithms classify it as red tide, otherwise it will be classified as non red tide. We had 5 algorithms and varied N from 1 to 5 in our experiments.

For weighted voting, each algorithm produces its weighted voting value for each pixel between 0 and 1. For random forests, its weighted voting value is the percentage of trees that classify a pixel as red tide. For support vector machines, the weighted voting is a pixel's probability of being the red tide class. For 3-nearest neighbors, the weighted voting is the percentage of a pixel's red tide neighbors among its 3 nearest neighbors. For the CHL anomaly method [12] , it is the linearly normalized distance between the pixel's CHL anomaly and its thresholding value. For the backscattering method [18], it is the linearly normalized distance between this pixel's CHL, FLH and BBP and their thresholding values. A pixel will be classified as red tide if the sum of all 5 weighted voting values is 2.5 or more.

III. Experimental setup

A. Data set and preprocessing

For each day, if clouds don't cover the whole area of interest, one MODIS image for the West Florida Shelf was used. Ground truth red tide data for the West Florida Shelf was collected by the Florida Fish and Wildlife Research Institute (FWRI). Data points with water depth less than 2 meters were discarded, and counts from the different *Karenia* species were combined for the same dates and coordinates. From Jan 1, 2003 to Apr 20, 2007, 17649 ground truth data points were available. *K. brevis* cell counts higher than 15000 cells/liter were regarded as red tide, and non red tide otherwise. Although the cell count data were collected from a point source in the ocean, each ground truth data point was used to represent the 1 km^2 area of water corresponding to the satellite image pixel size.

Due to frequent cloud cover, only 1969 of 17649 data points were associated with valid, concurrent, and co-located MODIS data. To have more satellite data available for algorithm training, we developed the ground truth approximation strategy, shown in Figure 1.

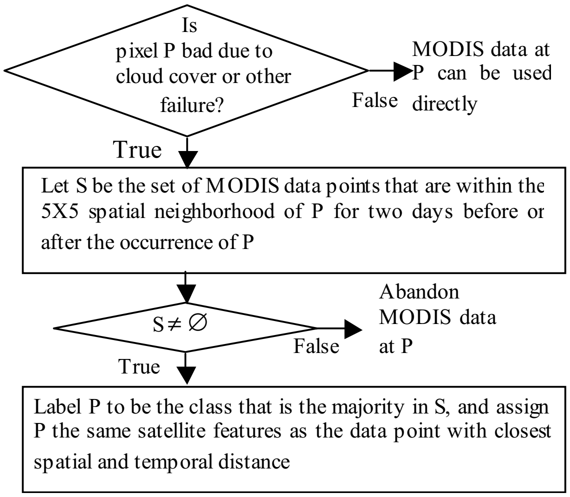

Is
pixel P bad due to
cloud cover or other
failure?
False
MODIS data at
P can be used
directly
True
Let S be the set of MODIS data points that are within the
5X5 spatial neighborhood of P for two days before or
after the occurrence of P
S ≠ ∅
False
Abandon
MODIS data
at P
True
Label P to be the class that is the majority in S, and assign
P the same satellite features as the data point with closest
spatial and temporal distance

Figure 1. Flow chart for ground truth approximation

After this process, we had 2695 ground truth red tide pixels and 8167 non red tide pixels across 832 days from Jan 1, 2003 to Apr 20, 2007.

About 75% of *in situ* cell counts data were labeled as "non red tide". Machine learning methods may easily produce models to classify all pixels as "non red tide" to obtain high accuracy on such skewed data [19]. To overcome this challenge, we randomly chose only B% of the majority class (non red tide) for training. Other *in situ* data labeled as "red tide" was 100% chosen. B was selected in the following way: we divided the training set into 5 chunks, using 4 chunks for training and 1 chunk for testing. Different B's from 10 to 100 were tested on each of 5 chunks. We used the B with the highest average F-measure on those 5 chunks for the whole training set.

For support vector machines, after the percentage B was selected, we selected the regularization constant C by doing a 5 fold cross validation on the training set. Different Cs from 0.5 to 4096 (increased by doubling itself on each new experiment) were tested. For each training set, we used the C with the highest average F-measure for the whole training set.

