Proceedings of KFIS Autumn Conference 2006 Vol. 16, No. 2.

PLS 기반 개선된 M5 알고리즘에 의한 수질 예측

The Water Quality Prediction using Improved M5 Algorithm Based on PLS

박진일[1], 이대종[2], 정남정[3], 박상영[3], 전명근[1]

[1] 충북대학교 전기전자컴퓨터 공학부
[2] 충북대학교 BK21 충북정보기술사업단
[3] 한국 수자원공사 수자원연구원
mgchun@chungbuk.ac.kr

요 약

본 논문은 모델트리 알고리즘인 M5에 부분최소법(PLS: Partial Least Square)을 적용하여 클로로필-a 농도의 예측 모델을 제안한다. 제안된 방법은 M5을 이용하여 모델트리를 구축한 후 잎노드에서 PLS를 적용하여 지역모델(local model)을 구축한다. 제안된 방법의 우수성을 보이기 위해 수질 데이터를 대상으로 실험한 결과 기존의 M5 방식에 비하여 향상된 성능을 보임을 알 수 있었다.

Key Words : Model tree, M5, PLS, Data prediction

1. 서 론

결정트리는 의사결정과정을 도표화하여 관심대상 집단을 몇 개의 소집단으로 분류하거나 예측하는 매우 효과적인 데이터 마이닝의 한 기법으로, 모형의 구축과정을 일종의 트리 형태로 표현한다. 결정트리를 형성하는 알고리즘으로 널리 사용되는 것으로 CHAID [1], CART [2], C4.5 [3] 등을 이용하는데 이들은 분리기준과 정지규칙 그리고 가지치기 등에서 서로 다른 형성과정을 가지고 있다. 이러한 결정트리방식은 말단의 잎노드로 부터 결과를 도출할 수 있을 뿐만 아니라 연결된 각 내부노드로부터 결과에 대한 과정을 추적할 수 있어 해석이 용이하고 구현이 비교적 간단한 장점을 지니고 있다. 그러나 분석용 자료에만 의존하기 때문에 새로운 자료의 예측에서는 불안정할 가능성이 높고, 이진분류 알고리즘을 적용한 경우 분리 가지 수가 많아지는 단점이 있다.

일반적으로 예측문제에서는 연속적인 입력변수 및 출력값을 갖는 데이터들이 대부분을 차지한다. 결정트리의 분류인 회귀트리는 말단노드에 위치한 잎노드에 속한 연속적인 출력값의 평균값을 계산함으로 예측력의 저하를 초래한다. 이러한 문제점을 해결하기 위해 모델트리 기반의 다양한 알고리즘이 제안되고 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. 이러한 모델트리도 회귀트리와 같이 데이터를 반복적으로 분리하여 트리구조를 생성하는 상-하 추론 모델트리(TIMT: Top-down Induction of Model Tree) 형식을 갖는다. 이러한 모델트리는 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.

본 논문에서는 M5'에 의하여 모델트리구조를 구축한 후 부분최소법에 의해 모델을 예측하는 PLS 기반 M5' 알고리즘을 제안한다. 부분최소제곱법은 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 [4]. 제안된 방법의 우수성을 보이기 위해 수질 데이터를 대상으로 실험한 결과 기존의 모델트리 방식에 비하여 향상된 성능을 보임을 알 수 있었다.

2. PLS 기반 M5′ 알고리즘

본 논문에서 제안된 방법은 M5에 의하여 모델트리구조를 구축한 후 부분최소법에 의해 모델을 예측하는 PLS 기반 M5 알고리즘으로서, 예측과정은 그림 1에서 보는 바와 같이 미리 구축된 트리 구조에 의해 잎 노드에 존재하는 모델을 선택하여 출력결과를 얻는다.

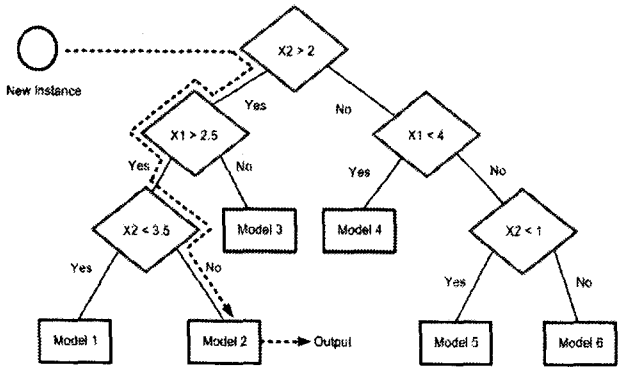


그림 1. PLS 기반 M5′ 알고리즘

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, 6].

