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

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.

We use a model to depict the relationship between spectrum and water quality variable. The traditional statistical regression methods are often used to establish parameter model 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 new method--semi-supervised support vector regression (SVR) to deal with the problem of insufficient paired samples and model accuracy. 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 advantage of 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 Weihe River in Shaanxi province, China are presented, and compared with results of multivariate statistical regression.

2. METHOD

2.1 Support vector regression model

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 [2].

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 estimated output variable. The estimated function $f(\boldsymbol{x}) = \boldsymbol{\omega}^T \phi(\boldsymbol{x}) + b$, where $\phi(\boldsymbol{x})$ is nonlinear mapping from the input space to a certain high dimensional space. $\boldsymbol{\omega}$ is weight vector, b is 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. where $\Gamma(\cdot)$ is loss function. Constant $C > 0$ is a fixed penalty parameter. When use ε-insensitive loss function $L^{\varepsilon}(\boldsymbol{x}, y, f) = |y - f(\boldsymbol{x})|_{\varepsilon} = \max(0, |y - f(\boldsymbol{x})| - \varepsilon)$, 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 $K(\boldsymbol{x}_i, \boldsymbol{x})$, the regression result can be expressed as [2]:

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

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

We use radial basis kernel function $K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp\{-\|\boldsymbol{x} - \boldsymbol{x}_i\|^2 / \sigma^2\}$ in this paper. 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 SVR model. We adopt genetic algorithm to choose the optimal parameters of SVR model.

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 SVR model. The model employs two support vector regressors. Initial SVRs $h1$ and $h2$ are obtained by training set $L1$, $L2$ selected from the labeled sample set L, and prepare initial unlabeled sample set $U1$, $U2$ for $h1$ and $h2$ from unlabeled sample set U. Then enter the learning process: $h1$ estimates unlabeled data in $U1$, puts the most confidently unlabeled data and its estimate result to the training set $L2$ of $h2$. Do the same work

for $h2$. Next 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 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, denote MSE of the new regressor h' as MSE*. Let:

$$\Delta_u = MSE - MSE^* = \sum_{x_i \in L} ((y_i - h(x_i))^2 - (y_i - h'(x_i))^2) \tag{2}$$

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

Theoretical analysis shows that: after a number of learning rounds, the co-training process could not improve the performance further since the difference between the two regressors become smaller and smaller; If the two initial regressors have small difference, $|U|/|L|$ ($|set|$ represents number of *set*)should be small [5]. Hence, the number of iteration and unlabeled data should be appropriate.

3. EXPERIMENTAL RESULTS

In this paper, water quality of Weihe River (the largest branch of Yellow River) near city Xi'an (the capital of Shaanxi province) is as a case study. Thirteen pairs of quasi-synchronous SPOT5 remote sensing data and in situ measurements constitute the labeled sample set.

Preprocessing of remote sensing images include atmospheric correction and geometric correction. The objective radiance is obtained and all the four SPOT5 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. Table 2 shows the results of multivariate linear regression (MLR) and semi-supervised SVR co-training regression (SS-SVR).

From the table, the results of SS-SVR are obviously better than that of MLR. This is due to the superiority of nonlinear mapping by SVR, and SS-SVR may do better through a number of unlabeled samples and two support vector regressors co-training, it is operational for the actual situation.

Table 1: The results of 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

4. CONCLUSION

More attention should be paid to rapidly developing remote sensing techniques and theoretical studies on water quality data retrieval include for inland water bodies 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).

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.