Two thirds of the ground truth data were randomly selected for training, and for the remaining one third, only points measured *in situ* without approximation were used for testing. 3-nearest neighbors, random forests and support vector machines used the same testing and training set. This process of randomly selecting the training and testing data was repeated 30 times. We present the averaged results of the 30 testing sets from each machine learning method.

Satellite data of chlorophyll (CHL), fluorescence light height (FLH), normalized water leaving radiance (NLW) and particulate backscatter (BBP) have been used previously to detect red tides [8, 12, 17, 18]. In our experiments with machine learning approaches we used CHL, FLH, NLW412, NLW551, NLW678, NLW869, and BBP551.

B. Data normalization

Each attribute was normalized to a value between 0 and 1 for each image by (1):

$$x_{ij} = \frac{\max_j(v_{ij}) - v_{ij}}{\max_j(v_{ij}) - \min_j(v_{ij})}. \qquad (1)$$

where v_{ij} is the original value of channel j for pixel i. x_{ij} is its value after normalization. $\max_j(v_{ij})$ and $\min_j(v_{ij})$ are the maximum and minimum values for channel j in the training set.

Some of the satellite data can be abnormally high due to satellite algorithm errors. To filter those extreme cases, we used all MODIS images from 2003 to 2007 and sorted the data in each channel j from high to low. Then, we set the maximum value of this channel ($\max_j(v_{ij})$) as the value ranking at the *k*th position ($k = round(number\ of\ all\ pixels \times 0.3\%)$) from the highest value. Any v_{ij} bigger than $\max_j(v_{ij})$ was normalized to 1.

C. Accuracy assessment

To understand how the algorithms work for both red tide and non red tide waters, we used a confusion matrix as shown in Table I, where A is true positive, D is true negative, B is false negative, and C is false positive.

TABLE I. NOTATION FOR ACCURACY ASSESSMENT

	Classified as red tide	Classified as non red tide
Red tide	A	B
Non red tide	C	D

To describe an algorithm's overall performance, considering correct recognition on both red tide and non red tide cases, we used the F-measure [21] as shown in (2).

$$\text{F-measure (FM)} = 2*A/(2*A+C+B) \qquad (2)$$

We also used the receiver operating characteristic (ROC) curve [14] to evaluate different methods in our experiments. Different true positive rates and false positive rates for each algorithm were generated by varying their respective thresholds. For random forests, we varied the threshold on the number of trees that predict the pixel as red tide. For support vector machines, we varied the threshold on the probability of the red tide class. For the CHL anomaly method, we varied its threshold. For the backscattering method, we varied its CHL, FLH and BBP thresholds.

Area under the ROC curve (AUC) [15] is proposed as a single-number measure for algorithm performance. AUC for random forests, support vector machines, the CHL anomaly method, and the backscattering method were computed for comparison in our study.

IV. RESULTS

A. Machine learning approaches

The three machine learning approaches had a higher F-measures than all thresholding methods, as shown in Table II. Under a two-sided Wilcoxon significance rank test [13] at a confidence interval of 95%, F-measures of all methods are significantly different except support vector machines and random forests. F-measures are ranked by (from high to low): support vector machines, random forests, 3-nearest neighbors, backscattering method, and CHL anomaly method.

TABLE II. F-MEASURES OF 3-NEAREST NEIGHBOR, RANDOM FORESTS, SUPPORT VECTOR MACHINES, THE CHL ANOMALY METHOD AND BACKSCATTERING METHOD.

Methods	F-measure
Support vector machines	0.590
Random forests	0.581
3-Nearest neighbor	0.562
Backscatter method	0.480
CHL Anomaly method	0.463

B. Hybrid approaches

F-measures of weighted voting and unweighted voting with N of 2 and 3 outperformed support vector machines (0.591), as shown in Table III. Unweighted voting with N=2 achieved the best F-measure of 0.607 among all voting strategies, higher than 0.597 from the weighted voting. The F-measure of voting method with N=2 was significantly higher than support vector machines under a two-sided Wilcoxon significance rank test at a confidence interval of 95%. Unweighted voting with N=5 (a pixel will be classified as red tide as long as it gets not less than N votes as red tide from all 5 algorithms) has the lowest F-measure of 0.271.