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

여기에서 T는 도달한 마디의 예제들의 집합이고, $T_1, T_2, \ldots$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(pruning) 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 부분최소법을 이용하여 출력값을 예측한다.

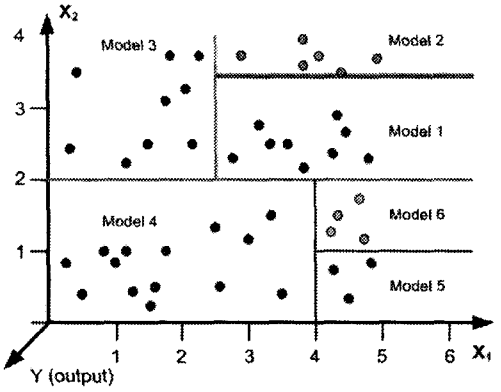


그림 2. 모델트리의 생성

모델트리 방식에서 말단의 잎노드에서 선형모델을 구하기 최소자승법에 기반을 둔 다중선형회귀분석(multiple linear regression)을 이용한다. 그러나 수처리 데이터와 같이 데이터의 개수가 충분하지 못할 경우 트리구조에 의해 입출력 데이터의 분할이 지속될 경우 는 일 년에 한번 입출력 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.

부분최소법은 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 회귀식 측면에서 입력변수만으로 주성분을 만들어 내는 주성분회귀와 비슷한 방법이지만, 주성분회귀에서 주성분은 입력변수의 데이터값만으로 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 부분최소법은 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, T와 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 가중치행렬 $W \in R^{p \times c}$를 이용하여 요인점수행렬 $T = XW$ ($T \in R^{n \times c}$)를 계산한다. 여기서, 가중치행렬 W는 식 (3)과 같이 출력값과 대응되는 요인점수 사이의 공분산이 최대화 되도록 설정한다.

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

3. 실험 및 결과

제안된 방법의 타당성을 검증하기 위하여 금강의 최상류에 위치한 용담댐의 수질데이터를 적용하였다. 용담댐 유역의 수위관측소는 8개소가 설치돼 있으며, 이 중에서 용담댐의 댐수위 관측소와 천천, 동향 등 2개의 하천수위 관측소는 댐 운영을 위한 실시간 통신 관측소이며, 주자천의 주천교, 정자천의 석정교, 진안천의 상도치교, 장계천이 동정교, 구량천의 주고교 등 5개 관측소는 시험유역 운영을 위해 하천에 설치되어 있다. 용담호 수질거동의 분석을 위해 2005년 1월부터 2005년 12월까지 저수지 내 10개소에서 유량 및 수질을 측정하였다. 수질측정항목은 DO, pH, EC, 수온, Secchi-depth, Turbidity, SS, BOD, CODMn, CODCr, Chlorophyll-a, TP, PO4-P, NH4-N, NO3-N, NO2-N, DOC 이다.

데이터를 취득한 station 별로 클로로필-a 농도 예측모델을 구축하기 위해 11개의 입력성분(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.

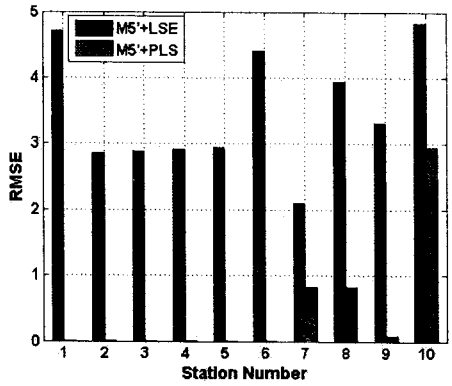
트리를 구축하기 하기 위한 M5 모델에서는 최소 데이터의 개수를 7개로 설정하였고, 평활조건을 선택하였다. 모델의 성능을 평가하기 위하여 훈련 데이터와 검증 데이터로 구분하여 실험하였다. 이를 위해서는 실측 데이터의 개수가 충분해야 하지만, 일 년 단위로 데이터를 취득하는 수처리 데이터의 경우 충분한 데이터를 얻는 데는 한계가 있다. 특히 본 논문에서는 station 별로 2005년도의 데이터만을 취득하였기 때문에 제안 방법의 성능을 평가하는데 한계가 있다. 이를 보완하기 위한 한 방법으로서, 본 논문에서는 취득된 데이터를 기준으로 하여 station 별로 검증 데이터를 생성했다. 즉, Station 별로 취득한 데이터 중 입력성분에 ±2%, 출력성분인 클로로필-a 농도에 ±5% 값을 갖는 10개의 데이터를 생성하여 검증 데이터로서 이용하였으며, 훈련데이터는 2005년도의 실측 데이터를 이용하여 적용방법을 평가하였다. 실험은 MATLAB 환경에서 실험하였으며, PLS 알고리즘은 N-way 툴박스를 이용하였다 [7].

