퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 6, pp. 777-784

PLS기반 c-퍼지 모델트리를 이용한 클로로필-a 농도 예측

Chlorophyll-a Forcasting using PLS Based c-Fuzzy Model Tree

이대종*, 박상영**, 정남정**, 이혜근**, 박진일***, 전명근***

Dae-Jong Lee, Sang-Young Park, Nahm-Chung Jung,
Hye-Keun Lee, Jin-Il Park, Meung-Geun Chun

* 충북대학교 BK21 충북정보기술사업단
** 한국 수자원 공사 수자원연구원
*** 충북대학교 전기전자컴퓨터 공학부

요 약

본 논문에서는 부분최소법 (PLS: Partial least square)과 c-퍼지 모델트리를 적용하여 클로로필-a 농도의 예측 모델을 제안한다. 제안된 방법은 모든 입력속성을 고려하여 퍼지 클러스터에 의해 계산된 중심벡터를 설정한 후, 각각의 중심벡터들과 입력속성간의 소속도를 이용하여 내부 노드를 형성하고, 형성된 내부노드에서 PLS를 적용하여 지역모델(Local model)을 구축한다. 노드의 분리기준으로서 부모노드(parent node)에서 구축된 모델에서 계산된 에러값이 자식노드(child node)에서 계산된 에러값보다 클 경우에 분기가 이루어진다. 최종 단계에서는 임의의 입력데이터와 잎노드에서 계산된 클러스터 중심값과 비교하여 소속도가 높은 클러스터에 속한 지역모델을 선택하여 출력값을 예측한다. 제안된 방법의 우수성을 보이기 위해 수질 데이터를 대상으로 실험한 결과 기존의 모델트리 방식에 비하여 향상된 성능을 보임을 알 수 있었다.

키워드 : 모델트리, 퍼지 클러스터, 퍼지 모델트리, 부분최소법

Abstract

This paper proposes a c-fuzzy model tree using partial least square method to predict the Chlorophyll-a concentration in each zone. First, cluster centers are calculated by fuzzy clustering method using all input and output attributes. And then, each internal node is produced according to fuzzy membership values between centers and input attributes. Linear models are constructed by partial least square method considering input-output pairs remained in each internal node. The expansion of internal node is determined by comparing errors calculated in parent node with ones in child node, respectively. On the other hands, prediction is performed with a linear model having the highest fuzzy membership value between input attributes and cluster centers in leaf nodes. To show the effectiveness of the proposed method, we have applied our method to water quality data set measured at several stations. Under various experiments, our proposed method shows better performance than conventional least square based model tree method.

Key Words : Model Tree, Fuzzy Cluster, Fuzzy Model Tree, Partial Least Square

1. 서 론

다목적댐은 홍수조절 외에 식수와 용수의 공급원으로서 중요한 기능을 담당하고 있다. 그러나 물의 지속적인 유입 및 유출이 이루어지는 유수 생태계(lotic ecosystem)과 달리 호수, 저수지, 댐 등의 정수생태계(lentic ecosystem) 서는 물의 흐름이 원활하지 않아 자정작용이 약하고, 오염된 수질을 회복하는데 오랜 시간이 소요되는 문제점을 지니고 있다. 이러한 정수생태계에서 물이 한곳에 체류하는 시간이 증가할수록 영양분의 양이 많아지고 퇴적물도 증가하면서 심각한 부영양(Eutrophic) 상태의 부영양화 현상이 발생하고, 이로 인해 생·공용수 이용에 많은 악영향을 초래하고 있으며, 특히 상수원 수질오염으로 인한 피해는 심각한 실정이다. 부영양화가 진행되면 식물플랑크톤이 증가되어 빛의 투과도가 낮아져 수중에 사는 침수수초는 살 수가 없게 되며, 특히 산소가 부족하게 되어 유기물이 박테리아에 의해 분해되는 과정에서 메탄, 암모니아, 황화수소 등 환원기체가 발생함으로써 물에서 썩는 냄새가 난다. 일반적으로 식물플랑크톤의 성장은 영양염 농도, 일사량, 수온, 식물플랑크톤의 사멸속도, 침강속도 및 동물플랑크톤의 포식작용 등에 의해 좌우된다. 이러한 식물플랑크톤의 현존량을 나타내는 중요한 간접지표로 클로로필-a 농도를 이용하며, 따라서 효과적인 수질관리를 위해서는 클로로필-a의 농도를 분석 및 예측할 수 있는 모델 개발이 시급한 실정이다[1-4].

수질분석 및 특정 수질매개변수의 예측을 위해 다양한 데이터 마이닝 기법이 이용되고 있으며, 주된 알고리즘으로는 학습을 통한 최적화 기법으로 예측, 군집, 분류 등에 이용되고 있는 신경망 기법과 높은 적중률을 보이면서 해석 또한 용이한 결정트리(DT; Decision Tree), 모델트리(MT; Model

접수일자 : 2006년 10월 5일
완료일자 : 2006년 11월 30일
감사의 글 : 본 연구는 한국 수자원공사의 지원에 의해 이루어졌음.

Tree) 기법 등이 대표적이며, 이 밖에도 결과에 대해 수학적 설명이 가능한 통계적 기법으로 회귀분석, 군집분석, 시계열 분석 등이 있다.

모델트리는 말단의 잎노드에 속한 출력값의 평균값을 계산하는 회귀트리와 달리 연속적인 입력값과 출력값을 이용하여 예측 오차값이 최소화되는 계수값을 계산한 후, 계산된 계수값을 이용하여 출력값을 예측한다[5][6]. 이러한 모델트리도 회귀트리와 같이 데이터를 반복적으로 분리하여 트리구조를 생성하는 상-하 추론 모델트리(TIMIT:Top-down Induction of Model Tree) 형식을 갖는다. 이러한 트리구조의 추론방식은 몇 가지 문제점 즉, 다중 입력변수 중에서 뿌리노드의 기준이 되는 첫 번째 주요 입력속성을 선정하는 문제, 입노드들의 결정 그리고, 잎노드에서 모델의 선택 등이 선행되어야 한다. 특히, 다중입력 속성 중에서 하나의 선택된 입력속성에 의존하여 트리의 구조를 확장시키는 것은 몇 가지 문제점을 초래할 수 있다. 물론, 단일입력 속성을 선택함으로써 트리구조 자체가 간단하고 명료해 보이지만, 단일 속성만 선택함으로써 다중 속성을 선택한 경우보다 트리 가지의 수가 구조적으로 증가할 우려가 높다. 또한, 어떤 노드에서 주어진 분리기준에 의해 분할되지 않는 입력속성이 존재하는 경우 두 개 또는 그 이상의 입력속성을 동시에 고려하는 것이 분류문제에서 효과적인 것으로 보고되고 있다[7].

