Application of Data Mining Technology Based on Wireless Sensor Networks in Oceanographic Forecasting

https://doi.org/10.3991/ijoe.v14i04.8392

Wei Zhai
Xi'an Aeronautical University, Xi'an, China
2587842805@qq.com

Abstract—This paper aims to present a desirable prediction method for oceanographic trends. Therefore, an online monitoring scheme was prepared to collect the accurate oceanographic hydrological data based on wireless sensor network (WSN) and computer technology. Then, the data collected by the WSN were processed by support vector regression algorithm. To obtain the most important parameters of the algorithm, the particle swarm optimization was introduced to search for the global optimal solution through the coopetition between the particles. After that, an oceanographic hydrological data collection and observation system was created based on the hydrological situation of New York harbour. Then, the traditional support vector regression and the proposed method were applied to predict the oceanographic trends based on water temperature, salinity and other indices. The results show that the proposed algorithm enhanced the data utilization rate of the WSN, and achieved good prediction accuracy. The research provides important insights into the application of advanced technology in oceanographic forecast.

Keywords—Oceanographic forecast, Wireless sensor network (WSN), Data mining, Support vector regression.

1 Introduction

Since its nascence in the late 1990s, the wireless sensor network (WSN) has developed into a collaborative network capable of collecting and processing object information and transmitting the processed information to users. The WSN usually consists of several sensors in ad hoc mode. Over the years, the network has been applied to various fields, such as environmental monitoring, traffic management, risk monitoring, wildlife tracking, and security monitoring [1-2].

In particular, the proliferation of the WSN has reshaped the monitoring of water resources. With the WSN, it is possible to flexibly adjust the monitoring tasks, and realize real-time monitoring of water resources. The WSN is especially suitable for monitoring the water quality in large water bodies, thanks to the following advantages. First, the network nodes can automatically form a network in the monitoring area; second, the WSN has a wide scope of monitoring; third, the network leaves a small environmental impact [3-4]. Many real-time water resource monitoring systems have been developed based on the WSN in foreign countries, namely, the EmNet of Heliosware (US) and the Fleck system of CSIRO (Australia) [5-6]. In China, scholars have also explored the WSN-based real-time monitoring systems for water resources [7].

Through the above analysis, it is not surprising that the WSN has also been introduced to the monitoring of oceanographic hydrological data. These data have long been the focal point in theoretical and empirical research. For instance, Wooley et al. employed data mining, classification and clustering techniques to process oceanographic hydrological data [8]. Guo et al. extracted and visualized the features of oceanographic hydrological data, and relied on them to detect the flow and vortex in the ocean [9]. Hun proposed a method to capture association results in remote sensing images, and created a phylogenetic tree structure of oceanographic hydrological data. To predict typhoons, Liu et al. analysed the typhoon images with various data mining techniques (e.g. principal component analysis, case-based learning, and similarity-based image retrieval) [11].

To sum up, many advanced data processing methods are yet to be applied to the management and utilization of oceanographic hydrological data. Therefore, this paper prepares an online monitoring scheme that can effectively collect oceanographic hydrological data based on the WSN and computer technology. Then, the data collected by the WSN were processed by support vector regression algorithm. To obtain the most important parameters of the algorithm, the particle swarm optimization was introduced to search for the global optimal solution through the coopetition between the particles. After that, an oceanographic hydrological data collection and observation system was created based on the hydrological situation of New York harbour. Then, the traditional support vector regression and the proposed method were applied to predict the oceanographic trends based on water temperature, salinity and other indices. The results show that the proposed algorithm enhanced the data utilization rate of the WSN, and achieved good prediction accuracy.

2 Wireless sensor networks for marine hydrology

2.1 Basic concept

The hardware structure of the seawater detection system is composed of sensor nodes, gateway nodes, and a central computer system. The wireless sensor nodes are distributed in the sea area which needs to be detected, and performs the tasks of data collection, simple processing, and communication with the sink node. The collected data are transmitted progressively along the sensor node, and then collected at the sink node after multi-hop. Then, it is transmitted to the control center through the general packet radio service (GPRS) network, so as to become first-hand information for expert studies. At the same time, administrators can manage the sensor network nodes through the GPRS network to achieve two-way communication. In this paper, we briefly introduce the sensor nodes and sink nodes in the WSNs.

