

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

Y. R. Ding^{*1}, Y. J. Cai², P. D. Sun³ and B. Chen⁴

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 $n \times p$ (2680×23) order is constituted. The original sample data

$$X = \begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{Bmatrix}$$

matrix is X , The new variable target denotes as vector $y_1, y_2, y_3, \dots, y_m$ ($m \leq p$). Y is linear combination of the data X .

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

In the Eq. 1, the loading vector $a_i = \{a_{i1}, a_{i2}, \dots, a_{ip}\}$ ($i=1, 2, \dots, m$) 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 ($i \neq j$).
- (2) $a_i^T \sum a_i$, the variance of y_i , is maximized.
- (3) $a_i^T a_i = 1$, 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 .

$\lambda_i / \sum_{j=1}^p \lambda_j$ ($i=1, 2, \dots, p$) 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

method of repeated trial calculation are mostly adopted. Here, the hidden layer neuron number is determined according to the experimental Eq. 2. The different Q are tested and 8 is more appropriate on condition that the goal and gradient are met as possible. In the end, the final network structure is 15-8-1.

$$Q = \sqrt{(\text{input nodes} + \text{output nodes})} + C \quad C = [1, 10] \quad (2)$$

In our experiment, the output value is limited to the range of [0, 1], and we select logsig as the transfer function from the input to the hidden layers and from the hidden to the output layers. BPNN training Levenberg-Marguardt (LM) is applicable to the centre network of sufficient memory. Applying the LM optimization algorithm to water quality prediction may shorten the learning time and improve the training speed. BPNN has problems in slow convergence rates and appearances of a “local minimum” in convergence learning. The big challenge of water quality prediction is that there is a complex non-linear recessive relationship between input and output data. Then, it is very practical to obtain an useful model through a large amount of sample learning and training.

2.4 The combined model of PCA, BPNN and GA

We combined PCA, BPNN and GA algorithms to establish a water quality prediction system. PCA is used to remove some redundant information to reduce data dimensionality and obtain principal components. Using obtained principal components as network input neurons has many advantages: (1) reducing node number of the network input layer, (2) simplifying neural network structure, (3) improving both BPNN training speed and model prediction rate accuracy with GA optimization network parameters. Figure 1 is the simplified flowchart of the combined model.

The steps of the combined PCA, BPNN and GA algorithm to predict water quality are: (1) Converted 2680 groups of sample data into their corresponding 2680 groups of aggregative index sample data; the data were normalized and labeled; (2) Conducted PCA in input samples $X_1, X_2, \dots, X_{23}$; converted them into aggregative index $Z_1, Z_2, \dots, Z_m$ ($m < 23$); (3) Selected BPNN hidden layer neuron number from repetitive

random testing; (4) Randomly divided the 2680 normalized aggregative index sample data into training (2000 sample data) and testing sets (680 sample data); BPNN is carried out with GA (GA optimizes BPNN weights and threshold values).

2.5 Verify and test the combined model

We randomly selected 2000 groups of data for training, with 1000 groups of polluted samples and non-polluted samples. A five-fold cross validation is used to estimate the performance of the hybrid intelligent algorithm. The predictive value outputted by BPNN with GA approached 1 or 0, which could predict whether the water quality is polluted. The local and global prediction accuracies are computed according to Eqs. 3 and 4.

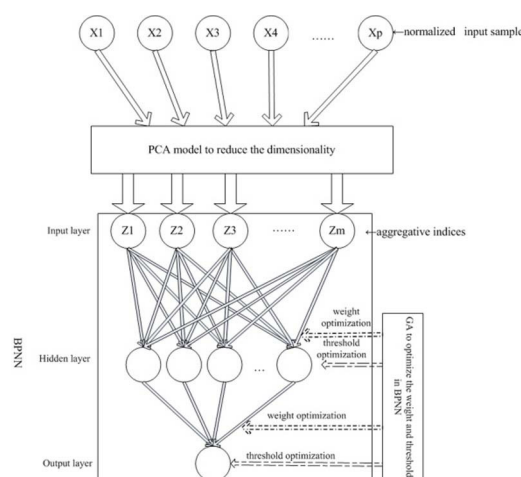


Figure 1. The simplified flowchart of the combined PCA, BPNN and GA algorithm.