이러한 모델트리의 문제점 이외에도 표본의 수보다 특징변수의 수가 많은 경우 트리 말단의 잎노드에서 얻어진 선형모델의 신뢰성이 저하되는 단점을 지니고 있다. 특히, 년 단위로 측정되는 수질데이터의 경우 취득된 특징변수의 수가 많은 반면에 표본의 개수는 충분하지 못하다. 따라서 모델트리로 분할을 지속할 경우 말단 잎노드에 속한 특징변수의 수가 표본변수의 수보다 적은 문제점이 발생한다. 이럴 경우 일반적인 회귀문제에서 부분최소제곱(PLS:Partial Least Square) 방법이 적용된다. 이 방법은 예측하고자 하는 종속변수와 특징변수 사이의 관계를 모형화하는 방법으로, 특징변수의 수가 많아서 특징변수들 간의 상관관계가 높을 경우에도 다른 방법에 비해 우수한 성능을 나타내는 것으로 보고되고 있다 [8-10].

본 논문에서는 PLS와 c-퍼지 모델트리를 적용하여 수질예측에 중요한 판단기준으로 사용되는 클로로필-a 농도를 예측할 수 있는 모델을 제안하고자 한다. 기존의 모델트리 방식과 달리 클러스터링 기반 퍼지모델트리인 c-FMT(Clustering based Fuzzy Model Tree)는 클러스터링 기법을 이용하여 단일 입력속성만을 고려하지 않고 모든 입력속성을 고려하여 분리기준을 판정함으로써 트리구조의 간결성뿐만 아니라 성능에서도 모델트리기법에 비해 우수한 것으로 나타났다[7]. 제안된 방법은 모든 입력속성을 고려하여 퍼지 클러스터에 의해 계산된 중심벡터를 설정한 후, 각각의 중심벡터들과 입력속성간의 소속도를 이용하여 내부 노드를 형성하고, 형성된 내부노드에서 PLS를 적용하여 지역모델(Local model)을 구축한다. 노드의 분리기준으로서 부모노드(parent node)에서 구축된 모델에서 계산된 에러값이 자식노드(child node)에서 계산된 에러값보다 클 경우에 분기가 이루어진다. 최종 단계에서는 임의의 입력데이터와 잎노드에서 계산된 클러스터 중심값과 비교하여 소속도가 높은 클러스터에 속한 지역모델을 선택하여 출력값을 예측한다.

논문의 구성은 2장에서는 모델트리방식과 PLS 기반 c-퍼지 모델트리에 대하여 설명한다. 3장에서는 제안된 방법을 수질데이터에 적용한 실험 및 고찰을 설명하고, 마지막으로 4장에서 결론을 맺는다.

2. 모델트리와 PLS기반 c-퍼지 모델트리

2.1 모델 트리

모델트리는 회귀트리와 구조적으로 동일하지만 회귀트리는 말단의 잎노드에 속한 연속적인 출력값의 평균값을 계산함으로써 예측력의 저하를 초래한다. 이러한 문제점을 해결하기 위해 모델트리 기반의 다양한 알고리즘이 제안되고 있으며, 주된 차이점으로는 각각의 잎노드에 속한 평균값을 취하는 회귀트리와 달리 연속적인 입력과 출력값을 대상으로 예측값과 실제 출력값과의 에러가 최소화되는 선형모델을 생성하고, 생성된 선형모델을 이용하여 출력값을 예측한다.

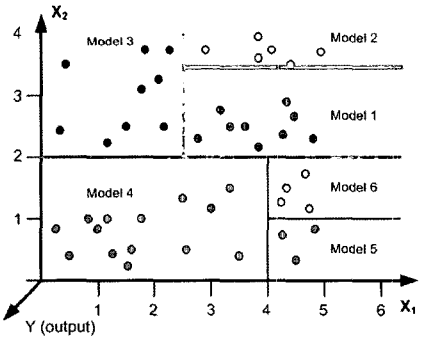
모델트리를 추론하기 위해 M5 알고리즘이 이용된다. M5 모델트리 알고리즘은 식 (1)에서 보인 바와 같이 해당되는 내부마디에 존재하는 입출력 데이터의 표준 편차와 상위노드와 하위노드와의 감소율에 기인한 SDR (Standard Deviation Reduction)을 분리기준으로 사용하고 있다[5].

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \qquad (1)$$

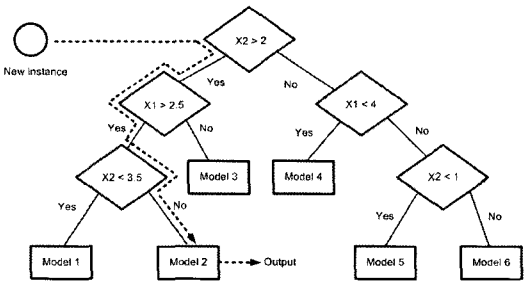
여기에서 T는 도달한 마디의 예제들의 집합이고, $T_1, T_2, \cdots$ 들은 선택된 속성에 따라 분리된 마디로 부터의 결과 집합들이다.

모델트리는 입력속성들 중에서 하나의 속성만을 대상으로 SDR을 계산한 후, SDR이 최대가 되는 수치값을 기준점으로 하여 입력공간을 분할한다. 동일한 방법으로 설정된 조건을 만족할 때까지 노드의 분기는 지속된다. 그러나 지나치게 많은 마디를 가지는 트리구조는 새로운 자료를 적용할 때 예측 오차가 매우 커지는 경향이 있다. 따라서 모델트리구조가 형성된 후, 주어진 트리구조에서 적절하지 않은 마디를 제거하여 적당한 크기의 구조를 갖도록 가지치기(Prunning) 과정을 수행하게 된다. 마지막 단계에서는 가지치기 과정을 수행한 각각의 말단에 위치한 잎노드에서 계산된 선형 모델들 사이에서 필연적으로 발생할 수 있는 비연속적인 값들을 보상해주는 평활화(smoothing)과정이 수행된다. 이런 모든 과정을 거친 후, 임의의 입력데이터에 대하여 루트노드로부터 말단의 잎노드까지 경로를 탐색한 후, 잎노드에서 계산된 선형계수값을 이용하여 출력값을 예측한다.