The sensor node is at the forefront of data acquisition. It is responsible for the collection of environmental parameters, and then transmits the information to the user. Its structure is shown in Figure 1.

The sensor unit collects the characteristic signal of the object which is to be detected. In general, each sensor unit contains a variety of sensors, and they work simultaneously. The microcontroller is the control core of the sensor node, which is responsible for the control of data collection and transmission, and the communication with the sink node. The sink node is the control center of the system, which is mainly composed of a wireless communication module, microcontroller, power supply module, keyboard and liquid-crystal display (LCD). The structure of the system is shown in Figure 2.

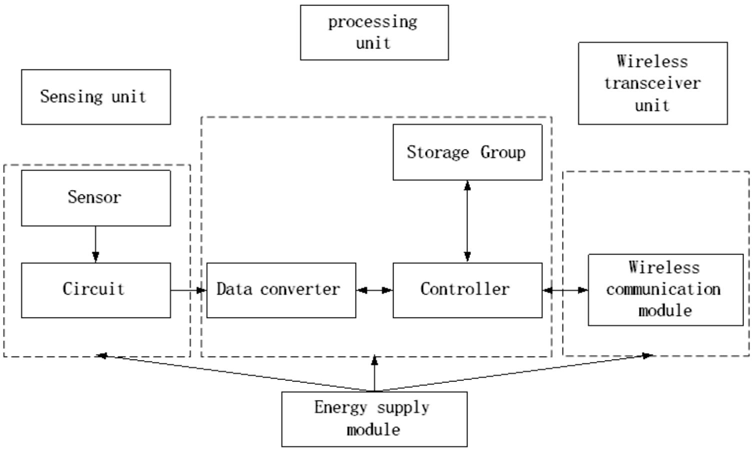

Fig. 1. Architecture of sensor node system

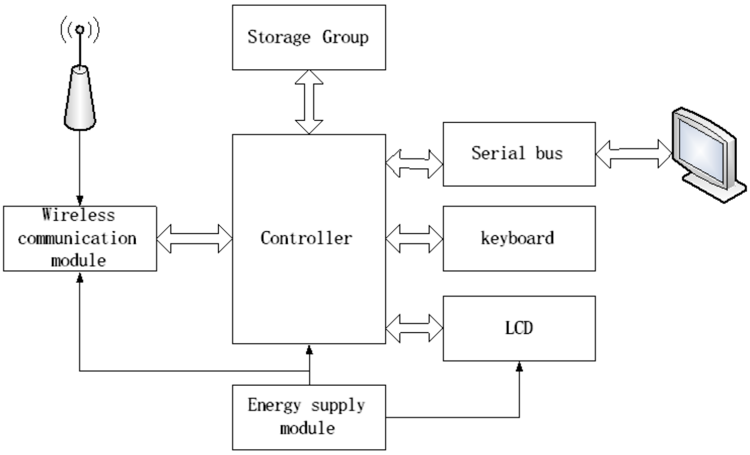

Fig. 2. Diagram of sink node

The sink node controls the data acquisition of wireless sensor nodes through the wireless communication module. Then, the collected data are transmitted to the host computer through the serial data bus, and a friendly monitoring interface is provided on the host computer. It forms the core part of the marine hydrological monitoring and analysis system.

2.2 Real-time monitoring and analysis system for marine hydrology

The ultimate purpose of ocean observation is to understand the current situation of marine law, and even predict the future of the marine situation. This requires more precise data. A data quality control platform has been generated for this demand, and it plays a very important role in the real-time data management system of marine hydrological observation. The improvement of data quality is the basis of other work. If the data quality is not high, even if there are better models and methods, our predictions are inaccurate. The data quality control platform is scalable and can be implemented by adding new data quality control methods. Figure 3 is a real-time data management system for marine hydrology.

