APPLY SEMI-SUPERVISED SUPPORT VECTOR REGRESSION FOR REMOTE SENSING WATER QUALITY RETRIEVING

Xili Wang[1], Lei Ma[1], Xilin Wang[2]

1 School of computer science, Shaanxi Normal University, Xi'an 710062, P.R. of China
2 School of soil and water conservation, Beijing Forestry University, Beijing 100083, P.R. of China

ABSTRACT

This paper proposes a novel semi-supervised regression model with co-training algorithm based on support vector machines, which retrieves water quality variables from SPOT5 remote sensing data. Nonlinear relationship between water quality variables and SPOT5 spectrum are described by two support vector regression (SVR) models. Semi-supervised co-training algorithm for the two SVR models is established. The method is used for retrieving four representative water quality variables of the Weihe River in Shaanxi Province, China. The results show that the new method has better performance than the statistical regression method. Through integrating two SVR models and using unlabeled samples, an operational method when paired samples are limited is obtained. Combining techniques of machine learning and remote sensing, it provides an effective approach for remote sensing water quality retrieving.

Index Terms—semi-supervised learning, support vector regression, water quality variables, retrieving, SPOT5

1. INTRODUCTION

Studies on pollutants' spectral features and the improvement of retrieval algorithms have shown that it is possible to perform water quality monitoring through remote sensing on more water quality variables and with higher precision. Retrieving water quality from remote sensing data is time and cost efficient and feasible over a large area although it might not be as precise as traditional water quality monitoring methods.

Though bio-optical model is ideal for water quality remote sensing retrieving, it needs various optical properties about water quality parameters. Except few parameters, such as chlorophyll-a (chl-a), many parameters' optical properties are not clear. Therefore empirical models are well studied. We use such experiential model to depict the relationship between the spectrum and water quality variables. The traditional statistical regression methods are often used to establish parameter models to implement retrieving. Recently, artificial neural network (ANN) is used for water quality remote sensing. It is convenient for nonlinear modeling and has better performance than statistical regression. Both of the methods need lots of paired samples (inputs and corresponding outputs are all known) to construct reliable and accurate model. This is known as supervised learning in the field of machine learning. In most cases, there are not enough paired samples for modeling since abundant in situ measurements are too cost. We seek a new method--semi-supervised support vector regression (SVR) to deal with the problem of insufficient paired samples and model accuracy.

Based on the statistical learning theory, SVR can implement any nonlinear mapping without specify the form of the mapping function. It can attain the best generalization capability (namely, predict precision) using limited samples by tradeoff between the model complexity and learning ability.

In machine learning, supervised learning refers to the technique that construct (regression) model using paired samples, like statistical and ANN regression. Different from it, semi-supervised learning can get more accurate models using a number of unlabeled examples in addition to labeled samples. Semi-supervised learning not only use paired samples (also called labeled samples) but also exploit unlabeled samples (samples only inputs are known), and could get more accurate models [1]. Therefore it is helpful for remote sensing retrieving modeling since we have lots of unlabeled samples (i.e. remote sensing data) but limited paired samples.

Nonlinear support vector regression model is established for water quality variables retrieval, co-training algorithm is designed to take advantages of the semi-supervised learning. Using the proposed method and SPOT5 data, four water quality organic pollution indicators' (potassium permanganate index (CODmn), ammonia nitrogen (NH_3-N), chemical oxygen demand (COD) and dissolved oxygen (DO)) retrieving results for the Weihe River in Shaanxi Province, China and water quality image mappings for these variables are presented, and compared with the results of the multivariate statistical regression. The results show that the proposed method provides the

opportunity for the interpretation of water quality variations over large body of water even if field data are limited.

2. METHOD

2.1 Support vector regression model

Based on the statistical learning theory, SVR can implement any nonlinear mapping without specifying the form of the mapping function. It can attain the best generalization capability using limited samples by tradeoff between the model complexity and learning ability [2]. It is studied and used more and more because of its solid theoretical foundation and good performance.