그림 1에서는 모델트리의 생성 및 추론과정을 나타냈다. 구조를 나타냈다. 그림 1(a) 에서 보는 바와 같이 분리기준에 의해 2차원의 입력공간이 6개의 모델을 갖는 부분공간으로 분할되었다. 여기서, 각각의 분할모델은 선형회귀모델 ($y = a_0 + a_1 x_1 + a_2 x_2$)을 갖는다. 그림 1(b)에서는 새로운



(a) 모델트리의 생성
(a) Building a model tree



(b) 모델트리의 추론

(b) Inducing a model tree

그림 1. 모델트리의 생성 및 추론 과정

Fig. 1. Building and inducing process of a model tree

입력속성에 대한 추론과정을 나타냈다. 새로운 입력속성은 뿌리노드의 조건 ($X_2 > 2$)으로부터 시작하여 내부노드를 거쳐 말단의 잎노드인 Model 2를 탐색하였다. 그림 1(b)에서 점선이 새로운 속성이 뿌리노드에서 입노드까지의 추적경로(path)를 나타낸다. 최종적으로 입력속성에 대한 출력값은 Model 2에서 미리 계산된 선형계수값을 이용하여 예측한다.

2.2. PLS 기반 c-퍼지 모델트리

기존의 모델트리 방식[5]과 c-퍼지 모델트리 방식[6]은 말단의 잎노드에서 선형모델을 구하기 최소자승법에 기반을 둔 다중선형회귀분석(Multiple linar regression)을 이용한다. 그러나 수처리 데이터와 같이 데이터의 개수가 충분하지 못할 경우 트리구조에 의해 입출력 데이터의 분할이 지속될 경우는 일년에 한번 일출력 데이터를 얻을 수 있으므로 얻는 수처리 데이터의 경우 구하고자 하는 독립변수보다 데이터의 개수가 적어질 가능성과 변수간의 상관성이 높게 나타나 비정칙(singularity) 문제가 발생하여 정확한 선형모델을 구축하는데 어려움이 있다. 이를 해결하기 위한 한 방법으로서 상관성이 강하고 노이즈가 많이 함유된 데이터를 상관성이 적은 저차원의 모델로 해석하기 위해 주성분분석기법(PCA:Principal Component Analysis)을 이용한 주성분 회귀(PCR: Principal Component Regression)이 제안되었으나, 입력변수의 분산만을 이용하여 회귀모델을 구축함으로써 입력과 출력간의 관계를 효과적으로 나타내는 데는 한계가 있다 [8-10].

부분최소자승법은 기존의 선형회귀모델에서 문제시 되는 제약조건 없이 입력변수의 분산을 고려하면서 출력변수와의 상관성도 최대화 할 수 있는 방법으로, 다양한 분야에서 널리 적용되고 있으며, 특히 입력(설명)변수의 수가 출력(종속)변수의 수보다 많을 때 효과적인 것으로 보고되고 있다. 물론 회귀식측면에서 입력변수만으로 주성분을 만들어 내는 주성분회귀와 비슷한 방법이지만, 주성분회귀에서 주성분은 입력변수의 데이터값 만으로 만들어지는데 반하여 부분최소법에서의 성분은 입력과 출력의 관계를 이용하여 만들어낸다는 큰 차이점이 있다.

m차원의 출력변수 $Y \in R^{n \times m}$와 상관성이 매우 높은 n개의 데이터를 갖는 p차원의 입력변수 $X \in R^{n \times p}$ $(n \ll p)$의 입출력 데이터를 고려하자. 다중 선형회귀모델의 한 부류인 부분최소자승법은 입출력간의 관계인 $Y = XB + E$를 표현해 줄 수 있는 회귀계수 행렬값 $B \in R^{p \times m}$을 구하는 문제로부터 시작된다. 이를 위해 부분최소법에서는 식 (2)에 나타낸 바와 같이 잠재변수(latent variable) 행렬 $T(T \in R^{n \times c})$, $U(U \in R^{n \times c})$ 와 적재(loading) 행렬 $P(P \in R^{p \times c})$, $Q(Q \in R^{m \times c})$를 이용하여 입출력 공간상에서 두 부분으로 분해한 후, 두 수식사이의 관계를 이용한다.

$$X = TP^T + E = \sum_{h=1}^{a} t_h p_h^T + E \qquad (2)$$

$$Y = UQ^T + F = \sum_{h=1}^{a} u_h q_h^T + F$$

여기서, T 와 U는 추출된 요인점수(factor score)행렬이고, E와 F는 오차항이다.

부분최소법은 가능한 오차항 F를 최소화 하면서 출력 Y를 잘 설명할 수 있게 함과 동시에 입력 X와 출력 Y의 유용한 관계를 얻어내는 것이 목표이다. 이를 위해 부분최소법과 같은 요인추출(factor extraction)을 이용한 회귀법은 가중치 행렬 $W \in R^{p \times c}$를 이용하여 요인점수(factor score)행렬 $T = XW$ ($T \in R^{n \times c}$)를 계산한다. 여기서, 가중치행렬 W는 식 (3)과 같이 출력값과 대응되는 요인점수 사이의 공분산이 최대화 되도록 설정한다.

$$W = \arg max(cov(t,u)), cov(t,u) = t^T u/n \qquad (3)$$

본 논문에서는 LSE 기반 c-퍼지 모델트리의 문제점을 개선하기 위하여 PLS 기반 c-퍼지 모델트리를 제안한다. 제안된 PLS기반 c-퍼지 모델트리는 FCM을 알고리즘을 이용한 퍼지 클러스터를 이용하여 부분 모델트리를 형성한다. 즉, FCM에 의해 c개의 클러스터의 중심벡터를 설정한 후, 각각의 중심벡터들과 입력속성간의 소속도를 이용하여 내부 노드를 형성한다[7]. 지역모델(Local model)은 형성된 내부노드에서 PLS를 적용하여 구축한다. 모델의 가지를 확장하는 하위 노드의 분리기준으로는 표 1에서 보인 바와 같이 네 가지 조건을 고려한다.

