유전자 알고리즘과 신경망 이론을 이용한 수질예측

하수정[1] · 김동렬[2] · 김용구[3] · 박성천[4]

1. 서 론

급속한 산업화와 도시화에 따라 용수의 사용량이 증대되고, 배출되는 산업폐수와 생활 오수 등은 심각한 수질 오염을 일으키고 있으며, 오염물질의 부하량 증대 및 하천의 부영양화를 유발하여 하천 수계의 자연정화능력이 한계에 이르렀다. 따라서 하수와 폐수로 인한 수질오염을 방지하고 보다 효과적인 수질관리에 대한 대안이 요구되고 있다. 본 연구는 영산강 유역의 대표지점인 나주지점을 선정하여 비선형적인 하천의 수질을 유전자 알고리즘과 신경망 이론을 이용하여 모형을 구성하여 수질을 예측하고자 하였다. 구성된 자료는 입력층의 수가 n이라고 하면 은닉층의 수는 n~3n까지 변화시키면서 신경망의 연결강도 등의 매개변수를 유전자 알고리즘을 통해 최적화시켰다.

일련의 과정을 거쳐 탐색되어진 모형들에 대한 평가기준은 도식적인 기준과 수치적 기준을 적용하였다. 수치적 기준으로는 *CC*(Correlation Coefficient)와 *RMSE(Root Mean Square Error)*를 구하여 모형을 선택하고 평가하였다.

2. 연구의 이론적 배경

2.1 유전자 알고리즘과 신경망의 결합

유전자 알고리즘은 Darwin의 진화론에 발상의 기본을 두고 생물 진화의 과정을 추상화시킨 알고리즘이다. 신경망 이론은 두뇌를 구성하는 신경 회로망을 추상화하고, 몇 가지 태스크를 수행시키도록 하자는 발상이다. 이 두 가지에는 많은 공통점이 있다. 먼저 양자는 모두 넓은 의미에서 학습과 적응에 관한 모델이다. 또한 양자 모두 실제 생물에 관련된 원리를 추상화한 것이라든지, 병렬도가 높은 수법이라는 점등은 매우 유사하다. 그러나 신경망이 한 개의 개체의 학습을 다루고 있는데 반하여, 유전자 알고리즘은 종의 적응을 다루고 있다는 면에서 차이점을 가지고 있다. 본 연구에서 실시한 유전자알고리즘과 신경망의 결합 방법은 각 염색체에 실치표현(real value encoding)을 사용하였으며, 그 방법의 도식적인 표현은 Fig. 1과 같다.

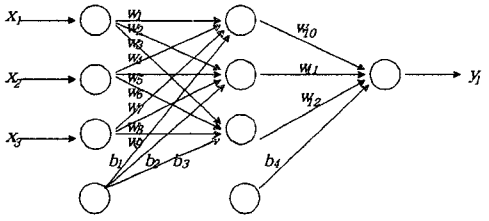


Fig. 1 연결강도 등의 염색체 표현

유전자 알고리즘을 이용한 신경망의 학습을 수행할 때에는 각 염색체를 네트워크에 사상시켜, 다집단유전자 알고리즘(Multiple Populations Genetic Algorithm; *MPGA*)을 이용하여 최적해 근방까지 학습시킨다.

모든 네트워크의 적응도가 결정되면, 다음 세대에 남기기 위해 선택, 교배, 돌연변이의 유전자 조작이 적용된다. 이와 같은 절차를 반복하는 중에 적응도는 향상되어 간다. 이렇게 *MPGA*에서 1차 수렴단계를 거쳐 가장 우수한 적응도를 가진 염색체는 신경망(Neural Network ; NN)으로 사상된다. NN에서는 최급 강하법, 적응식 학습율과 역전파 알고리즘(back propa -gation algorithm; BPA)의 early stopped training approach를 이용하여 학습을 실시한 후 최종 수렴단계를 거쳐 출력값을 갖는다.