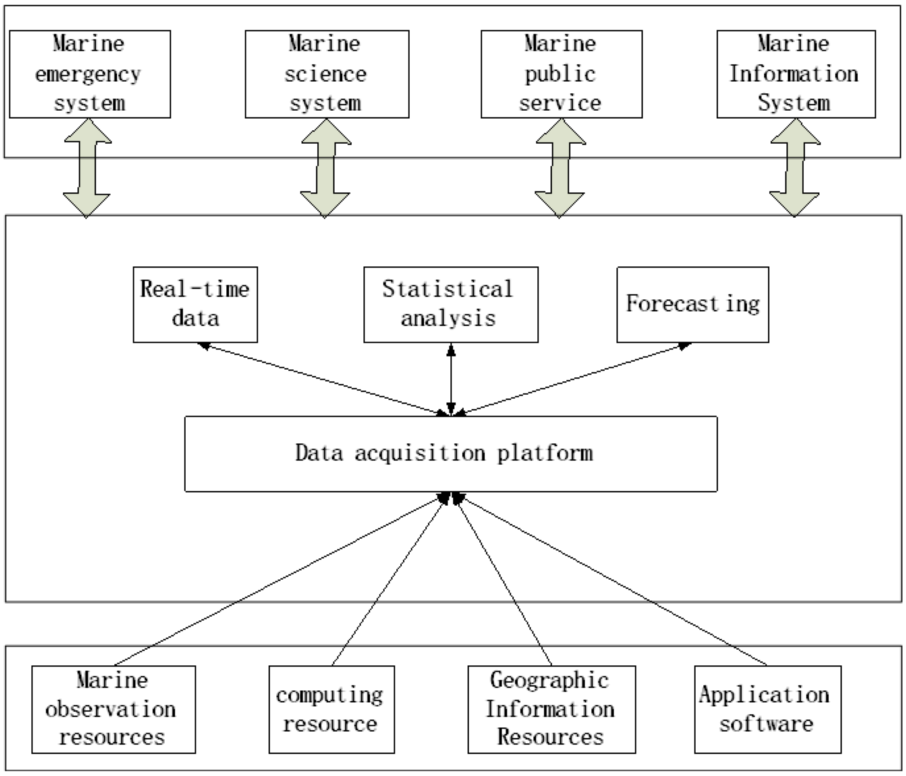

Fig. 3. Real-time data management system for marine hydrology

3 Particle swarm optimization SVM algorithm

We start out by introducing the principle and improvements of particle swarm optimization (PSO). There is a sample set: $T=(x_i, y_i)$, 1, 2, L, n. Among them, x_i is the *i*th input value in the *n*-dimensional sample set, and y_i is the *i*th output value. Based on the support vector machine SVM learning process, it can be interpreted as the solution *f(•)* between the input and output. For any $T=(x_i, y_i)$, 1, 2, L, n, it enables $f(x_i)=y_i$ to be set up. Generally, *f(•)* can be expressed as:

$$f(x)=\omega_i\phi_i(x)+b \tag{1}$$

In formula (1), ω is the weight, and b is called the threshold. The function of $\varphi: x^n \to \Omega$ is to map the low dimensional input vector to the high dimensional feature space. In order to improve the prediction accuracy of the water level near the port, to obtain the appropriate parameters, and to minimize the structural risk, we can construct the empirical risk function as shown in formula (2):

$$R=\frac{1}{L}\sum_{i=1}^{m} \tag{2}$$

Among them, L is called the loss function, and its expression is:

$$L(\bullet)=\begin{cases}0 & |y_i-f(x_i)|<\varepsilon \\ |y_i-f(x_i)|<\varepsilon & |y_i-f(x_i)|\geq\varepsilon\end{cases} \tag{3}$$

Where ε is greater than 0. When learning samples are small, it is not appropriate to use empirical values to replace the actual risk value, so the support vector machine algorithm considers the generalization risk which is called the SRM criterion. Then, according to the SRM criteria, the selected nonlinear regression function is:

$$f(x)=\operatorname{sgn}(\sum_{i=1}^{m}\omega_i\phi_i(x)+b) \tag{4}$$

According to the principle of maximum interval of the SVM, the nonlinear regression problem can be transformed into the following functional programming problem:

$$\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}(\xi_i+\xi_i^*) \tag{5}$$

And it satisfies the constraint of formula (6):

$$\text{s.t.}\begin{cases} y_i - f(x_i) \le \varepsilon + \xi_i \\ f(x_i) - y_i > \varepsilon + \xi_i^* \\ \xi_i \ge 0 \\ \xi_i^* \ge 0 \end{cases} \tag{6}$$