표 1. 분기조건

Table 1. Split criterion

- 분기 전 예측 오차값이 설정된 값 (S_1)이상일 때
- 분기 후 모든 클러스터에 포함되는 데이터의 개수가 설정된 값 (S_2)이상일 때
- 분기 전과 분기 후의 오차값 향상이 설정된 값 (S_3)이상일 때
- 분기된 트리의 깊이 (depth)가 설정된 값 (S_4)이하일 때

PLS 기반 c-퍼지 모델트리를 이용하여 데이터 모델을 구하는 과정을 단계별로 설명하면 다음과 같다.

[단계 1] 표 1에 언급된 분기조건에 적용되는 값 S_1, S_2, S_3, S_4 을 설정한다.

[단계 2] 모델트리의 특정 노드에 존재하는 $h\,(h \geq S_2)$ 개의 입출력 데이터 $\{X, Y\} \in R^{q \times h}$에 대하여 앞의 식 (2)의 PLS 기법을 이용하여 예측값 $\hat{y}(k)$ 을 구하고, 실제 출력값과 예측값과의 오차값을 다음과 같이 산출한다. 식 (6)으로부터 구한 오차값 E_b 값이 S_1 이상일 때 다음 단계를 실행하고 그렇지 않을 경우 분기를 정지한다.

$$E_b = \sqrt{\sum_{k=1}^{h} (\hat{y}(k) - y(k))^2 / h} \qquad (6)$$

[단계 3] FCM 알고리즘을 이용하여 [단계 1]의 노드에 존재하는 입출력 데이터를 이용하여 c 개의 클러스터 중심값을 산출한 후, 다음과 같이 입력값을 c 개의 중심값 중에서 소속도가 높은 클러스터로 하위노드 X_i의 입출력 클러스터를 형성한다.

$$\begin{cases} X_i = \{x(k) \mid u_i(x(k)) > u_j(x(k))\},\ all\ i \neq j \\ Y_i = \{y(k) \mid (x(k)) \in X_i\} \end{cases} \quad (7)$$

여기서, U_i는 아래와 같이 상위노드에 있는 데이터와 i번째 하위노드의 중심벡터에 대해 계산되는 소속값을 나타낸다.

$$U_i = [\ u_i(x(1)), u_i(x(2)), \cdots, u_i(x(h))\ ] \quad (8)$$

[단계 4] 각각의 하위노드인 $X_1, X_2, \cdots, X_f$에 존재하는 데이터의 개수 $n_1, n_2, \cdots, n_f$를 계산한 후, 각각의 데이터의 개수 중 하나라도 설정된 개수(S_2) 이하이면 분기를 정지하고 상위노드를 말단의 잎노드(leaf node)로 간주한다. 그렇지 않을 경우 [단계 5]를 실행한다.

[단계 5] 하위노드 중 클러스터 i에 해당하는 입출력 데이터 $\{X_i, Y_i\}$만을 이용하여 [단계 1]에서 계산된 방법과 마찬가지로 실제 출력값과 예측값과의 오차값을 각각 산출한 후, 하위노드에 존재하는 모든 데이터를 이용하여 에러값 E_f를 구한다.

$$E_f = \sqrt{\left(\sum_{i=1}^{c}\sum_{j=1}^{n_i}(\hat{y}_i(j) - y_i(j))^2\right) / \left\{\sum_{i=1}^{c} n_i\right\}} \quad (9)$$

여기서, E_f은 모든 클러스터에 해당되는 데이터들을 이용하여 예측된 출력값과 실제 출력값과의 오차값을 나타낸다.

분기전 상위노드에서 식 (6)에서 계산된 오차값 E_b와 분기 후 모든 하위노드에서 계산된 에러값 E_c간의 차 $\delta = E_b - E_f$를 계산한 후 δ값이 음의 값을 갖거나 아주 적은 값을 갖는 임계값 (S_3) 이하의 값을 가질 경우 분기과정을 정지한다. 즉, δ가 음의 값을 갖는다는 의미는 분기를 하였음에도 불구하고 오차값이 증가함을 의미하고 또한 δ가 임계값 이하로 감소하지 않는다는 의미는 분기를 했음에도 오차 측면에서는 큰 효과가 없음을 의미한다.

[단계 6] 표 1의 분기조건을 만족하는 하위노드를 대상으로 분기를 시작하며, 그 과정은 [단계 1]~ [단계 5] 과정을 반복한다. 단, 트리의 깊이(depth)가 설정된 값 (S_4)를 초과할 경우 분기는 정지한다.

3. 실험 및 결과

3.1 데이터 취득 및 구성

본 연구는 용담댐 저수지를 대상으로 수행되었다. 용담다목적댐은 전주권(전주, 군산, 이리)을 포함한 서해안 지역의 안정적 용수공급과 수자원의 효율적 개발을 목적으로 건설되었다. 비교적 수자원이 풍부한 금강 상류에 건설된 유역 변경식 댐으로써 2001년 준공되었다. 용담댐의 만수위는 EL. 263.5m로서 유역 전체가 고지대에 위치하고 있다. 용담댐 유역은 금강의 최상류에 위치하며, 북위 36°00'~35°35', 동경 127°20'~127°45'와 무주군, 진안군, 장수군을 포함한 충정남도, 전라북도, 경상남도의 경계에 걸쳐있다. 이 유역은 산업화가 덜 이루어진 편이어서, 토지이용의 대부분이 산림이거나 농경지이며 점오염 및 비점오염원의 분포가 많지 않아 시각적으로 보이는 하천의 수질은 대체적으로 맑은 편이다.