3. 대상유역 및 지점

본 연구에서 사용된 영산강 유역은 동경 126°27′~127°05′과 북위 34°40′~35°29′사이의 우리나라 남서부인

1) 동신대학교 공업기술연구소 연구원
2) 동신대학교 토목공학과 석사과정
3) 동신대학교 토목공학과 박사과정
4) 동신대학교 토목 · 환경공학부 부교수

호남지방에 위치하고 있으며, 유역면적은 3,429km^2, 동서와 남북간의 최장거리는 각각 61.3km 및 89.7km이며, 유역의 평균폭이 26.5km이다. 영산강의 발원지인 용추봉에서 하구지점까지의 유로연장은 129.5km이다.

4. 입력자료의 처리 및 구성

4.1 입력 자료의 처리

산정된 입력자료와 출력자료는 유전자 알고리즘과 신경망(Genetic Algorithm & Neural Network : GANN)에 적용시키기 위해서 전처리과정(pre-processing)과 후처리과정을 거쳤다. 본 연구에서 실시한 전처리과정을 위한 방법으로 평균이 0, 그리고 표준편차가 1인 표준정규분포 N(0, 1)이 되도록 식 (2)와 같이 정규화 하였다. 유전자알고리즘과 신경망에 의해 결정된 매개변수를 이용하여 모의된 자료는 후처리과정을 거쳐 실제값으로 복원된다.

$$Z = \frac{X_i - \mu}{\sigma} \quad \text{----- 식 (1)}$$

(Z : 정규화된 자료, X_i : 각 자료의 i번째 자료, μ : 각 자료의 평균, σ : 각 자료의 표준편차)

4.2 모형의 구성

본 연구에서는 영산강 유역의 대표지점인 나주지점에 대한 DO농도, BOD농도, T-N농도, T-P농도 수질항목을 예측하기 위하여 유전자 알고리즘과 신경망을 조합한 모형을 구성하였다. 각 모형은 입력층의 노드의 수를 n개라 할 때 은닉층의 노드의 수를 n~3n개까지 변화시키면서 모형을 구성하였으며, 이러한 과정을 거친 모형들은 수치적 평가기준인 *CC*와 *RMSE*를 적용하여 최적의 모형을 선별하였다.

1) DO농도 모형

DO농도를 예측하기 위하여 시행착오방법에 의해 DO농도의 시차를 4로 고정하고 BOD, Temp, Q_min, Q-max은 시차를 1~4까지 변화를 주어 모형을 구성하였다. 구성한 모형 중 DO농도의 예측에 적용성이 뛰어난 모형은 ModelⅡ로 판별되었다.

2) BOD농도 모형

BOD농도를 예측하기 위하여 시행착오방법에 의해서 BOD농도의 시차를 4로 고정하고, T-N농도, T-P농도, Q_avg, Q_max을 시차 1~4까지 변화를 주어 모형을 구성하였다. 구성한 모형 중 BOD농도의 예측에 뛰어난 모형은 ModelⅡ로 판별되었다.

3) T-N농도 모형

T-N농도를 예측하기 위하여 시행착오방법에 의해 T-N농도의 시차를 3으로 고정하고 T-P, Q_min, Q_max을 시차 1~3까지 달리하여 모형을 구성하였다. 구성된 모형 중 T-N농도의 예측에 뛰어난 모형은 ModelⅡ로 판별되었다

4) T-P농도 모형

T-P농도를 예측하기 위하여 시행착오방법에 의해 T-P농도의 시차를 3으로 고정하고 T-P, temp, ss, Q_max을 시차 1~3까지 달리하여 모형을 구성하였다. T-P농도의 예측에 뛰어난 모형은 ModelⅡ로 판별되었다.

