Application of Decision Tree and Neural Network Algorithm in Water Quality Assessment Forecast

Bing-Xiang LIU[1,a,*], Xiang WAN[1], Xu-Dong WU[2,b], Ying-Xi LI[2], Hui-Qiu ZHU[2]

[1]Jingdezhen Ceramic Institute, Jingdezhen, 333403, China

[2]Jiangxi Ceramic & Art Institute, Jingdezhen, 333000, China

[a]lbx1966@163.com, [b]wxd20077@126.com

*Corresponding author

Keywords: Neural Network, Decision Tree, Water Quality Model.

Abstract. The paper builds water quality assessment forecast model by adopting four algorithms which are decision tree, BP neural network, Logistic recursion and RBF neural network and analyzes the result. The forecast result shows that RBF neural network is the fittest method and its rate of accuracy is high that can be used widely.

Introduction

Water is an important matter that people depend on living and it is important resource that is indispensible and irreplaceable. With the rapid development of economy and speeding up of industrialization process, the increasingly water pollution becomes an import factor that restricts our national strategic of sustainable development. Water environment has become an important problem that many professionals and experts study. Currently, there are many assessment methods studying water environment quality both at home and abroad. There are dozens of methods about reactive literature discussing water quality assessment which present very active situations. Because of the uncertainty of all kinds of factors in water environment and the unknown of water body, the traditional certain assessment method is hard to adapt study. Although some water quality model considers many factors influencing water quality change and the simulative forecast effect is idea, it is rather sophisticated and needs numerous basic material and data which make the study's further development and application be restricted. Currently, more common situation is to forecast water quality's future variation tendency according to water condition of the current water of the current water body migration traits of pollutant, discharge situation of pollutant in the river basin and it is very necessary to find a suitable model for mater quality pollution forecast.

Relative Theory Analysis

Decision Tree

Decision is one of the most widely used inductive inference algorithms and it is a kind of approaching discrete value function method which has good robustness for noise and can study an extract expression. Decision tree is usually generated from above to below and it uses greedy searching traversal method to traverse. Every decision or event may extract two or more events and cause different results. This kind of decision braches can be drawn in fig.s which look like the branches of tree and that's why it is called decision tree. The decision tree algorithms in common use age include C&RT algorithm,ID3 typical algorithm,C4.5 algorithm,C5.0 algorithm, CHAID algorithm and QUEST algorithm.

The basic algorithm of decision tree is greedy algorithm and it builds decision tree with rout one by one recursively from above to below. The most famous decision tree algorithm is ID3 algorithm which mainly directs at attribute selection and is the most influential and typical algorithm among decision tree learning methods. The ID3 method uses information gain degree selection test attribute.

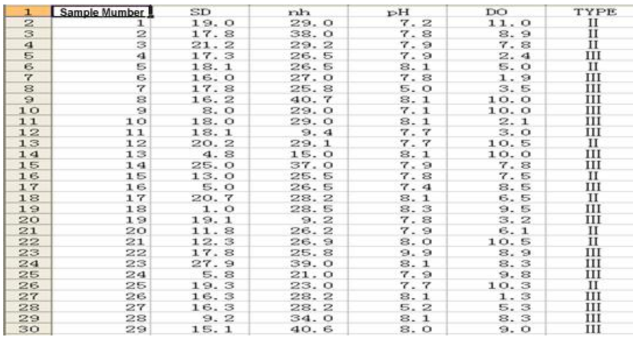
Radial Basis Function (RBF)Neural net work

The structure of RBF network is similar as multilayer feed forward network and it is three-layer feed forward network. The input layer is constituted by signal nodes; the second layer is implicit layer and the number of implicit units change according to the questions they describe. The transformation function of implicit units is RBF that is non-minus and non-linear function gathering towards the center decrease. The third layer is output layer which responds the effect of input model. The transformation from input space to implicit space is no-linear, while the transformation from implicit layer space to output layer space is linear.