용담호 수질거동의 분석을 위해 그림 2에서 보는 바와 같이 2005년 1월부터 2005년 12월까지 저수지 내 10개소에서 유량 및 수질을 측정하였으며, 수질분석방법은 다음과 같다. 수온, EC, pH, DO, Turbidity는 TROLL 9500을 이용하여 측정하였고, 투명도는 세키판(secchi disc)을 이용하여 측정하였으며, DOC는 TOC Analyzer(Phoenix 8000)을 이용하였다. CODcr은 Ultra Low COD Reagent(HACH)를 이용하여 측정하였으며, SS, BOD5, CODMn, 총질소 농도(T-P), 인산염 농도(PO4-P), 총질소 농도(T-N), 암모니아성질소 농도(NH4-N), 질산성질소 농도(NO3-N), 아질산성질소 농도(NO2-N), Chlorophyll-a 농도는 수질오염공정시험법에 준

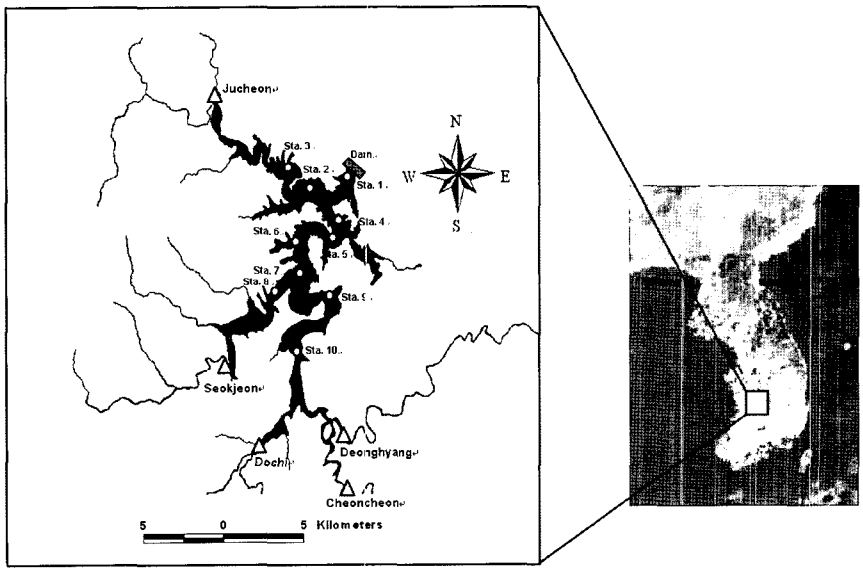


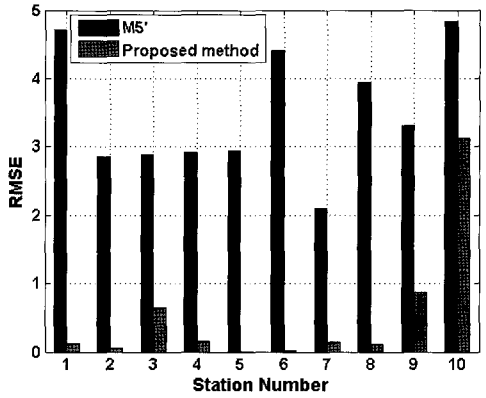
그림 2. 용담댐의 위치 및 데이터 취득 지점 (10 개소)
Fig. 2. Youngdam reservoir and location of sampling station

하여 실시하였다. 표 2에서는 각각의 측정지점별로 관측된 용담호 표층 수질자료의 통계학적 특성을 정리하여 나타냈다. 데이터 분석결과, 수질변화 특성은 1월부터 5월까지는 각각의 측정지점간에 큰 변화를 보이지 않았으나, 강우가 집중하는 6월, 7월에는 급격한 변화 양상을 보이는 것으로 나타났다.

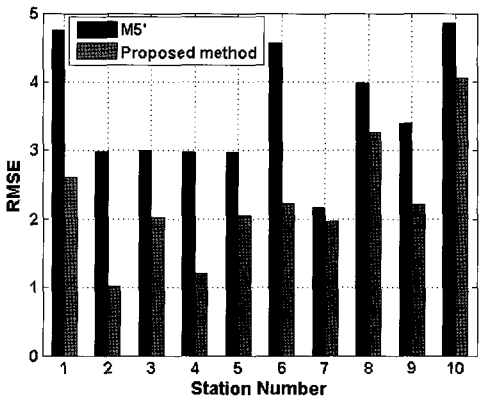
4.2 Station별 클로로필-a 농도 예측모델

데이터를 취득한 Station 별로 클로로필-a 농도 예측모델을 구축하기 위해 11개의 입력 성분(water temperature, DO, pH, Turbidity, SS, BOD, CODcr, TN, TP, PO4-P, Secchi depth)를 이용하였다. 제안된 방법을 평가하기 위하여 LSE에 의해 선형모델을 구축한 M5′ 모델과 비교하였다. 트리를 구축하기 하기 위한 M5′모델에서는 최소 데이터의 개수를 7개로 설정하였고, 평활조건을 선택하였다. 모델의 성능을 평가하기 위하여 훈련데이터와 검증 데이터로 구분하여 실험하였다. 이를 위해서는 실측데이터의 개수가 충분해야 하지만, 일 년 단위로 데이터를 취득하는 수처리 데이터의 경우 충분한 데이터를 얻는 데는 한계가 있다. 특히 본 논문에서는 Station 별로 2005년도의 데이터만을 취득하였기 때문에 제안 방법의 성능을 평가하는데 한계가 있다. 이를 보완하기 위한 한 방법으로서, 본 논문에서는 취득된 데이터를 기준으로 하여 Station 별로 검증 데이터를 생성했다. 즉, Station 별로 취득한 데이터 중 입력성분에 ±2%, 출력성분인 클로로필-a 농도에 ±5% 값을 갖는 10개의 데이터를 생성하여 검증데이터로서 이용하였으며, 훈련데이터는 2005년도의 실측데이터를 이용하여 적용방법을 평가하였다. 제안된 PLS 기반 c-퍼지 클러스터 모델트리는 MATLAB 환경하에서 실험하였으며, PLS 알고리즘은 N-way 툴박스를 이용하였다 [11].

그림 3에서는 성능지표를 RMSE(root mean square error)를 기준으로 하여 각각의 방식을 비교하여 나타냈다. 그림 3에서 보는 바와 훈련데이터와 검증 데이터 모두 M5′ 모델보다 제안된 모델이 우수한 결과를 보였다. 표 3에서는 훈련 데이터에 대하여 M5′ 방식과 제안된 방법의 결과를 여러 가지 척도를 이용하여 분석한 결과를 나타냈다. 모든 성능지표에서 제안된 방법이 우수한 것으로 나타냈으며, 특히 RMSE 값을 고려하며 M5′ 방식은 최소 2.09에서 최대 4.8의 오차값을 나타낸 반면에 제안된 방법은 대부분 0.1이하의 오차를 보여 높은 예측결과를 나타냈다. 표 4에서는 station 별로 얻어진 10개의 검증데이터에 대한 성능비교를 나타냈다. 검증 데이터의 경우에도 모든 성능척도에서 M5′ 모델보다 제안된 PLS 기반 c-FMT 모델이 우수한 결과를 나타냄을 알 수 있다.



