The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality

Y. R. Ding*[1], Y. J. Cai[2], P. D. Sun[3] and B. Chen[4]

[1] Department of computer science and technology,
JiangNan University, Wuxi, China.
* yr_ding@jiangnan.edu.cn
[2] School of Biotechnology,
JiangNan University, Wuxi, China.,
[3] School of Chemical and Material Engineering,
JiangNan University, Wuxi, China.
[4] Environmental Monitoring Station of Binhu District,
Wuxi, China.

ABSTRACT
To effectively control and treat river water pollution, it is very critical to establish a water quality prediction system. Combined Principal Component Analysis (PCA), Genetic Algorithm (GA) and Back Propagation Neural Network (BPNN), a hybrid intelligent algorithm is designed to predict river water quality. Firstly, PCA is used to reduce data dimensionality. 23 water quality index factors can be compressed into 15 aggregative indices. PCA improved effectively the training speed of follow-up algorithms. Then, GA optimizes the parameters of BPNN. The average prediction rates of non-polluted and polluted water quality are 88.9% and 93.1% respectively, the global prediction rate is approximately 91%. The water quality prediction system based on the combination of Neural Networks and Genetic Algorithms can accurately predict water quality and provide useful support for real-time early warning systems.

Keywords: back propagation neural network, genetic algorithm, principal component analysis, water quality prediction.

1. Introduction

Rapid economic growth inevitably causes water pollution. To effectively control water pollution, automatic water quality monitoring stations are built in many important districts. Accurate water quality prediction methods are very important to monitor and control water pollution timely. Therefore, a powerful water quality prediction methods are vital when automatic water quality monitoring systems are established

So far, many methods are used to predict water quality including grey relational method [1], mathematical statistics method [2], model-based approach [3], Bayesian approach [4], neural network model [5-8], and Genetic Algorithm (GA) [9-11]. Approximately, 85%-90% of the water quality prediction work have been completed using Neural Network. Neural network has many favourable characteristics, including mass information processing, distributed association, and the ability of self-learning and self-organizing [12-16]. As a high non-linear system, it also has a good fault-tolerance ability and a good applicability to complex problem. However, the non-linear transfer function of Neural Network has multiple local optimum solutions. Generally, the optimization process is influenced by the selection of initial point. If the initial point is closer to the local optimum point than to the global optimum point, it will cause the multi-layer network failing to obtain global optimum solutions. However, GA can avoid these problems easily. GA cannot be restricted by search space, it can obtain a global optimum solution of discrete, multi-extremum high-dimensional problems with noise. GA has been used in water quality model calibration [9], river water quality management model optimization [10], and water quality monitoring networks optimization [11]. Then, combining BP Neural Network (BPNN) with GA can improve prediction accuracy and speed of BPNN [16-18]. In this paper, GA is used to optimize BPNN parameters to speed the prediction process. The difference from other works is that we apply Principal Component Analysis (PCA) in the system to reduce data dimensionality and speed the learning process.

Many factors affect water quality (There are 23 factors in our work, see materials and methods section). These factors have complex non-linear relationship with water quality. Then, the data dimensionality should be reduced to extract the most important factors. PCA is a technology that can compress multiple original indices into a few aggregative variable indices, which can represent original data information. PCA has been successfully applied in environmental data analysis [19,20]. Here, PCA is applied to optimize and select the sample set.

In this work, we combined PCA, BPNN and GA to predict water quality. By integrating the advantages of these algorithms, the water quality prediction system can not only ensure the prediction accuracy of water quality, but also can improve prediction speed.

2. Materials and methods

2.1 Dataset

Experimental data are from the detection data of rivers flowing into Taihu Lake, China. There are 2680 sample data. They were categorized into two groups, that is, non-polluted and polluted water. The ratio is approximately 1:1. 23 influencing factors of water quality are pH, NH3-N, volatile phenol, TN, Cr6+, CODMn, TP, BOD5, TCN, COD, petroleum, Cd, Cu, Zn, Pb, Hg, As, Se, F-, sulfide, dissolved oxygen, electrical conductivity, and LAS.

2.2 Principal component analysis (PCA)

PCA applies the idea of dimensionality reduction under the premise that the minimum original data loss is guaranteed. It can compress multiple original indices into a few aggregative variable indices. In this paper, we assume the water sample number is n (here n=2680), the number of factors affecting the water quality is p (here p=23); thus, a water quality data matrix of n*p (2680*23) order is constituted. The original sample data matrix is $X=\left\{\begin{matrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{matrix}\right.$, The new variable target denotes as vector y_1, y_2, y_3, y_m (m≤p). Y is linear combination of the data X.

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mp}x_p \end{cases} \left(m \le p\right) \qquad (1)$$

In the Eq. 1, the loading vector $a_i = \left\{a_{i1}, a_{i2}, ..., a_{ip}\right\} \ (i=1,2,...,m)$ is determined by $(\sum \ -\lambda_i I)a_i = 0$, satisfying the following conditions:

(1) y_i is uncorrelated to y_j to form the orthogonal subspace (i≠j).

(2) $a^T{}_i \sum a_i$, the variance of yi, is maximized.

(3) $a^T{}_i a_i = 1$, a_i is standardized.

Eigenvalue decomposition of the covariance matrix of X determines the loading vector a_i as an eigenvector associated with eigenvalues λ_i. $\lambda_i / \sum_{j=1}^{p} \lambda_j (i=1,2,...,p)$ is the contribution of PC*i*. The PC*i* contribution indicates the ability of PCs to represent the original data. After ranking the value of λ_i (usually in descending order), the first PCs with the largest eigenvalues are selected. The criterion is the cumulative value up to 85%. The selected PCs are aggregative indices that are used in BPNN.

2.3 Optimize BPNN using GA

The BP network model contains one hidden layer. For the determination of hidden layer node number, empirical formula estimating or trial

Download English Version:

https://daneshyari.com/en/article/725301

Download Persian Version:

https://daneshyari.com/article/725301

Daneshyari.com