Experiment Process

This paper gathers and extracts a group of users basic data and pre-processes the data which is shown in Table 1.

Tab. 1 Information Chart



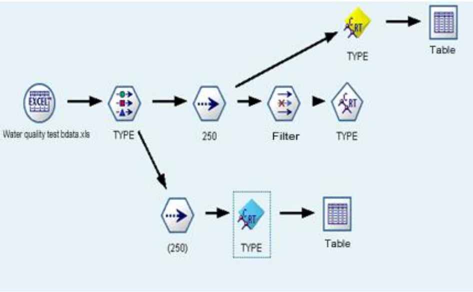
1	Sample Mumber	SD	nh	pH	DO	TYPE
2	1	19.0	29.0	7.2	11.0	II
3	2	17.8	38.0	7.8	8.9	II
4	3	21.2	29.2	7.9	7.8	II
5	4	17.3	26.5	7.9	2.4	III
6	5	18.1	26.5	8.1	5.0	II
7	6	16.0	27.0	7.8	1.9	III
8	7	17.8	25.8	5.0	3.5	III
9	8	16.2	40.7	8.1	10.0	III
10	9	8.0	29.0	7.1	10.0	III
11	10	18.0	29.0	8.1	2.1	III
12	11	18.1	9.4	7.7	3.0	III
13	12	20.2	29.1	7.7	10.5	II
14	13	4.8	15.0	8.1	10.0	III
15	14	25.0	37.0	7.9	7.8	III
16	15	13.0	25.5	7.8	7.5	II
17	16	5.0	26.5	7.4	8.5	III
18	17	20.7	28.2	8.1	6.5	II
19	18	1.0	28.5	8.3	9.5	III
20	19	19.1	9.2	7.8	3.2	III
21	20	11.8	26.2	7.9	6.1	II
22	21	12.3	26.9	8.0	10.5	II
23	22	17.8	25.8	9.9	8.9	III
24	23	27.9	39.0	8.1	8.3	III
25	24	5.8	21.0	7.9	9.8	III
26	25	19.3	23.0	7.7	10.3	II
27	26	16.3	28.2	8.1	1.3	III
28	27	16.3	28.2	5.2	5.3	III
29	28	9.2	34.0	8.1	8.3	III
30	29	15.1	40.6	8.0	9.0	III

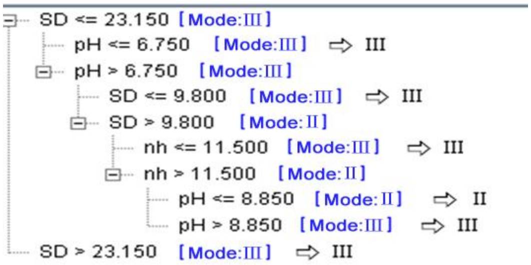
The main indicators influencing water quality are :(1)dissolved oxygen(DO) that measures self-purity ability of the water body sensor network collect automatically.(2)temperature: water temperature changes according to weather(sensor network collect automatically).(3)PH value: effects water quality's PH value(sensor network collect automatically).(4)ammonia nitrogen: represents the content of nutrition pollutants (sensor holder by hands collect manually). (5)oxidation reduction potential(ORP):measure indicator for ox ideation reducing power of aqueous solution(sensor holder by hands collect manually).(6)current situation factor (environmental situation of weather water body).Among that records from 1 to 250 are training samples date and records from 251-300 are measured sample date. This paper forecasts more accurate models by building various models to find a set of fittest assessment methods.