(a) 훈련 데이터



(b) 검증데이터

그림 3. 적용 기법별 예측오차
Fig. 3. RMSE calculated by each method
(a) training data (b) testing data

표 2. 용담댐의 수질데이터의 통계적 분석
Table 2. Basic Statistics of the Water Quality in Youngdam Reservoir

	Temp. (℃)	EC (μs)	pH	DO (mg/L)	Turb. (NTU)	SS (mg/L)	BOD (mg/L)	CODMn (mg/L)	CODCr (mg/L)
Min.	2.9	53	6.10	0.4	0.3	0.4	0.5	1.9	3.0
Max.	26.1	161	8.92	16.9	155.5	77.0	6.6	6.5	14.8
Mean	14.9	97	7.23	10.3	12.8	7.6	1.4	3.2	5.6
StdDev.	6.5	27	0.61	3.8	30.0	13.5	0.8	1.1	1.6
	TP (mg/L)	PO4-P (mg/L)	TN (mg/L)	NH4-N (mg/L)	NO3-N (mg/L)	NO2-N (mg/L)	DOC (mg/L)	TOC (mg/L)	Chl-a (mg/m³)
Min.	0.004	0.000	1.16	0.01	0.57	0.003	1.0	1.0	0.8
Max.	0.17	0.042	5.27	0.70	3.18	0.674	3.3	2.2	28.6
Mean	0.03	0.007	1.82	0.08	1.29	0.027	1.5	1.5	4.1
StdDev.	0.03	0.012	0.61	0.11	0.45	0.073	0.4	0.2	4.3

표 3. 훈련 데이터에 대한 성능 비교
Table 3. Comparing proposed method with M5′ for training data

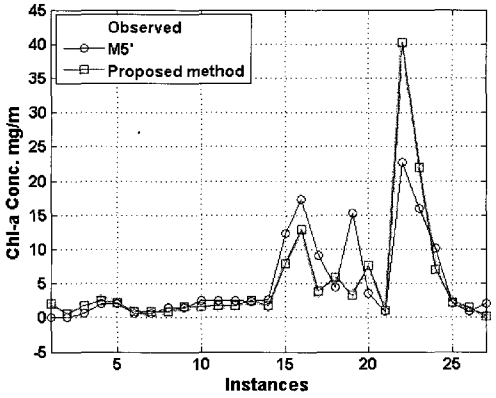
Station	Correlation coefficient		Mean absolute error		Root mean squared error		Relative absolute error(%)		Root relative squared error(%)	
	M5P	Proposed method	M5P	Proposed method	M5P	Proposed method	M5P	Proposed method	M5P	Proposed method
1	0.8253	0.9999	2.5903	0.0775	4.7195	0.1244	51.5922	1.5443	57.0993	1.5051
2	0.9128	1.0000	1.9977	0.0243	2.8529	0.0471	41.5648	0.5054	41.2111	0.6797
3	0.9078	0.9953	1.8358	0.4338	2.8895	0.6444	47.5963	11.2465	43.5333	9.7083
4	0.8338	0.9996	1.9962	0.0832	2.9172	0.1541	56.7298	2.3647	55.4080	2.9265
5	0.8256	1.0000	1.9163	0.0008	2.9496	0.0022	51.3746	0.0211	56.9177	0.0423
6	0.7731	1.0000	2.7371	0.0152	4.4079	0.0269	58.5078	0.3243	64.9161	0.3968
7	0.8959	0.9996	1.5771	0.0734	2.0940	0.1378	40.8529	1.9012	44.5931	2.9337
8	0.9080	0.9999	2.9364	0.0368	3.9437	0.1063	42.6151	0.5334	43.1252	1.1622
9	0.8540	0.9905	2.2667	0.4054	3.3106	0.8659	46.5508	8.3249	52.6565	13.7726
10	0.8227	0.9268	3.2594	2.0385	4.8389	3.1173	57.3736	35.8826	58.2949	37.5539

표 4. 검증 데이터에 대한 성능비교
Table 4. Comparing proposed method with M5′ for testing data