Given sample data $(\boldsymbol{x}_i, y_i), i = 1, 2, \cdots, l$, where $\boldsymbol{x}_i$ denotes input vector, $y_i = f(\boldsymbol{x}_i)$ is the estimated output variable. The estimated function is $f(\boldsymbol{x}) = \boldsymbol{\omega}^T \phi(\boldsymbol{x}) + b$. Here $\phi(\boldsymbol{x})$ is some nonlinear mapping from the input space to a certain high dimensional space. $\boldsymbol{\omega}$ is the weight vector, b is the offset. The regression target is to find the parameters $\boldsymbol{\omega}$ and b which make the regression risk function $R_{reg}(f) = C\sum_{i=1}^{l} \Gamma(f(\boldsymbol{x}_i) - y_i) + \frac{1}{2}\|\boldsymbol{\omega}\|^2$ smallest, Constant $C > 0$ is a fixed penalty parameter. $\Gamma(\cdot)$ is a loss function. If ε-insensitive loss $L^{\varepsilon}(\boldsymbol{x}, y, f) = |y - f(\boldsymbol{x})|_{\varepsilon} = \max(0, |y - f(\boldsymbol{x})| - \varepsilon)$, which indicates that if the difference between the true value and the predicted value is less than ε the loss is 0, is used, solving the regression function can be expressed as a constrained optimization problem, and its dual optimization problem leads to a quadratic programming (QP) solution by Lagrange optimization method. Moreover, with the help of kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$ which satisfies Mercer's Conditions, the regression result can be expressed as [2]:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} (\bar{a}_i - \bar{a}_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + \bar{b} \quad (1)$$

where a_i, a_i^* are Lagrange multipliers. $\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \cdots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$, $\bar{b}$ is the optimal solution.

Radial basis kernel function $K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\{-\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 / \sigma^2\}$ is used in this paper. Thus, model parameters include penalty coefficient C, parameter of kernel function σ and width of the insensitive loss function ε. They are key factors affecting the performance of the SVR model. These parameters are often selected by trial and test, and the optimal value is difficult to obtain. Some new intelligent search techniques can find the global optimal solution in large search space. As one of such techniques, genetic algorithm (GA) has the advantages of fast and parallel search at complicated search space. Therefore we adopt GA to choose the optimal parameters of the SVR model and use RBF in this paper.

2.2 Semi-supervised SVR co-training algorithm

In order to implement semi-supervised learning, we borrow ideas from literatures [3] and [4] to design co-training algorithm for two SVR models.

Let $L = \{(\boldsymbol{x}_1, y_1) \ldots (\boldsymbol{x}_L, y_L)\}$ denotes the labeled example set. U denotes the unlabeled example set. $h1$, $h2$ denote the two SVR regressors and must be different, otherwise the algorithm will become self-training. The difference of the SVRs' parameters reflects the difference of $h1$ and $h2$. GA can be used to choose the parameters of the two regressors.

The retrieving model employs two support vector regressors. At first, Initial $h1$ and $h2$ are obtained by training set $L1$ and $L2$ respectively. $L1$ and $L2$ are selected from the origin labeled sample set L. Initial unlabeled sample set $U1$, $U2$ for $h1$ and $h2$ are selected from the unlabeled sample set U. Then enter the iterative learning process: firstly, $h1$ estimates the unlabeled data in $U1$, and puts the most confidently unlabeled data and its estimate result to the training set $L2$ of $h2$. Do the same for $h2$. Secondly, uses respective updated training set re-training the corresponding regressor. The process is repeated for a pre-set number of learning rounds. Finally, the regression result is acquired as the mean value of the two regressors' outcomes.

Estimating the labeling confidence is crucial for the algorithm. The labeling confidence is come from the influence of the labeling of unlabeled samples on the labeled samples. The sample that has the best labeling confidence should be the sample that makes the error of the regressor on the labeled sample set decreasing. Hence, the mean squared error (MSE) of the regressor on the labeled sample set can be evaluated and used for labeling confidence. If $\boldsymbol{x}_u$ is an unlabeled sample, $\hat{y}_u$ is the estimation result of $\boldsymbol{x}_u$ by regressor h. Add $(\boldsymbol{x}_u, \hat{y}_u)$ to the labeled sample set and re-training regressor h, denote the re-trained regressor as h' and the MSE of the re-trained regressor h' as MSE*. Let:

$$\begin{aligned} \Delta_u &= MSE - MSE^* \\ &= \sum_{x_i \in L} ((y_i - h(x_i))^2 - (y_i - h'(x_i))^2) \end{aligned} \quad (2)$$

Then, choose those $(\boldsymbol{x}_u, \hat{y}_u)$ that have the maximum Δ_u value to be the most confidently labeled sample.