Four key indictors DO、PH、NH、SD are extracted based on water assessment quantification indicator to build water body environment quality assessment quantification model and achieves water quality intelligent resolution. According to the regulations of 《surface water environmental quality standards》(GB3838-2002) on using purpose and protection barges ,there are five kinds of surface water in china: the first kind is mainly used for headstream water and national-level nature reserve, the second is mainly used for centralization water for living and drink、first level conservation district for water head size on the earth's surface, rare aquatic organism habitat, pawing sit for fish and shrimp and nursery ground for larvae and jovial and so on .The third is mainly used for centralization water for living and drink, second conservation district for water head size on the earth's surface, overwinter and migration route for fish and shrimp, aquiculture area and fisheries area and swimming area. This paper studies the three kinds of water quality and conducts real time forecast, classification and simulation on water quality.

Model Building by Decision Tree

Fig.1 shows the data stream build by decision tree algorithm. The core algorithm of decision tree is C&RT model algorithm which can quickly build overall data flow chart. The advantage of this algorithm is that it can initiate mutual conversation as model's building choice. The decision tree can be compiled before the model is generated and knowledge model can be used to generate and trim decision tree.





```
SD <= 23.150 [Mode:III]
    pH <= 6.750  [Mode:III]  ⇨ III
    pH > 6.750  [Mode:III]
        SD <= 9.800  [Mode:III]  ⇨ III
        SD > 9.800  [Mode:II]
            nh <= 11.500  [Mode:III]  ⇨ III
            nh > 11.500  [Mode:II]
                pH <= 8.850  [Mode:II]  ⇨ II
                pH > 8.850  [Mode:III]  ⇨ III
SD > 23.150  [Mode:III]  ⇨ III
```

Fig. 1 Decision Tree Data Flow Chart

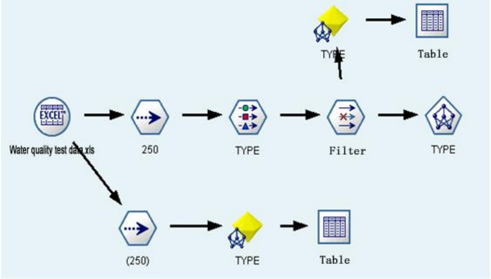
Fig. 2 Decision Tree C&RT Model

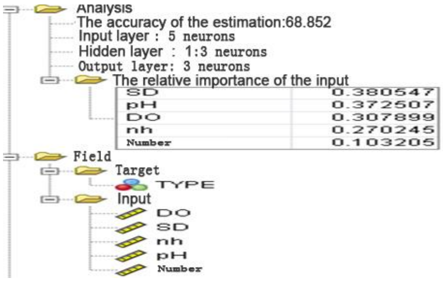
As is shown in fig.2 decision tree C&RT model, which rank water quality lies in can be imaged by the model according to the content of four key indicators: DO、PH、NH、SD. Among that when SD>23.15,it is the third water equation, when SD<23.5 and PH<=6.75,it is the third water quality; when SD<=9.8 and PH>6.75,it is the third water quality; when 9.8<SD<=23.15 and PH>6.75,it is the second water quality; when 9.8<SD<=23.15 AND PH>6.75 and NH<=11.5,it is the third water quality; when 9.8<SD<23.15 and PH>6.75 and PH>11.5, it is the second water quality, when 9.8<SD<=23.15 and 6.75<PH<=8.85 and NH>11.5,it is the second water quality ;when9.8<SD<23.15 and 6.75<PH<=8.85 and NH>11.5,it is the second water quality;when9.8<SD<23.15 AND PH>8.85 and NH>11.5,it is the third water quality.

Design tree model is built by sample data and it is used from 250-3000 recording test data to get following results: forecast accuracy rate is 76%,the variance is 0.021,the standard deviation is 0.144 and the standard error of mean is 0.02.