다음의 식 (2)~식 (5)은 각각의 선택된 모형 ModelⅡ를 나타내고 있다.

$$\text{Model Ⅱ}\ DO_k = \phi \begin{bmatrix} do_{k-1},\ do_{k-2},\ do_{k-3},\ do_{k-4} \\ bod_{k-1},\ bod_{k-2} \\ temp_{k-1},\ temp_{k-2} \\ qmin_{k-1},\ qmin_{k-2} \\ qmax_{k-1},\ qmax_{k-2} \end{bmatrix} \quad \text{식 (2)}$$

$$\text{Model Ⅱ}\ BOD_k = \phi \begin{bmatrix} bod_{k-1},\ bod_{k-2},\ bod_{k-3},\ bod_{k-4} \\ t\text{-}n_{k-1},\ t\text{-}n_{k-2} \\ t\text{-}p_{k-1},\ t\text{-}p_{k-2} \\ qmax_{k-1}, qmax_{k-2} \\ qavg_{k-1}, qavg_{k-2} \end{bmatrix} \quad \text{식 (3)}$$

$$\text{Model Ⅱ}\ T\text{-}N_k = \phi \begin{bmatrix} tn_{k-1},\ tn_{k-2},\ tn_{k-3} \\ tp_{k-1},\ tp_{k-2} \\ qmin_{k-1}, qmin_{k-2}, \\ qmax_{k-1}, qmax_{k-2}, \end{bmatrix} \quad \text{식 (4)}$$

$$\text{Model Ⅱ}\ T\text{-}P_k = \phi \begin{bmatrix} tp_{k-1},\ tp_{k-2},\ tp_{k-3} \\ temp_{k-1},\ temp_{k-2} \\ ss_{k-1},\ ss_{k-2}, \\ qmax_{k-1}, qmax_{k-2}, \end{bmatrix} \quad \text{식 (5)}$$

5. 모의결과 및 고찰

본 연구에서 실시한 유전자 알고리즘과 신경망의 이론에 의한 하천수 수질예측 모의 결과의 정확성 평가하기 위해서 상관계수(*CC* : Correlation Coefficient)와 평균제곱오차의 평방근(*RMSE* : Root Mean Squared Error), 그리고 선형도시법을 사용하였다. *CC*는 0과 1사이의 범위 값을 가지며 값이 1에 가까울수록 모형의 정확도와 적합성이 뛰어난 것으로 판별한다. *RMSE*는 관측값과 계산값의 제곱오차로부터 평균제곱오차를

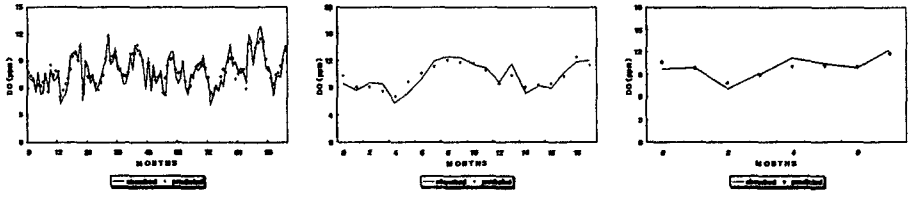
구하여 아래 식과 같이 재곱근을 구한 것으로 그 값이 작을수록 모형의 오차가 작을 것으로 판별한다.

$$CC = \frac{Cov(X, Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}} \quad \text{식 (7)} \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - Y_i)^2} \quad \text{식 (8)}$$

모형의 선별과정은 각 농도의 예측모형 중 훈련 보정 과정에서 예측의 적용성이 뛰어난 모형을 선택하였고, 검증과정은 훈련 및 보정 과정에서 선별된 모형을 대상으로 검증을 실시하여 최종 모형을 선택하였다.

5.1 DO농도 모형

본 연구에서 DO농도의 모형을 개발하기 위하여 모형을 구성한 결과 ModelⅡ가 타 모형에 비하여 예측력의 우수성이 인정되었다. 그 결과는 Table 1에서 나타난 바와 같으며, 훈련과 보정, 검증과정에서 *CC*와 *RMSE*값이 우수한 모형을 판별한 결과, 입력층의 노드의 수가 12개, 은닉층 노드의 수가 22, 출력층이 1개인 DO_GANN(12,22,1)이 선택되었다. Fig. 2는 본 모형에 대한 도식적인 표현을 나타내고 있다.