In the above formula, the parameters ξ_i and ξ_i^* are relaxation variables, and C is generally called the capacity factor. In general, formula (6) is transformed into the Lagrange dual problem to be solved:

$$f(x) = \sum (\alpha_i^* - \alpha_i) K(x_i, x) + b \tag{7}$$

Among them, a_i is a Lagrange multiplier, and K(•) is called a kernel function which satisfies the following formula:

$$K(x_i, x_j) = \varphi(x_i) \bullet \varphi(x_j) \tag{8}$$

In this paper, the Gaussian kernel function is taken as the kernel function of SVM. That is:

$$K(x_i, x) = \exp(\frac{\|x_i - x\|}{2\sigma^2}) \tag{9}$$

In this paper, an improved PSO algorithm is adopted in the training process. The PSO algorithm is used to search for the global optimal solution through the cooperation and competition among the particles. Each particle represents a possible solution vector. For SVM training, it is possible to make the current position of the particle the current value of the parameter vector:

$$x = (C, \varepsilon, \sigma) \tag{10}$$

Then, the optimal position obtained by the optimization process is the optimal solution of the elements in the vector, that is, the optimal parameters of the corresponding SVM. The basic principle of the improved PSO can be expressed as:

$$\begin{cases} v_i^{k+1} = \omega v_i^k + c_1 r_1 (x_{pbest} - x_{id}^k) + c_2 r_2 (x_{gbest} - x_i^k) \\ x_i^{k+1} = x_i^k + v_i^{k+1} \end{cases} \tag{11}$$

In formula (11), v_i^{k+1} v_i^{k+1} is the current particle speed, v_i^{k+1} is the current position of the particle, r is a random constant, c is the learning factor, and ω is the inertia weight. The function is as follows:

$$\omega = \omega_{\max} - k \frac{\omega_{\max} - \omega_{\min}}{k_{\max}} \tag{12}$$

In formula (12), ω_{max} is considered to be the initial weight, ω_{min} is considered to be the final weight, k is expressed as the current number of iterations, and k_{max} is the maximum number of iterations.

Next, we will introduce the methods and steps of using PSO to optimize the penalty parameter and kernel function parameter in the SVM algorithm.

4 Simulation experiment and result analysis

4.1 Preparation of data and description of correlation factors

Firstly, a set of WSN nodes are set up in the New York port and the upper reaches of the river and estuary, so as to construct an observation and prediction system for the harbor's oceanographic hydrological environment. The system combines the fixed point data collection, data fitting, and the prediction model of this paper to achieve the purpose of using hydrological indicators to predict the water level near the port.

Secondly, we give the water level diagram of New York Harbor near Red Hook at 0:00 local time on October 1, 2016 in Figure 4. As can be seen from Figure 4, the water level outside the port of the city is relatively low, while the water level inside the port is relatively high. The real-time control of the water level is a very meaningful thing, as it is conducive to enhancing the safety level near the port.