Model building by BP neural network

This paper user data mining software and water quality assessment forecast model of BP neural network algorithm. The output variation is based on water quality assessment quantification indicator DO、PH、NH、SD. This paper builds water body environment quality assessment quantification model and achieves intelligent recognition of water quality. This paper users 1-250 records as training sample data and 251-300 records as test sample data.





```
Analysis
    The accuracy of the estimation:68.852
    Input layer : 5 neurons
    Hidden layer : 1:3 neurons
    Output layer: 3 neurons
    The relative importance of the input
        SD        0.380547
        pH        0.372507
        DO        0.307899
        nh        0.270245
        Number    0.103205
Field
    Target
        TYPE
    Input
        DO
        SD
        nh
        pH
        Number
```

Fig. 3BP Neural Network Tree Data Flow Chart

Fig. 4 BP Neural Network Model Analysis

As is shown in fig. 3 and 4 ,BP neural network model judges water quality lies in which level according to the content of four key indicators. DO、PH、NH、SD Neural network training model is built through sample data and it is used from 250-300 recording test data and gets the following result: forecast accuracy rate is 68.852%,variance is 0.103 standard deviation is 0.321 and standard error of mean is 0.045.

Building Forecast Logistic Regression Model

Water quality assessment forecast model is built by data mining and Logistic regression algorithm. Output variation is based on water quality assessment qualification indicator DO、PH、NH、SD to build water body environment quality assessment quantification model and achieve intelligent recognition of water body. This paper uses 1-250 records as training sample data and 251-300 records as test sample data.

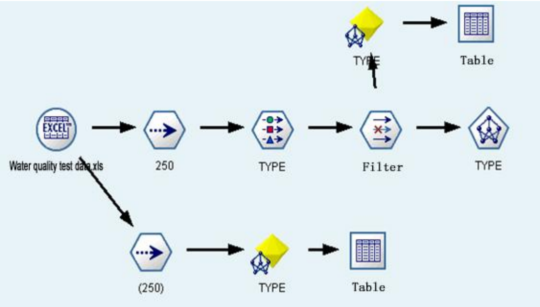
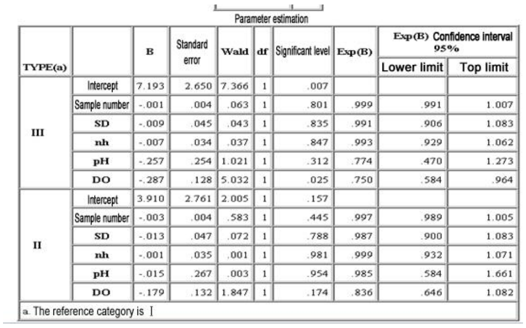


Fig. 5 Logistic Algorithm Data Flow Chart



Parameter estimation

TYPE(a)		B	Standard error	Wald	df	Significant level	Exp(B)	Exp(B) Confidence interval 95%	
								Lower limit	Top limit
III	Intercept	7.193	2.650	7.366	1	.007			
	Sample number	-.001	.004	.063	1	.801	.999	.991	1.007
	SD	-.009	.045	.043	1	.835	.991	.906	1.083
	nh	-.007	.034	.037	1	.847	.993	.929	1.062
	pH	-.257	.254	1.021	1	.312	.774	.470	1.273
	DO	-.287	.128	5.032	1	.025	.750	.584	.964
II	Intercept	3.910	2.761	2.005	1	.157			
	Sample number	-.003	.004	.583	1	.445	.997	.989	1.005
	SD	-.013	.047	.072	1	.788	.987	.900	1.083
	nh	-.001	.035	.001	1	.981	.999	.932	1.071
	pH	-.015	.267	.003	1	.954	.985	.584	1.661
	DO	-.179	.132	1.847	1	.174	.836	.646	1.082

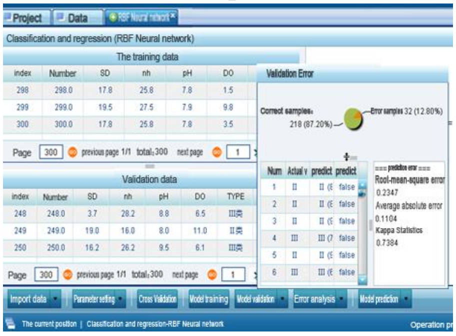
a. The reference category is Ⅰ

Fig. 6 Logistic Model Analysis

As is shown in fig. 5 and 6,Logistic regression algorithm model includes water quality lies in which level according to the content of four key indicators DO、PH、NH、SD. Neural network training model is built by simple data and it is used from 250-300 recording test data and gets the following results: forecast accuracy rate is 90%,variance is 0.011,standard derivation is 0.105 and standard error of mean is 0.015.