TABLE III. F-MEASURES OF HYBRID APPROACHES BY VOTING

Voting method	F-measure
Weighted voting	0.597
Unweighted Voting, N=1	0.579
Unweighted Voting, N=2	0.607
Unweighted Voting, N=3	0.605
Unweighted Voting, N=4	0.484
Unweighted Voting, N=5	0.271

C. *ROC analysis*

Figure 2 shows the ROC curves of support vector machines, random forests, CHL anomaly method, and the backscattering method. Different true positive rates and false positive rates were generated by varying their thresholds as discussed in Section III.C.

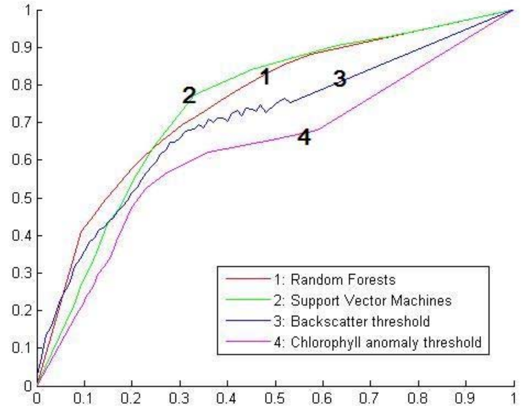


Figure 2. ROC curves for different methods

The AUC was computed for all 4 algorithms as shown in Table IV. The AUC for random forests was higher than other methods and had higher true positive than other methods when false positives were between 0.08 and 0.28. For an application that requires a false positive rate between 0.08 and 0.28, random forests can be a good algorithm to use.

TABLE IV. AUC FOR DIFFERENT METHODS

Method	AUC
Random forests	0.754
Support vector machines	0.747
Backscattering method	0.699
CHL anomaly method	0.629

D. *Confusion matrices*

Tables V, VI and VII show the confusion matrices for the support vector machines, backscattering method and weighted voting. The backscattering method did not detect as many red tide pixels as the machine learning approaches or weighted voting.

TABLE V. CONFUSION MATRIX FOR SUPPORT VECTOR MACHINES

	Classified as red tide	Classified as non red tide
Red tide	419	186
Non red tide	395	970

TABLE VI. CONFUSION MATRIX FOR BACKSCATTERING METHOD

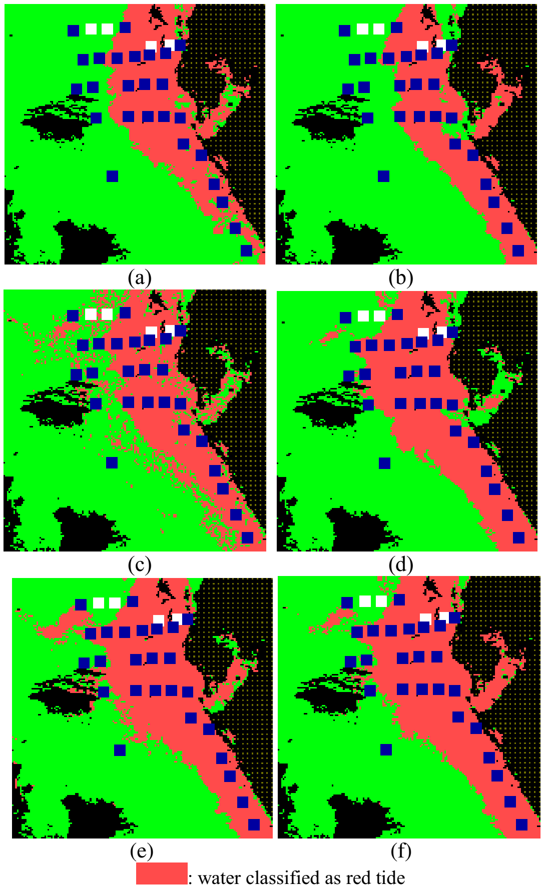
	Classified as red tide	Classified as non red tide
Red tide	242	363
Non red tide	160	1205

TABLE VII. CONFUSION MATRIX FOR UNWEIGHTED VOTING, N=2

	Classified as red tide	Classified as non red tide
Red tide	472	132
Non red tide	477	888

E. *An example*

Figure 3 shows an example of the results from these detection algorithms.