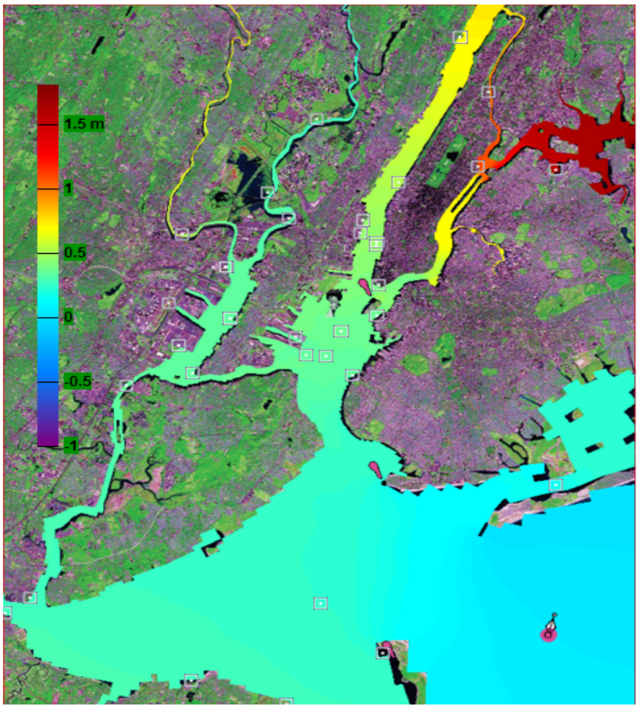

Fig. 4. Water level of New York Harbor near Red Hook

Thirdly, we illustrate the hydrological indicators of the marine environment. The real-time hydrological element data of the marine environment observation mainly include the temperature, salinity, and density of seawater, waves, currents and sea ice, water color, transparency, etc.

1. The water temperature is one of the basic physical factors of the ocean, and many marine phenomena and even the Earth's phenomena are related to the temperature of the seawater. The sea surface temperature depends on solar radiation, therefore, the seawater temperature is high in low latitudes, and low in high latitudes.
2. The seawater salinity is another important physical factor in the ocean. In addition to the salinity of various seas not being the same, the vertical distribution of the salinity is also different. The surface salinity is low in cold sea areas, and the salinity of seawater increases with depth.
3. All factors that affect the temperature and salinity of seawater will affect the density of seawater. The distribution of density of seawater is complex, and it varies with geographical location, the depth of the sea, and the time. There is a certain relationship between the density of seawater and temperature, and between the density of seawater and salinity. It follows the law that the higher the salinity, the higher the density, and the higher the temperature, the smaller the density.

To sum up, the salinity, atmospheric temperature, and sea surface temperature are chosen as sample data sets, and then they are put into the improved SVM prediction model which is established in this paper.

Finally, according to the WSN which is established in this paper, we can collect ocean hydrological data of the city port from October 1 to October 31, 2016. The data set is collected every four minutes by the WSN, and it contains a total of 2,400 data. Then, we use the data as a sample data set, and the specific situation is shown in Figures 5, 6, 7.