Model building by RBF neural network

This paper uses data mining software and water quality assessment forecast model of RBF neural network algorithm. Output variation is on water quality assessment qualification indicator DO、PH、NH、SD to build water body environment quality assessment qualification model and active intelligent cognition of water body. This paper uses 1-250 record as training sample data and 251-300 records as test sample data which are shown in fig. 7 and 8.



Classification and regression (RBF Neural network)

The training data

index	Number	SD	nh	pH	DO
298	298.0	17.8	25.8	7.8	1.5
299	299.0	19.5	27.5	7.9	9.8
300	300.0	17.8	25.8	7.8	3.5

Page 300 previous page 1/1 total：300 next page 1

Validation data

index	Number	SD	nh	pH	DO	TYPE
248	248.0	3.7	28.2	8.8	6.5	III类
249	249.0	19.0	16.0	8.0	11.0	II类
250	250.0	16.2	26.2	9.5	6.1	III类

Page 300 previous page 1/1 total：300 next page 1

Validation Error

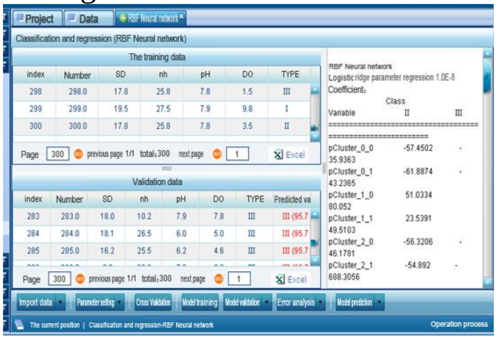
Correct samples：218 (87.20%)

Error samples 32 (12.80%)

Num	Actual v	predict	predict
1	II	II (8	false
2	II	II (8	false
3	II	II (9	false
4	III	III (7	false
5	II	II (9	false
6	III	III (8	false

```
=== prediction error ===
Root-mean-square error
0.2347
Average absolute error
0.1104
Kappa Statistics
0.7384
```

Fig. 7 Neural Network Training Sample Model



Classification and regression (RBF Neural network)

The training data

index	Number	SD	nh	pH	DO	TYPE
298	298.0	17.8	25.8	7.8	1.5	III
299	299.0	19.5	27.5	7.9	9.8	I
300	300.0	17.8	25.8	7.8	3.5	II

Page 300 previous page 1/1 total：300 next page 1

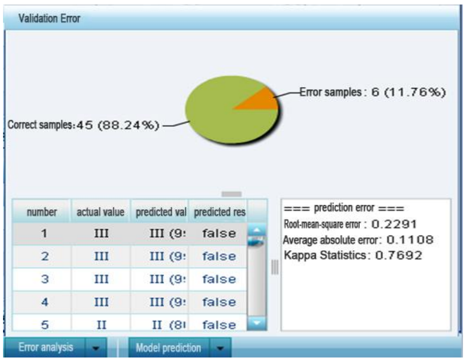
Validation data

index	Number	SD	nh	pH	DO	TYPE	Predicted va
283	283.0	18.0	10.2	7.9	7.8	III	III (95.7
284	284.0	18.1	26.5	6.0	5.0	III	III (95.7
285	285.0	16.2	25.5	6.2	4.6	III	III (95.7

Page 300 previous page 1/1 total：300 next page 1

```
RBF Neural network
Logistic ridge parameter regression 1.0E-8
Coefficient：
                    Class
Variable            II            III
========================================
=========================
pClusler_0_0        -57.4502      -
35.9363
pCluster_0_1        -61.8874      -
43.2365
pCluster_1_0        51.0334
80.052
pCluster_1_1        23.5391
49.5103
pCluster_2_0        -56.3206      -
46.1781
pCluster_2_1        -54.892       -
688.3056
```