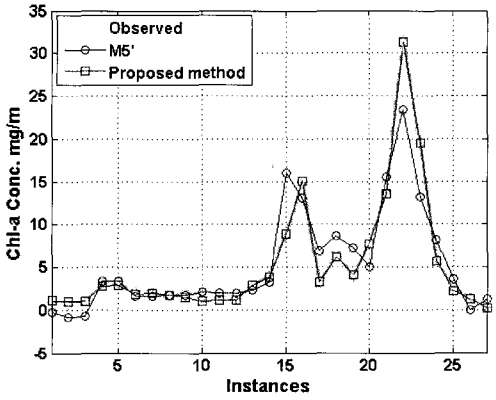
Station	Correlation coefficient		Mean absolute error		Root mean squared error		Relative absolute error(%)		Root relative squared error(%)	
	M5P	Proposed method	M5P	Proposed method	M5P	Proposed method	M5P	Proposed method	M5P	Proposed method
1	0.8204 ±0.018	0.9905 ±0.002	2.9559 ±0.156	1.0228 ±0.117	4.7613 ±0.223	1.1724 ±0.112	59.0397 ±3.980	20.6310 ±2.475	57.7301 ±2.258	14.2315 ±1.417
2	0.9067 ±0.007	0.9921 ±0.002	2.0919 ±0.116	0.7958 ±0.107	2.9847 ±0.171	0.9125 ±0.105	43.4833 ±2.776	16.4555 ±2.251	42.6595 ±1.564	13.0642 ±1.577
3	0.9080 ±0.009	0.9871 ±0.003	1.9540 ±0.106	0.9387 ±0.147	2.9981 ±0.178	1.1131 ±0.142	48.9681 ±3.681	23.4973 ±2.932	44.0733 ±1.752	16.3420 ±1.878
4	0.8262 ±0.013	0.9923 ±0.001	2.1156 ±0.079	0.5690 ±0.053	2.9819 ±0.106	0.6603 ±0.058	59.3746 ±2.370	16.0676 ±1.740	56.4909 ±1.896	12.4783 ±1.145
5	0.8258 ±0.012	0.9944 ±0.001	1.9686 ±0.102	0.5043 ±0.064	2.9770 ±0.118	0.5772 ±0.060	51.4564 ±2.501	13.3168 ±1.515	56.9751 ±1.628	11.0469 ±1.067
6	0.7573 ±0.019	0.9925 ±0.001	2.9072 ±0.102	0.7451 ±0.089	4.5710 ±0.138	0.8574 ±0.071	61.7776 ±2.977	15.8900 ±1.887	66.5690 ±1.745	12.4904 ±1.018
7	0.8904 ±0.009	0.9946 ±0.001	1.6006 ±0.079	0.4345 ±0.061	2.1694 ±0.092	0.5076 ±0.058	41.1405 ±2.333	11.1670 ±1.599	45.7343 ±1.731	10.7046 ±1.233
8	0.9048 ±0.007	0.9955 ±0.001	3.0192 ±0.150	0.7753 ±0.118	3.9809 ±0.142	0.8902 ±0.104	43.6414 ±1.473	11.2578 ±1.710	43.6817 ±1.286	9.7682 ±1.113
9	0.8465 ±0.012	0.9846 ±0.003	2.4052 ±0.125	0.8498 ±0.094	3.4028 ±0.146	1.1168 ±0.102	49.3595 ±2.544	17.4412 ±1.795	53.8130 ±1.805	17.6752 ±1.634
10	0.8204 ±0.010	0.9245 ±0.006	3.3644 ±0.099	2.1412 ±0.111	4.8549 ±0.171	3.1631 ±0.144	58.0917 ±1.898	37.0219 ±1.664	58.5458 ±1.401	38.1558 ±1.548

그림 4에서는 측정 station 1과 2에서 10개의 검증데이터에 대한 클로로필-a 농도의 평균 예측 결과를 나타냈다. 그림 5에서는 측정결과와 예측결과의 분포도를 나타냈다. 그림 4 및 그림 5로부터 예측 훈련데이터와 검증데이터 모두 M5′ 모델보다 제안된 방법이 우수한 것으로 나타났다. 제안된 방법이 기존 방법에 비하여 우 수한 성능을 보임을 정량적으로 분석하기 위하여 그림 6 및 그림 7에 WEKA 프로그램에서 수행한 M5′모델[12]과 제안된 PLS 기반 c-FMT에 의해 발생된 트리구조를 나타냈다. 그림 6의 M5′모델에 의해서 7개의 선형모델이 발생하였다. 말단의 잎노드에서 안정된 선형모델을 얻기 위해서는 각각의 모델에 존재하는 데이터의 개수가 입력의 수보다 충분해야 하지만, M5′에 의해 발생된 선형모델에 남아 있는 데이터의 개수는 12개 이하로 입력차원의 수보다 적은 데이터가 존재한다. 특히, 모델 1인 경우 존재하는 데이터의 개수는 두 개만 존재하므로 모델의 신뢰성이 저하될 우려가 높다. 그림 7에서는 제안된 PLS 기반 c-퍼지 모델트리에 의해 구축된 모델트리를 나타냈다. 그림 7에서 보는 바와 같이 선형 모델에 존재하는 데이터의 개수는 구하고자 하는 계수의 값보다 적은 것으로 나타났다. 이 경우 기존의 LSE방식에 의해 선형모델을 산출할 경우 모델의 성능이 저하될 가능성이 있다. 이러한 문제점을 해결하기 위하여 PLS에 의해 선형모델을 구축함으로써 예측성능의

향상을 가져온 것으로 분석되며, 또한 클러스터링 기반으로 모델 트리를 구축함으로써 기존의 M5′ 모델트리보다 트리의 구조가 간결해 짐을 확인할 수 있었다.



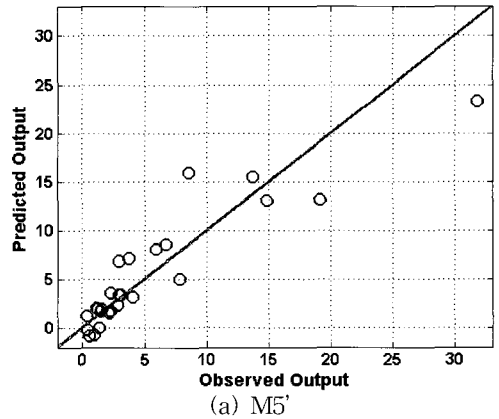
(a) station



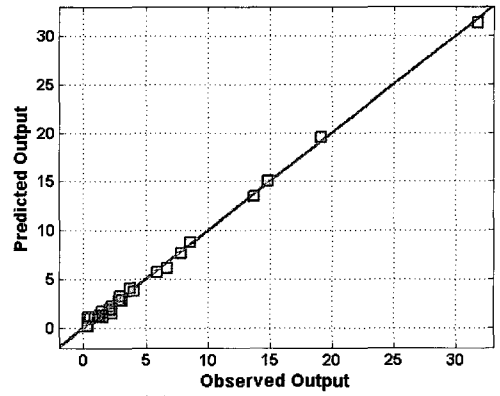
(b) station 2

그림 4. 측정지점 (1)과 (2)에서의 클로로필-a 농도 예측

Fig. 4. Chl-a concentration prediction output for sampling station (1) and (2)



(a) M5'