Local prediction accuracy (LA):

$$LA = \frac{\sum_{i=1}^{\rho} P_i}{\rho}, \text{ here, } P_i = \frac{T_i}{n_i} \quad (3)$$

Overall prediction accuracy (TA):

$$TA = \frac{\sum_{i=1}^{\rho} T_i}{N} \quad (4)$$

In the Eqs.3 and 4, N is the number of all sample data, ρ is class of sample data (non-polluted or

polluted water), n_i is number of class i , T_i is number of correctly predicted samples in class i . After that, the remaining 680 sample data is used to test the combined model.

3. Results and discussion

3.1 Principal component analysis

After conducting PCA, 23 original sample indices are compressed into 15 aggregative indices. Table S1 shows the related coefficient matrix in PCA. Table 1 shows eigenvalues and contribution rates.

Principal component	Eigenvalue	Contribution rate (%)	Cumulative rate (%)
1	3.90	16.97	16.97
2	2.26	9.85	26.81
3	1.92	8.34	35.15
4	1.65	7.19	42.34
5	1.41	6.13	48.48
6	1.29	5.60	54.07
7	1.18	5.12	59.20
8	1.05	4.58	63.78
9	1.01	4.38	68.16
10	0.89	3.86	72.03
11	0.83	3.61	75.64
12	0.76	3.29	78.93
13	0.74	3.20	82.13
14	0.63	2.75	84.88
15	0.59	2.55	87.43
16	0.57	2.49	89.92
17	0.47	2.05	91.97
18	0.42	1.82	93.80
19	0.35	1.51	95.31
20	0.33	1.41	96.72
21	0.27	1.17	97.90
22	0.26	1.12	99.02
23	0.23	0.98	100.00

Table 1. Eigenvalues and contribution rates.

The relevant matrices show that there is a strong correlation between volatile phenol and NH₃-N, TN and COD and NH₃-N, hexavalent chromium and volatile phenol, COD_{mn} and COD, TP and NH₃-N and TN, BOD₅ and COD_{mn}. Obviously, the information overlapped.

The characteristic values and contribution rates in Table 1 show that the first 15 principal components can represent 87.43% information of the original data. Then, 15 principal components can replace 23 primary data. And these 15 principal components are input neurons for BPNN. The

dimensionality reduction can speed the training process with less information loss.

3.2 The performance of the combined model

We use the remaining 680 sample data to test the performance of the combined model.

Table 2 shows that prediction accuracy of polluted water, prediction accuracy of non-polluted water, and global prediction accuracies are 93.1%, 88.9% and 91% respectively. And, the prediction accuracies of polluted water are all larger than that of non-polluted water. In this work, the river data are determined from 2001. In 2007, a large bloom of blue-green algae in Taihu Lake caused water quality to deteriorate severely. When we randomly choose the training data, if the number of the data in 2007 is larger, the prediction accuracy of polluted water is higher, while the non-polluted water is lower. The strong characteristic of heavily polluted water in this period may affect the result. At the same time, these prediction accuracies show that the combined model is suitable for predicting water quality. Most of all, this algorithm is very stable due to using GA to adjust BPNN connection weight and threshold values.

3.3 Comparison of BPNN performance with and without GA

We also compared BPNN prediction rates with or without using GA. Table 3 shows the results.

Table 3 and Figure 2 shows that BPNN search process with GA is unlikely to be entangled with the local optimum solution. Most predicted rates are approximately 90%, although the predicted accuracy is higher. BPNN predicted rate without using GA optimization sometimes achieves rates above 80%, but repeated experiments show that the trained model predicted rates float larger and sometimes converge to local optimum solutions in the BP network without genetic algorithm optimization. That can be proved by the MSE. The MSE with GA is significantly smaller than the MSE without GA. The smaller the MSE, the better the convergence. In the search process of BPNN without GA, the optimum solution cannot be searched, and the predicted accuracy declines. To overcome the disadvantages of BPNN, GA is necessary to optimize BPNN parameters.