Co-training algorithm uses two regressors to train simultaneously and improves the generalization ability of single regressor. In theory, it can improve the precision of a weak learner to any value. However, in practical use, the learning results may not be improved or even decreased after several rounds of co-training [5]. Hence, the algorithm only chooses a few examples from a number of unlabeled examples to add, which ensure the model parameters only have little change. In the experiment, we choose ten of the most confidently labeled samples from the forty unlabeled samples add to the training set.

The steps of co-training algorithm for the semi-supervised SVR are summarized as follows:

1. Produce initial training set $L1$, $L2$ for SVR $h1$ and $h2$ from the labeled example set L. Prepare initial unlabeled example set $U1$, $U2$ for $h1$ and $h2$ from the unlabeled example set U.
2. Use GA and labeled sample set $L1$, $L2$ to choose the parameters for the two SVRs respectively.
3. Regressor $h1$ estimates unlabeled example set $U1$, choose the most confidently labeled example and its estimation result join to the training set $L2$ of $h2$.
4. Regressor $h2$ estimates unlabeled example set $U2$, choose the most confidently labeled example and its estimation result join to the training set $L1$ of $h1$.
5. Update $L1$, $U1$, $L2$, $U2$, then retraining regressors $h1, h2$.
6. If the maximum number of iterations has not reached, go to step 3, else continue.
7. Use regressor $h1$, $h2$ to predict for new samples, the final regression result is the mean value of the two regressors' outputs.

3. EXPERIMENTAL RESULTS

In this paper, water quality of the Weihe River near city Xi'an (the capital of Shaanxi Province) is as a case study. The Weihe River is the largest branch of Yellow River. The total length is 818 kilometers, and drains a basin of 134,800 square kilometers. The drainage basin locates in the transition zone of the arid and humid region. The Weihe River basin in Shaanxi Province lies in central Shaanxi, China. It is the political, economic, cultural, financial and information center of Shaanxi Province. In recent years, with population growth and economic development, the Weihe River is seriously contaminated. Monitoring its water quality timely and large-scale and providing a basis for decision-making is significant.

Thirteen pairs of quasi-synchronous SPOT5 remote sensing data and in situ measurements from three monitoring stations on the Weihe River in Shaanxi constitute the labeled sample set. Shaanxi environmental quality bulletin show that the main pollution of the Weihe River is organic pollution in recent years [6]. Accordingly, we select four representative water quality variables CODmn, NH3-N, COD and DO as the retrieving parameters.

Preprocessing of remote sensing images include atmospheric correction and geometric correction. The objective radiance is obtained according to the correction formula:

$$RV = \frac{DC}{GAIN_\lambda} + BIAS_\lambda, DC = DN - L_d \quad (3)$$

DC is pixel gray value after calibration, L_d is the minimum value of each band. $GAIN_\lambda, BIAS_\lambda$: gain and offset of the remote sensor of the spectral band λ. All the four SPOT5 radiance bands are used for retrieving CODmn, NH_3-N, COD and DO.

The experiments use MSE and determination coefficient (R^2) to evaluate the results. They are defined as follows:

$$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ n$$
$$R^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \Big/ \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad (4)$$

y_i, $\hat{y}_i$ represent the i th sample's truth and retrieval value respectively. $\overline{y}$ is the mean value for y_i. MSE reflects the difference degree between true value and estimated value. R^2 reflects the fitting degree of the regression model.

Table 1 shows the thirteen samples' retrieving results by the multivariate linear regression (MLR) model and the semi-supervised SVR co-training regression (SS-SVR) model. The MLR model is obtained from the thirteen pairs of samples. In the SS-SVR model $|L1| = |L2| = 10$, the remaining three labeled samples are used as test samples in the SS-SVR establishing period, and they are different for the two regressors. $|U1| = |U2| = 40$.