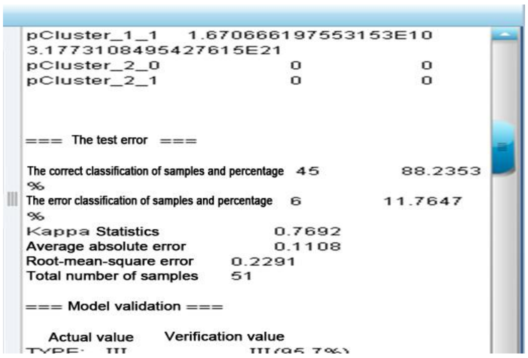
Fig. 8 Neural Network Test Sample Data



number	actual value	predicted val	predicted res
1	III	III (9:	false
2	III	III (9:	false
3	III	III (9:	false
4	III	III (9:	false
5	II	II (8l	false

```
=== prediction error ===
Root-mean-square error：0.2291
Average absolute error：0.1108
Kappa Statistics：0.7692
```

Fig. 9 Test Result Chart



```
pCluster_1_1    1.6706661975531 53E10
3.17731084954276 15E21
pCluster_2_0                 0                0
pCluster_2_1                 0                0



=== The test error ===

The correct classification of samples and percentage  45           88.2353
%
The error classification of samples and percentage   6           11.7647
%
Kappa Statistics                 0.7692
Average absolute error           0.1108
Root-mean-square error       0.2291
Total number of samples      51

=== Model validation ===

   Actual value      Verification value
TYPE:   III                  III(95.7%)
```

Fig. 10 Test Error

Fig. 9 and 10 show test result and error and among forecast error, the accuracy rate is 88.23%, error rate is 11.76%,kappa statistics is 0.7692,mean absolute error is 0.11,root mean square error is 0.2291. From the result, it can be seem that the model's self-study ability is strong and its precision is high and error band is small which suit the need of water quality assessment forecast model.

Conclusion

The result accuracy rates of the above four algorithm are shown in Table2:

Tab.2 Comparison of Accuracy Rates on Four Algorithms

Model algorithm	RBF neural network	Decision tree	Logistic regression	BP neural network
accuracy	88.23%	76%	70%	68.852%

It can be seem from above the date comparative analysis that the forecast result of RBF neural network algorithm is obviously better that of decision tree, Logistic regression algorithm and BP neural network algorithm.RBF neural network improves its algorithm model through many model learning and its quite convergence makes it very suitable for the building of water quality forecast model and further explore the law of water quality change process. While Logistic regression doesn't have the advantage of neural network from both the request for sample and forecast precision. Therefore, the application result shows that it is feasible to use RBF neural network to conduct water quality forecast and can offer a more efficient new method for water quality simulation forecast that can be popularized in water quality forecast.

References

[1] Wang Hailing,Cao Jing. The initially establish of pond water color discrimination system based on the Optimized Algorithm of L-M Neural Network [J]. Fishery Modernization. 2010, 37(5): 19-21.

[2] Zou Zhihong, Sun Jingnan, Ren Guangping. Study and Application on the Entropy method for Determination of Weight of evaluating indicators in Fuzzy Synthetic Evaluation for Water Quality Assessment [J]. 2005, 25(4).

[3] Liu Zhangun, Ye Liaoyuan. Fuzzy Probability Model and Its Application to Evaluation of Groundwater Quality [J]. Journal of basic science and engineering. 2007, 15(3).

[4] Wang Dongsheng, Li Shihua, Zhou Xingpeng. Assessment method of raw water quality based on PSO-RBF neural network model and its application [J].Journal of Southeast University(Natural Science Edition). 2011, 41(5).