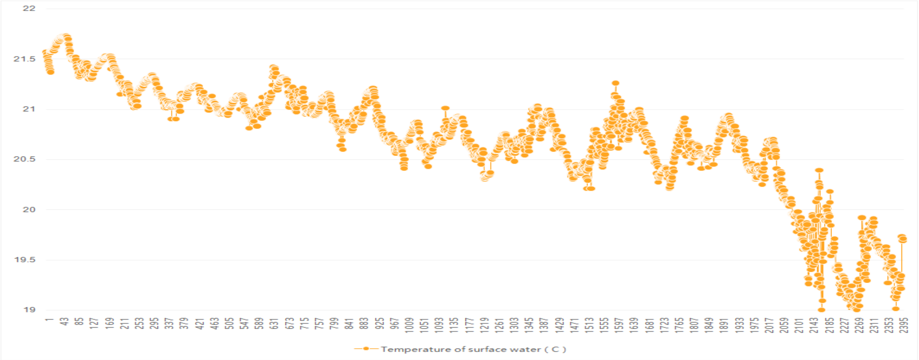

Fig. 5. Temperature of surface water in New York Harbor near Red Hook

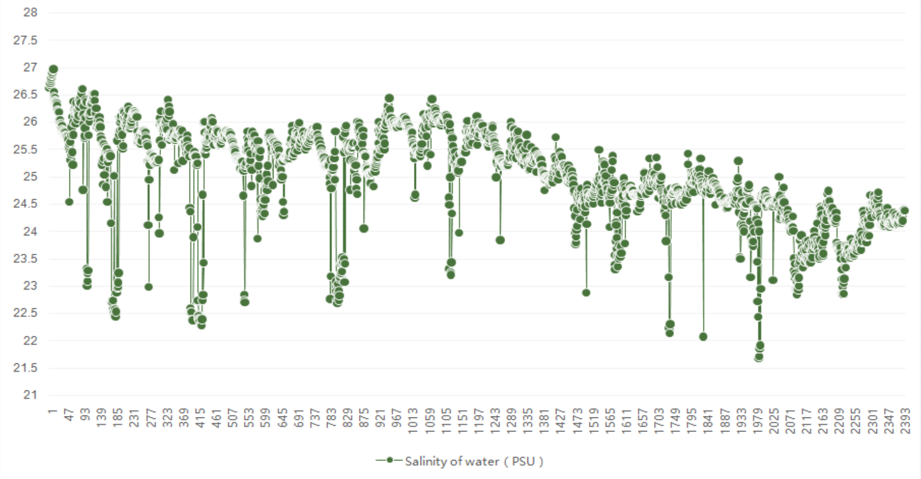

Fig. 6. Salinity of water in New York Harbor near Red Hook

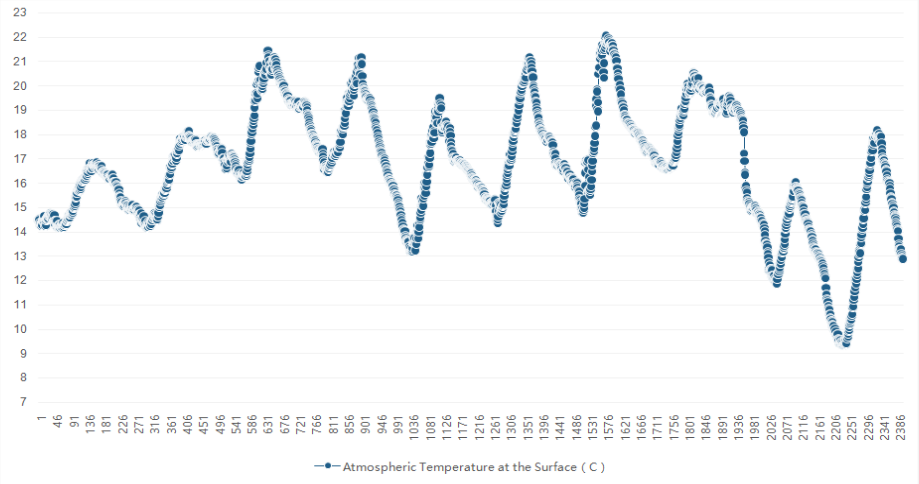

Fig. 7. Atmospheric surface temperature in New York Harbor near Red Hook

4.2 Experiment steps and result analysis

This paper uses the data of New York port from October 1 to October 10, 2016 as the sample data set. Then we use the SVM algorithm and the improved PSO optimization SVM algorithm proposed in this paper to predict the water level near the port in the next one day. Next, we will introduce the steps of the experiment, the experimental results, and the comparison of the experimental results.

The improved PSO optimization SVM algorithm is used for classification and prediction.

Parameter preparation:

The initial learning factor c1: The initial value is 2, which is used to control the local search ability of the PSO parameter;

The initial learning factor c2: The initial value is 2, which is used to control the global search ability of the PSO parameter;

The maxgen: The initial value is 100, which is used to control the maximum number of evolutions;

The sizepop: The inietial value is 20, which is used to control the maximum number of population;

K: The initial value is 0.6, which is used to control the relationship between speed and x;

Inertia weight w: The initial value is 1, ω_{min} is 0.8, and the ω_{max} is 1.2, which is the elastic coefficient in the velocity update formula;

The penalty parameter Cmax: The initial value is 100, which is used to control the maximum value of the SVM parameter C;

The penalty parameter Cmin: The initial value is 0.1, which is used to control the minimum value of the SVM parameter C;

The Gamma parameter gmax: The initial value is 100, which is used to control the maximum value of the SVM parameter g;

The Gamma parameter gmin: The initial value is 0.01, which is used to control the minimum value of the SVM parameter g;

The experimental results of the two methods are shown in Figure 8.

Figure 8 shows the results of the prediction of the New York Harbor's water level near Red Hook by the two different methods, as well as the results of the comparison between the calculated values and the observed values. From the 240 nodes, we can clearly see that the prediction error of this method is the smallest. In addition, it shows that the use of data is good and bad, not only related to the WSN, but also closely related to data mining technology.