Table 1: The results of the two retrieving models

	MLR (MSE/R^2)	SS-SVR (MSE/R^2)
CODmn	7.2283/0.8574	0.3172/0.9927
NH_3-N	4.5733/0.4833	0.0472/0.9831
COD	41.7274/0.7895	26.2338/0.9603
DO	0.9186/0.8132	0.0000/1.0000

From the table, the results of the SS-SVR are obviously better than that of the MLR. This indicates that the relationship between water quality parameters and remote sensing spectrum are typically nonlinear. The linear regression model can not represent such relations apparently. Benefit from the nonlinear mapping property of the SVR,

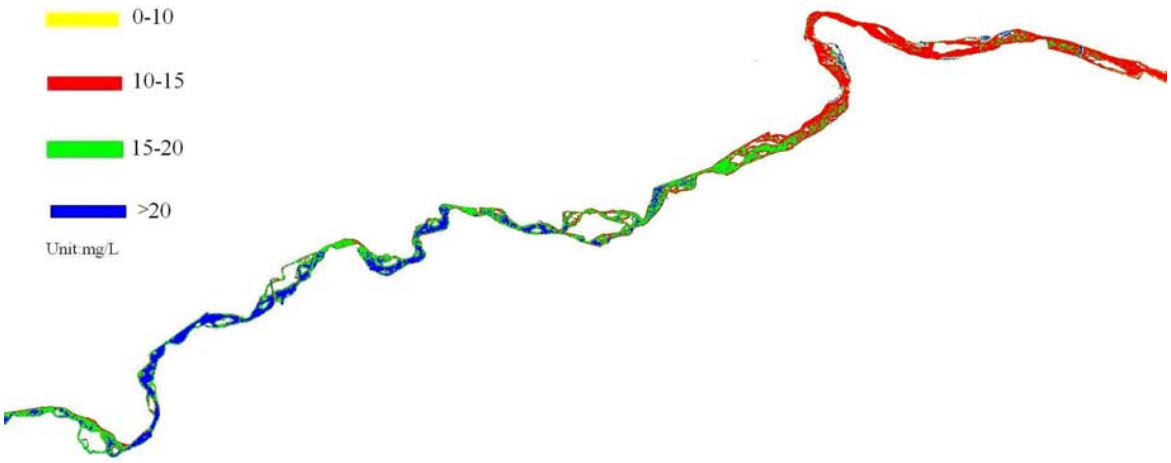

Figure 1. The concentration distribution map of CODmn

and further through a number of unlabeled samples and two support vector regressors co-training, the SS-SVR shows better performance.

We also use the SS-SVR model to retrieve the above four water quality variables for a part of the Weihe River. The remote sensing image was captured on 27th Jan, 2004. The results present the concentration distribution of the water quality variables of the Weihe River at that time, which reflects the pollution of the Weihe River directly. The concentration distribution maps use four different colors to represent different concentration ranges. Figure 1 shows the concentration distribution maps of CODmn. The maps show that this section is heavy polluted at that time. In situ measurements reflect the same situation. This district is near Xi'an and XianYang city. It is population densely, industry and agriculture developed, which water quality monitoring and management is more necessary. By Such operational method, we can give water quality map fast and large-scale through remote sensing data, and provide information for water quality managements. It is operational for the actual situation.

4. CONCLUSION

More attention should be paid to rapidly developing remote sensing techniques and studies on water quality data retrieval include for inland water and those water quality variables that are key to environmental management. We try to study retrieval method combining appropriate machine learning new technique, and propose a semi-supervised support vector regression model with co-training algorithm for remote sensing water quality retrieving. The model makes use of both labeled and unlabeled samples and two support vector regressors, improves the regression accuracy and has great advantages contrast to traditional regression methods when lack of paired samples. However, further studies are needed, such as: collect more data and does more space time analysis and validation; combine two different regressors and/or define labeling confidence by new suitable measurements to get better learning result.

5. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China (No.40671133), and the Fundamental Research Funds for the Central Universities (GK200902015).

6. REFERENCES

[1] O. Chapelle, B. Scholkopf, A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, Mass., USA, 2006.

[2] V. Vapnik, *The Nature of Statistical Learning*, Springer, New York, 1995.

[3] Z. H. Zhou, M. Li, *"Semi-supervised regression with co-training"*, In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Scotland, pp.908-913, 2005.

[4] A. Blum, T. Mitchell, *"Combining labeled and unlabeled data with co-training"*, In: Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98), Wisconsin, MI, pp.92-100, 1998.

[5] W. Wang, Z. H. Zhou, *"Analyzing co-training style algorithms"*, In: Proceedings of the 18th European Conference on Machine Learning (ECML'07), Warsaw, Poland, LNAI 4701, pp.454-465, 2007.

[6] http://www.snepb.gov.cn/.