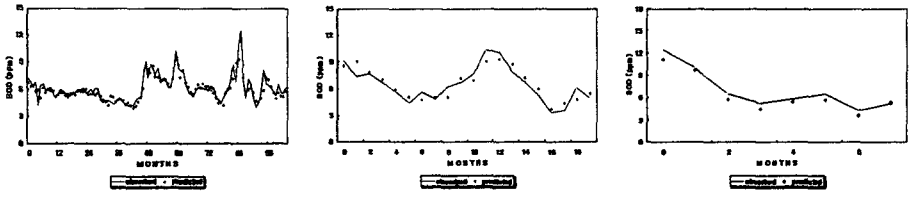


1) 훈련과정 2) 보정과정 3) 검증과정

Fig. 2 DO_GANN(12,22,1)모형에 대한 도시

5.2 BOD농도 모형

본 연구에서 BOD농도의 모형을 개발하기 위하여 모형을 구성하여 적용한 결과 ModelⅡ가 타 모형에 비하여 우수성이 인정되었다. 훈련과 보정, 검증과정에서 *CC*와 *RMSE*값이 우수한 모형을 판별한 결과, Table 1에서 나타난 바와 같이 입력층의 노드의 수가 12개, 은닉층 노드의 수가 32, 출력층이 1개인 BOD_GANN(12,32,1)이 선택되었다. Fig. 3은 선택된 BOD 모형에 대한 도식적인 표현을 나타내고 있다.

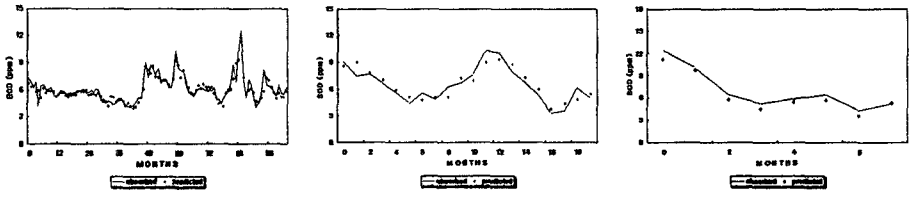


1) 훈련과정 2) 보정과정 3) 검증과정

Fig. 3 BOD_GANN(12,32,1)모형에 대한 도시

5.3 T-N농도 모형

본 연구에서 T-N농도의 모형을 개발하기 위하여 모형을 구성하여 적용한 결과 ModelⅡ가 타 모형에 비하여 우수성이 인정되었다. 훈련과 보정에서 *CC*가 0.8이상인 모형들을 1차적으로 선택하고 1차적으로 선택된 모형을 대상으로 검증과정을 실시하여 그 통계적 특성치를 나타내었다. 그 결과 Table 1에 나타난 바와 같이 입력층의 노드의 수가 9개, 은닉층 노드의 수가 22, 출력층이 1개인 T-N_GANN(9,22,1)이 선택되었다. Fig. 4은 선택되어진 모형에 대한 도식적인 표현을 나타내고 있다.



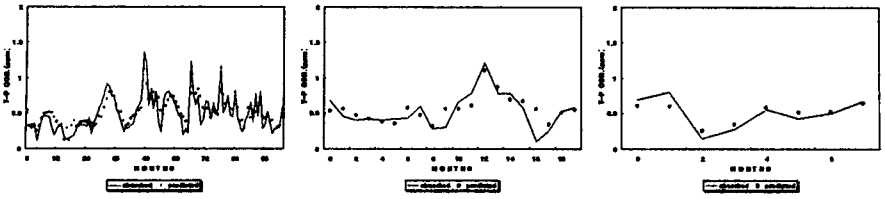
1) 훈련과정 2) 보정과정 3) 검증과정

Fig. 4 T-N_GANN(9,22,1)모형에 대한 도시

5.4 T-P농도 모형

본 연구에서 T-P농도의 모형을 개발하기 위하여 모형을 구성하여 적용한 결과 ModelⅡ가 타 모형에 비

하여 우수성이 인정되었다. 그 결과는 Table 1과 같으며 최종 모형의 선택은 1차적으로 선택된 모형 중 검증에서 *CC*와 *RMSE*가 우수한 T-P_GANN(9, 14, 1)의 모형을 선택하였다. Fig. 5는 선택된 T-P모형에 대한 도식적인 표현을 나타내고 있다.