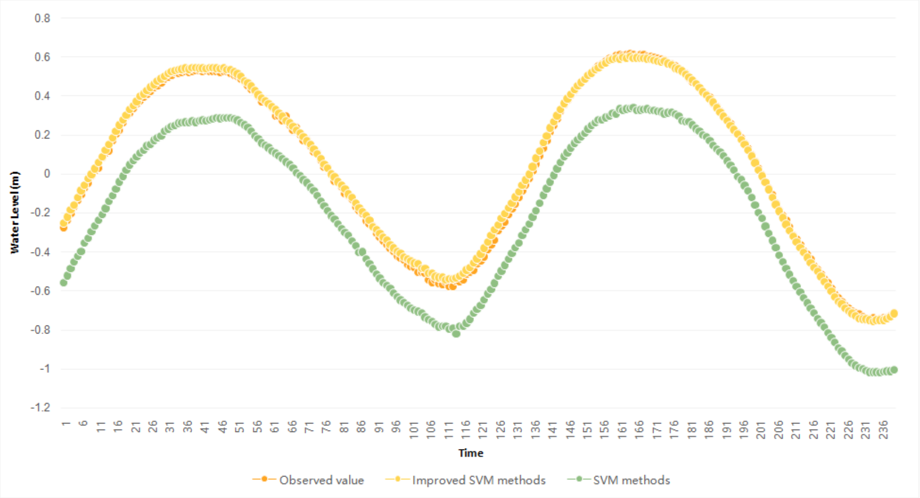

Fig. 8. Comparison of prediction results

5 Conclusions

In this paper, we present an online monitoring scheme by using WSN technology and computer technology, so as to collect hydrological data effectively. Moreover, this paper uses the method of support vector regression to process the data collected by the wireless sensor. Also, in order to obtain the most important parameters of the support vector regression algorithm, this paper uses the PSO algorithm to search for the global optimal solution through the cooperation and competition between the particles in the population. In the simulation experiment, a collection and observation system of New York Harbor is constructed based on the hydrological situation near the harbor. We use the traditional support vector regression method and the improved method to achieve our goal, namely, to use water temperature, salinity, and temperature to predict changes in the water levels. The experimental results show that the proposed algorithm can improve the data utilization rate of WSNs and has good prediction accuracy.

6 References

[1] Ren, F.Y., Huang H.N., Lin C. (2003). Wireless Sensor Networks, Journal of software, 14(7), 1282-1291.

[2] Mainwaring, A., Polastre J., Szewczyk R. (2003). Wireless sensor networks for habitat monitoring, ACM International Workshop on Wireless Sensor Networks and Applications, 88-97.

[3] Jiang, P., Kong, Y.F. (2008). Design of wireless sensor network data video base station for water environment monitoring, Chinese Journal of sensor technology, 21(9), 1580-1585.

[4] Du, X.M., Chen, Y. (2008). Research status and application of wireless sensor networks, Journal of Beijing Technology and Business University (Natural Science Edition), 2008, 26(1), 41-44.

[5] EmNet.LLC.Technology[EB/OL].http://www.heliosware.com/technology.html, 2008-01-16.

[6] The CSIRO ICT Centre. Wireless Sensor Network Devices [EB/OL]. http://www.ict.csiro.au/page.phpcid=87,2008-01-16.

[7] Jiang, P., Huang, Q., Wang, J., Dai, X., Lin, R. (2006). Research on wireless sensor networks routing protocol for water environment monitoring in wetlands, International Conference on Innovative, Computing, Information and Control, 251-254.

[8] Wooley, B., Bridges, S., Hodges, J., Skjellum, A. (2000). Scaling the Data Mining Step in Knowledge Discovery Using Oceanographic Data, IEA/AIE, 85-92.

[9] Guo, D., Evangelinos, C., Patrikalakis, N. (2004). Flow Feature Extraction in Oceanographic Visualization, Computer Graphics International, 162-173.

[10] Hun, C.X. (2002). Research progress of Marine Geographic Information System (MGIS), Marine geology letters, 18(1), 23-30.

[11] Liu, Z., Li, J. (2004). A preliminary study on GIS data warehouse model for digital seabed, Journal of Xi'an Petroleum University, 19(1), 77-81.

7 Author

Wei Zhai is Xi'an Aeronautical University of China, Xi'an 710077, China, she is manly engaged in the research of electronics and communication.

Article submitted 07 February 2018. Resubmitted 23 February 2018. Final acceptance 31 March 2018. Final version published as submitted by the author.