🟥: water classified as red tide

🟩: water classified as non red tide

Figure 3. Detection results on Oct 30, 2006. White square (□): ground truth non red tide points. Dark square (■): ground truth red tide points. (a): classification result by the CHL anomaly method; (b): classification result from the backscattering method; both CHL anomaly method and backscattering method missed some red tides in regions far from the shore; (c), (d), (e) and (f): classification result by 3-nearest neighbors, random forests, support vector machines and unweighted voting with N=2.

V. Conclusions

We proposed three machine learning approaches based respectively on 3-nearest neighbors, random forests, and support vector machines for red tide detection using MODIS imagery. Random forests and support vector machines achieved an F-measure 3% higher than 3-nearest neighbors and they outperformed the previous thresholding methods [12, 18] by more than 10% in terms of F measure, with statistical significance. The hybrid approach combining both machine learning and thresholding methods by voting improved the accuracy of machine learning approaches by less than 1.5% in terms of F-measure. Random forests and a support vector machine might be implemented as red tide detection in the eastern Gulf of Mexico. Future work includes combining visual interpretation from an image analyst to improve its accuracy.

Acknowledgement

This research was partially supported by the USF Research and Graduate Education Investment thrusts in Computational tools for discovery and Sustainable Healthy Communities-Water. We thank National Aeronautics and Space Administration (NASA) for the satellite data, and thank Florida Fish and Wildlife Research Institute for providing red tide ground truth data.

References

[1] R. K. Kiang. Classification of remotely sensed data using OCR-inspired neural network techniques, in IGARSS Symp., Houston, TX, 2:1081-1084, 1992.

[2] M. A. Friedl and C. E. Brodley. Decision tree classification of land cover from remotely sensed data, Remote Sensing of Environments, 61:399-409, 1997.

[3] F. A. Kruse, A. B. Lefkoff, and J. B. Dietz. Expert system-based mineral mapping in northern death valley, california/nevada, using the airborne visible/ infrared imaging spec-trometer(AVIRIS), Remote Sensing of Environments, 44:309-336, 1993.

[4] T. A. Warner, D. W. Levandowski, R. Bell, and H. Cetin. Rule-based geobotanical clas-sification of topographic, aeromagnetic, and remotely sensed vegetation community data,Remote Sensing of Environments, 50:41-51, 1994.

[5] L. Breiman. Random Forests. Machine Learning 45 (1), 5-32, 2001

[6] M. Zhang, L. O. Hall, and D. B. Goldgof. A Generic Knowledge-Guided Image Segmentation and Labeling System Using Fuzzy Clustering Algorithms, IEEE Transactions on Systems, Man, and Cybernetics, Part B, 32(5):571-582, 2002.

[7] J. C. Bezdek and S. K. Pal, Fuzzy models for pattern recognition, IEEE Press, Piscataway, NJ, 1992.

[8] C. Hu, F. E. Muller-Karger, C. Taylor, K. L. Carder, C. Kelble, E. Johns, and C. Heil. Red tide detection and tracing using MODIS fluorescence data: A regional example in SW Florida coastal waters. Remote Sens. Environ., 97:311-321, 2005.

[9] OpenDT's website: http://opendt.sourceforge.net

[10] V. N. Vapnik. The nature of statistical learning theory. Springer, 2000.

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] M. C. Tomlinson, R. P. Stumpf, V. Ransibrahmanakul, E. W. Truby, G. J. Kirkpatrick, B. A. Pederson, G. A. Vargo and C. A. Heil. Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico. Remote Sens Environ 91:293-303, 2004.

[13] F. Wilcoxon, Individual comparisons by ranking methods. Biometrics, 1, 80-83, 1945.

[14] F. Provost and T. Fawcett, Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997.

[15] S. J. Mason and N.E. Graham, Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q.J.R. Meteorol. Soc., 128, pp. 2145–2166, 2002.

[16] Belur V. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, 1991.

[17] Harmful Algal Bloom Forecasting System (HabFS) http://www.csc.noaa.gov/crs/habf.

[18] J.P. Cannizzaro, K.L. Carder, F.R. Chen, C.A. Heil, and G.A. Vargo. A novel technique for detection of the toxic dinoflagellate Karenia brevis in the Gulf of Mexico from remotely sensed ocean color data, Continental Shelf Research, 2008.

[19] I.H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques. (second edition). Morgan Kaufmann, San Francisco, CA. 2005.