	COD	pH	NH3-N	volatile phenol	TN	Cr6+	CODMn	TP	BOD5	TCN	petroleum	Cd	Cu	Zn	Pb	Hg	As	Se	F--	sulfide	LAS	dissolved oxygen	electrical conductivity
COD	1	-0.1	0.33	0.07	0.41	-0.05	0.41	0.28	0.19	0.07	0.18	-0.07	0	-0.1	-0.1	-0.1	0.06	0.09	0.31	-0.01	0.01	-0.08	0.29
pH	-0.1	1	-0.16	0.01	-0.13	0.05	-0.21	-0.15	0.02	-0.06	-0.1	0.17	-0.1	-0.3	-0	0.15	-0.1	-0.1	-0.02	-0.09	0.04	0.15	-0.14
NH3-N	0.33	-0.2	1	0.41	0.66	0.12	0.38	0.45	0.31	0.08	0.11	-0.1	0.2	0.08	-0	-0.1	0.22	0.02	0.29	0.14	0.02	-0.12	0.39
volatile phenol	0.07	0.01	0.41	1	0.3	0.5	0.16	0.13	0.1	0.02	-0.04	-0.12	0.34	0	0	-0.1	0.09	0.04	0.05	0.15	-0.01	-0.06	0.2
TN	0.41	-0.1	0.66	0.3	1	0.1	0.25	0.44	0.28	0.06	0.01	-0.02	0.09	0.05	-0.1	-0	0.09	0.15	0.23	0.11	0.11	-0.03	0.38
Cr6+	-0.1	0.05	0.12	0.5	0.1	1	0	0	0.04	-0.01	0	0.01	0.29	0	0	-0	0.08	0	-0.06	0.02	0	-0.01	-0.14
CODMn	0.41	-0.2	0.38	0.16	0.25	0	1	0.38	0.54	0.09	0.2	-0.16	0.22	0.01	0.01	-0.1	0.35	-0	0.2	0.19	-0.07	-0.12	0.27
TP	0.28	-0.2	0.45	0.13	0.44	0	0.38	1	0.33	0.04	0	-0.1	-0	-0	-0.2	-0	0.07	-0	0.27	0.27	0	-0.09	0.32
BOD5	0.19	0.02	0.31	0.1	0.28	0.04	0.54	0.33	1	0.02	0.12	0.04	0.18	-0.1	0.21	-0.1	0.26	-0	0.13	0.19	-0.04	-0.05	0.12
TCN	0.07	-0.1	0.08	0.02	0.06	-0.01	0.09	0.04	0.02	1	-0.02	-0.01	0.01	0.19	0	-0	0	0	0.07	0.03	0.02	0	0.08
petroleum	0.18	-0.1	0.11	-0.04	0.01	0	0.2	0	0.12	-0.02	1	-0.26	0	0.04	0.06	-0.1	0.4	-0	0.12	-0.05	-0.03	-0.04	0.11
Cd	-0.1	0.17	-0.1	-0.12	-0.02	0.01	-0.16	-0.1	0.04	-0.01	-0.26	1	0.03	0.11	0.01	0.14	-0.1	0.01	-0.16	-0.12	0.01	0.05	-0.19
Cu	0	-0.1	0.2	0.34	0.09	0.29	0.22	-0.04	0.18	0.01	0	0.03	1	0	0.02	0	0.48	-0	-0.26	0	0.01	-0.05	-0.16
Zn	-0.1	-0.3	0.08	0	0.05	0	0.01	-0.02	-0.06	0.19	0.04	0.11	0	1	-0.1	0	-0.1	0.06	-0.05	-0.01	0.08	0.04	-0.06
Pb	-0.1	-0	-0.03	0	-0.06	0	0.01	-0.15	0.21	0	0.06	0.01	0.02	-0.1	1	0.05	0.23	-0	-0.06	-0.01	0	-0.03	-0.11
Hg	-0.1	0.15	-0.1	-0.07	-0.03	-0.01	-0.07	-0.02	-0.06	-0.01	-0.05	0.14	0	0	0.05	1	0.02	0	-0.36	0.04	0.26	0.45	-0.23
As	0.06	-0.1	0.22	0.09	0.09	0.08	0.35	0.07	0.26	0	0.4	-0.1	0.48	-0.1	0.23	0.02	1	-0	-0.08	-0.03	0.03	-0.04	0.03
Se	0.09	-0.1	0.02	0.04	0.15	0	-0.03	-0.01	-0.02	0	-0.02	0.01	-0	0.06	-0	0	-0	1	-0.04	-0.01	-0.01	-0.01	0.06
F--	0.31	-0	0.29	0.05	0.23	-0.06	0.2	0.27	0.13	0.07	0.12	-0.16	-0.3	-0.1	-0.1	-0.4	-0.1	-0	1	0.11	-0.18	-0.3	0.52
sulfide	-0	-0.1	0.14	0.15	0.11	0.02	0.19	0.27	0.19	0.03	-0.05	-0.12	0	-0	-0	0.04	-0	-0	0.11	1	0	-0.01	0.14
LAS	0.01	0.04	0.02	-0.01	0.11	0	-0.07	0	-0.04	0.02	-0.03	0.01	0.01	0.08	0	0.26	0.03	-0	-0.18	0	1	0.68	-0.03
dissolved oxygen	-0.1	0.15	-0.12	-0.06	-0.03	-0.01	-0.12	-0.09	-0.05	0	-0.04	0.05	-0.1	0.04	-0	0.45	-0	-0	-0.3	-0.01	0.68	1	-0.12
electrical conductivity	0.29	-0.1	0.39	0.2	0.38	-0.14	0.27	0.32	0.12	0.08	0.11	-0.19	-0.2	-0.1	-0.1	-0.2	0.03	0.06	0.52	0.14	-0.03	-0.12	1