(b) Proposed method

Fig. 5. 측정지점 2에서의 예측 오차 분포도

그림 5. Distribution of predicted errors for sampling station 2

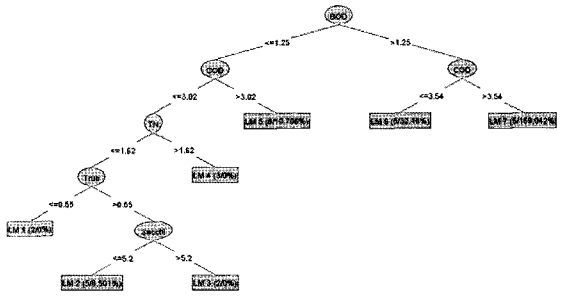


그림 6. 측정 지점 2에서 구축된 M5′ 모델트리

Fig. 6. M5′ model tree for station 2

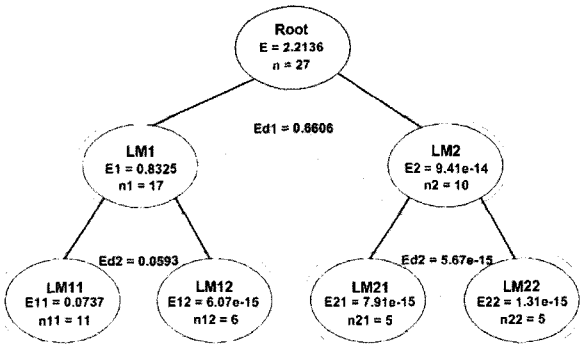


그림 7. 측정 지점 2에서 구축된 PLS기반 c-퍼지 모델트리

Fig. 7. PLS based c-fuzzy model tree for station 2

4. 결 론

본 논문에서는 퍼지 클러스터에 의한 산출된 중심값과 퍼지소속 정보를 기준으로 입출력 데이터를 분할 한 후, 부분최소법에 의해 수질데이터를 예측하는 방법을 제안하였다. 제안된 방법은 하나의 입력속성만을 선택하여 분기하는 기존의 모델트리 방식과 달리 모든 입력속성을 고려한 분기조건을 제안함으로써 예측 오차율을 줄 일 수 있었다. 최종 단계에서는 임의의 입력데이터에 대하여 루트노드로부터 하위노드 까지 각각의 클러스터 중심값 간에 계산된 소속도를 이용

하여 분기점을 추론하면서 각각의 말단에 위치한 잎노드를 탐색한 후 미리 계산된 선형계수값을 이용하여 출력값을 예측한다. 다양한 데이터를 대상으로 실험한 결과 기존의 모델트리 방식보다 향상된 인식 성능을 보임을 알 수 있었다. 특히, 기존의 모델형성을 위해 사용되는 최소자승법 보다 부분최소법을 적용한 결과, 단순한 예측 오차율뿐만 아니라 모든 비교척도에서 제안된 방법이 우수한 결과를 보임을 확인할 수 있었다.

참 고 문 헌

[1] 김좌관, 1995, 수질오염개론, 동화기술, pp.154-178.

[2] 김미숙, 정영륜, 서의훈, 송원섭, "낙동강 부영양화와 수질환경요인의 통계적 분석", Algae, Vol.17(1), pp. 105-115, 2002.

[3] 김호섭, 황순진, "육수학적 특성에 따른 국내 수지의 부영양화 유형분석-엽록소 a와 수심을 중심으로", orean J. Lomnol., Vol. 37(2), pp.213-226, 2004.

[4] Robert G.Wetzel, Limnology-lake and river ecosystems, third edition, Elsevier academic press, 2001.

[5] Quinlan J.R. "Learning with continuous classes" in Proceedings AI'92, Adams & Sterling (Eds.), World Sc -ientific, pp. 343-348, 1992.

[6] Wang Y., Witten I.H., "Inducing Model Trees for Continuous Classes", in Poster Paper of the 9th European Conference on Machine Learning (ECML 97)., M. van Someren, & G. Widmer (Eds.), Prague, Czech Republic, pp. 128-137, 1997.

[7] 이대종, 박진일, 박상영, 정남정, 전명근, "클러스터 기반 퍼지 모델트리를 이용한 데이터 모델링", 한국퍼지 및 지능시스템 학회 2006, Vol. 16, pp. 493-500, No. 5, 2006.

[8] Rasmus Bro, Age. K. Smide, Sijmen de Jong, "On the difference between low-rank and subspace approximation: improved model for multi-linear PLS regression", Chemometics and Intelligent Laboratory Systems, Vol. 58, pp. 3-13, 2001.

[9] Y.P. Zhou, J.H. Jiang, W.Q. Lin, L. Xu, H.L. Wu, G.L. Yu, "Artificial neural network-based partial least-square regression with application to QSAR studies", Talanta, 2006, not printed, available online at www.sciencedirect.com

[10] Y. Tian, L. Shi, W. Tong, G.T. Gene Hwang, C. Wang, "Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assesment of classification models", Computational Biology and Chemistry, Vol. 28, pp. 235-244, 2005.

[11] C. A. Andersson and R. Bro, "The N-way Toolbox for MATLAB", Chemometrics and Intelligent Laboratory Systems, Vol. 52, pp. 1-4, 2000.

[12] Ian H. Witten & Eibe Frank, DATA MININING-Practical machine learning tools and techniques, Morgan Kaufmann publisher, (2005).

저 자 소 개

이대종(Dae Jong Lee)
한국퍼지및지능시스템학회 논문지 제16권 제1호 참조

박상영(Sang Young Park)
1996년 : 충북대학교 도시공학과(학사)
1999년 : 충북대학교 도시공학과(공학석사)
2004년 : 충북대학교 도시공학과(공학박사)
2005년~현재 : 한국수자원공사 수자원연구원 선임연구원

관심분야 : 데이터마이닝, 시계열분석
E-mail : sypark119@kwater.or.kr

정남정(Nahm Chung Jung)
1981년 : 경북대학교 (학사)
1998년 : UNESCO-IHE 위생공학(공학석사)
2003년~현재 : UNESCO-IHE, TU Delft (박사과정)
1985년~현재 : 수자원공사 입사, 상하수도 연구소장

관심분야 : 저수지 수질모델링(물리적 수치모델과 데이터 마이닝)
E-mail : chung@kwater.or.kr

이혜근(Hye Keun Lee)
1977년 : 한양대학교 토목공학과 (학사)
1993년 : 미 플로리다주립대(공학박사)
1994년~현재 : 한국수자원공사 수자원연구원 수석연구원

관심분야 : 수질/수질모델링
E-mail : hklee@kwater.or.kr

박진일(Jin Il Park)
2001년 : 한밭대학교 제어계측공학과(학사)
2003년 : 한밭대학교 제어계측공학과 (공학석사)
2005년~현재 : 충북대 제어계측공학과 박사과정.

관심분야 : 지능시스템, 퍼지이론
E-mail : moralskr@yahoo.co.kr

전명근(Myung Geun Chun)
한국퍼지및지능시스템학회 논문지 제16권 제1호 참조