1) 훈련과정 2) 보정과정 3) 검증과정

Fig. 5 T-P_GANN(9, 14, 1)모형에 대한 도시

Table 1. 선택된 모형들에 대한 통계적 특성치

Model	Hidden layer	Training		Validation		Verification	
		RMSE	CC	RMSE	CC	RMSE	CC
DO_GANN(12, 22, 1)	n+10	0.688	0.907	1.099	0.856	0.392	0.917
BOD_GANN(12, 32, 1)	2n+8	0.402	0.902	1.016	0.864	1.379	0.939
T-N_GANN(9, 22, 1)	2n+4	1.109	0.934	1.625	0.889	1.830	0.929
T-P_GANN(9, 14, 1)	n+5	0.022	0.810	0.022	0.792	0.028	0.873

6. 결 론

본 연구에서는 영산강 유역 나주지점의 DO, BOD, T-N, T-P농도를 예측하기 위하여 유전자 알고리즘과 신경망이론을 이용한 이론적 모형을 개발하였다. 개발된 모형은 1990년~2000년도까지의 자료를 이용하여 훈련과 보정과정을 거쳤으며 2001년도 자료로 검증과정을 거쳤다. 개발된 DO농도 모형은 DO, BOD, Q_max, Q_min, Temp을 입력 자료로 하여 구성하였고, 입력층의 노드의 수는 12, 은닉층의 노드의 수는 12~36까지 변화를 주었다. 그 결과 검증과정에서의 *CC*는 0.815~0.917, *RMSE*는 0.392~1.364로 예측성능이 우수한 것으로 평가되었다. BOD농도 모형은 BOD, T-N, T-P, Q_avg, Q_max을 입력 자료로 모형을 개발하였다. 입력층의 노드의 수는 12, 은닉층의 노드의 수를 12~36까지 변화를 주어 구성한 결과, 검증과정의 *CC*는 0.804~0.939, *RMSE*는 0.843~1.379로 탁월한 예측 결과를 보여주고 있다. T-N농도를 예측하기 위해 T-N, T-P, Q_min, Q-max을 입력 자료로 모형을 개발하였다. 입력층의 노드의 수는 9, 은닉층의 노드의 수를 9~27까지 변화를 주어 모형을 구성한 결과 검증과정에서 *CC*는 0.882~0.929, *RMSE*는 1.830~2.219로 우수한 예측력을 보여주고 있다. T-P를 예측하기 위해 개발한 모형은 T-P, Temp, SS, Q_max을 입력 자료로 구성하였다. 입력층의 노드의 수는 9, 은닉층의 노드의 수는 9~27까지 변화를 주어 모형을 구성하였다. 그 결과 검증과정에서의 *CC*는 0.620~ 0.873, *RMSE*는 0.028~0.063로 비교적 양호한 결과를 보여주고 있다. 따라서 본 연구에서 유전자 알고리즘과 신경망 이론을 이용하여 개발한 모형을 하천에 적용하여 일자료 또는 시자료의 구축이 선행된다면 더 우수한 모형을 개발하여 하천의 수질을 보다 적극적으로 관리할 수 있을 것으로 판단된다.

참고문헌

진영훈(2000). 하천의 유출량 예측을 위한 인공신경망의 적용. 석사학위 논문, 전남대학교.

노경범(2001). 유전자 알고리즘과 신경망의 결합에 의한 유출량 예측. 석사학위 논문, 전남대학교.

오창렬(2001). 신경망 이론을 이용한 하천의 수질예측. 석사학위 논문, 동신대학교.

박성천외 2인(2001). GANN에 의한 하천의 일유출량 예측. 대한토목학회논문집, 대한토목학회, 제 21권 6-B pp.609-617.

박성천 외 3인(2001). 신경망을 이용한 영산강 수질예측. 대한 토목학회 논문집, 대한토목학회, 제 22권 6-B pp.371-382 2001.

Chipperfield, A. et al(1994). Genetic Algorithm Toolbox for use with 5. Matlab. Version1.2. Department of Automatic Control and Systems Engineering, University of Sheffield,