Table S1. The related coefficient matrix.

Number of times	Size of training set	Size of testing set	Percent of accuracy in predicting polluted water	Percent of accuracy in predicting non-polluted water	Percent of overall prediction accuracy
1	2000	680	0.910	0.875	0.891
2	2000	680	0.895	0.805	0.850
3	2000	680	0.970	0.940	0.955
4	2000	680	0.920	0.890	0.905
5	2000	680	0.960	0.935	0.948
Average			0.931	0.889	0.910

Table 2. Prediction accuracy of polluted water and non-polluted water.

No.	Average prediction rate with GA	Average prediction rate without GA	Mean Squared Error with GA	Mean Squared Error without GA
1	0.891	0.711	0.0036	0.0620
2	0.850	0.841	0.0510	0.0530
3	0.955	0.684	0.0032	0.0760
4	0.905	0.856	0.0035	0.0210
5	0.948	0.703	0.0032	0.0740

Table 3. Average prediction rates and Mean Squared Error with and without GA in BPNN.

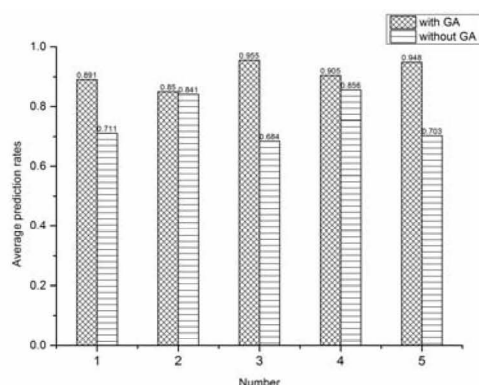


Figure 2. The prediction rate of water quality with and without GA algorithm.

Table 3 and Figure 2 shows that BPNN search process with GA is unlikely to be entangled with the local optimum solution. Most predicted rates are approximately 90%, although the predicted accuracy is higher. BPNN predicted rate without using GA optimization sometimes achieves rates above 80%, but repeated experiments show that the trained model predicted rates float larger and sometimes converge to local optimum solutions in the BP network without genetic algorithm optimization. That can be proved by the MSE. The MSE with GA is significantly smaller than the MSE without GA. The smaller the MSE, the better the convergence. In the search process of BPNN without GA, the optimum solution cannot be searched, and the predicted accuracy declines. To overcome the disadvantages of BPNN, GA is necessary to optimize BPNN parameters.

4. Conclusions

We present a water quality prediction model that combines PCA, BPNN and GA. Using BPNN model to study water classification and prediction can overcome disadvantages including the large workload of traditional evaluation methods and strong subjectivity. This model possesses objectivity, universality and practicality. PCA converts the multi-indices into a few aggregative indices with little original data information loss and reduces the input data to speed the training process. Using GA to optimize network parameters can effectively prevent the search process from converging to local optimum

solutions, optimize global optimal network parameters, and significantly improve the accuracy of water quality prediction. This model makes full use of the advantages and characteristics of PCA, BPNN and GA algorithms to predict water quality. This model can obtain high training speed and good prediction rate and can be extended to other classification problem.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (21001053), the National High Technology Research and Development Program (2009AA02C210) and the Fundamental Research Funds for the Central Universities (JUSRP11126).

References

- [1] W. C. Ip, et al., "Applications of grey relational method to river environment quality evaluation in China," *Journal of Hydrology*, vol. 379, no. 3-4, pp. 284-290, 2009.
- [2] E. M. Smeti, et al., "Treated water quality assurance and description of distribution networks by multivariate chemometrics," *Water Research*, vol. 43, no. 18, pp. 4676-4684, 2009.
- [3] D. A. Brydon and D. A. Frodsham, "A model-based approach to predicting BOD5 in settled sewage," *Water Science and Technology*, vol. 44, no. 2-3, pp. 9-15, 2001.
- [4] Liu Y, et al., "Water quality modeling for load reduction under uncertainty: a Bayesian approach," *Water Research*, vol. 42, no. 13, pp. 3305-3314, 2008.
- [5] V. Chandramouli, et al., "Backfilling missing microbial concentrations in a riverine database using artificial neural networks," *Water Research*, vol. 41, no. 1, pp. 217-227, 2007.
- [6] H. G. Han, et al., "An efficient self-organizing RBF neural network for water quality prediction," *Neural Network*, vol. 24, no. 7, pp. 717-725, 2011.
- [7] E. O'Connor, et al., "A neural network approach to smarter sensor networks for water quality monitoring," *Sensors (Basel)*, vol. 12, no. 4, pp. 4605-4632, 2012.
- [8] O. Senkal, et al., "Precipitable water modelling using artificial neural network in Cukurova region," *Environmental Monitoring and Assessment*, vol. 184, No. 1, pp. 141-147, 2012.

- [9] J. H. Cho, et al., "A river water quality management model for optimising regional wastewater treatment using a genetic algorithm," *Journal of Environmental Management*, vol. 73, no. 3, pp. 229-242, 2004.
- [10] R. Zou and W. S. Lung, "Robust water quality model calibration using an alternating fitness genetic algorithm," *Journal of Water Resources Planning and Management*, vol. 130, no. 6, pp. 471-479, 2004.
- [11] Y. Icaga, "Genetic algorithm usage in water quality monitoring networks optimization in Gediz (Turkey) river basin," *Environmental Monitoring and Assessment*, vol. 108, no. 1-3, pp. 261-277, 2005.
- [12] L. L. Rogers, et al., "Optimal field-scale groundwater remediation using neural networks and the genetic algorithm," *Environmental Science & Technology*, vol. 29, no. 5, pp. 1145-1155, 1995.
- [13] S. Ledesma-Orozco, et al., "Hurst Parameter Estimation Using Artificial Neural Networks," *Journal of Applied Research and Technology*, vol. 9, no. 2, pp. 227-240, 2011.
- [14] A. Baharodimehr, et al., "Capacitive MEMS accelerometer wide range modeling using artificial neural network," *Journal of Applied Research and Technology*, vol. 7 no. 2, pp. 185-192, 2009.
- [15] M. R. Arab, et al., "Electroencephalogram Signals Processing for the Diagnosis of Petit mal and Grand mal Epilepsies Using an Artificial Neural Network," *Journal of Applied Research and Technology*, vol. 8 no.1, pp. 120-129, 2010.
- [16] J. Rivera-Mejía, et al., "PID based on a single artificial neural network algorithm for intelligent sensors," *Journal of Applied Research and Technology*, vol. 10, no. 2, pp. 262-282, 2012.
- [17] V. Petridis, et al., "A hybrid neural-genetic multimodel parameter estimation algorithm," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 862-876, 1998.
- [18] J. T. Kuo, et al., "A hybrid neural-genetic algorithm for reservoir water quality management," *Water Research*, vol. 40, no. 7, pp. 1367-1376, 2006.
- [19] S. G. Dalal, et al., "Evaluation of significant sources influencing the variation of water quality of Kandla creek, Gulf of Katchchh, using PCA," *Environmental Monitoring and Assessment*, vol. 163, no. 1-4, pp. 49-56, 2010.
- [20] R. L. Olsen, et al., "Water quality sample collection, data treatment and results presentation for principal components analysis--literature review and Illinois River Watershed case study," *Water Research*, vol. 46, no. 9, pp. 3110-3122, 2012.