표 1. M5와 제안된 방법의 성능비교

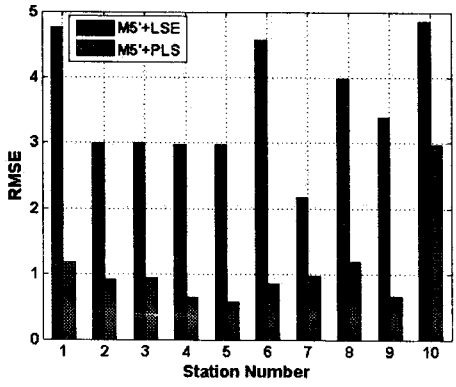
Station	Training data		Test data	
	M5	Proposed method	M5	Proposed method
1	4.7195	4.51e-15	4.7613 ±0.223	1.1708 ±0.105
2	2.8529	4.19e-15	2.9847 ±0.171	0.9087 ±0.103
3	2.8895	2.41e-15	2.9981 ±0.178	0.9405 ±0.101
4	2.9172	1.43e-15	2.9819 ±0.106	0.6441 ±0.059
5	2.9496	7.93e-15	2.9770 ±0.118	0.5772 ±0.060
6	4.4079	3.02e-15	4.5710 ±0.138	0.8584 ±0.072
7	2.0940	0.8316	2.1694 ±0.092	0.9789 ±0.083
8	3.9437	0.8260	3.9809 ±0.142	1.1908 ±0.179
9	3.3106	0.0835	3.4028 ±0.146	0.6607 ±0.044
10	4.8389	2.9447	4.8549 ±0.171	2.9750 ±0.126

그림 3및 표 1에서는 성능지표를 RMSE(Root Mean Square Error)를 기준으로 하여 각각의 방식을 비교하여 나타냈다. 그림 3에서 보는 바와 훈련데이터와 검증 데이터 모두 M5 모델보다 제안된 방법이 우수한 결과를 보였다. 특히 RMSE 값이 M5 방식은 최소 2.09에서 최대 4.8의 오차값을 나타낸 반면에 제안된 방법은 대부분 0.1이하의 오차를 보여 높은 예측결과를 나타냈다.

그림 4에서는 측정 station 1에서 검증데이터에 대한 클로로필-a 농도 예측 결과를 나타냈다. 그림 4에서 보는 바와 같이 LSE 기반의 M5 보다 제안된 PLS 기반의 M5방식이 우수한 것으로 나타났다. 그림 5에서는 실제 출력값과 예측값과의 분포도를 나타냈다. 그림 5에서 보는 바와 같이 제안된 방법이 실제 출력값을 예측하는데 효과적인 것으로 나타났다.



(a) 훈련 데이터



(b) 검증 데이터

그림 3. 적용 기법별 예측 오차

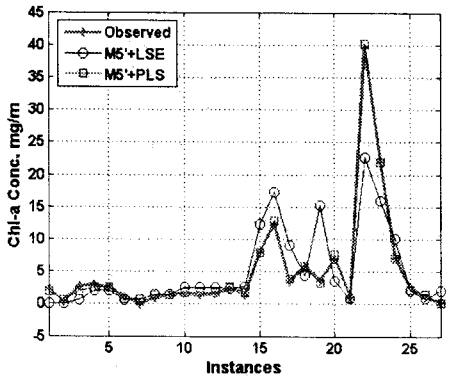
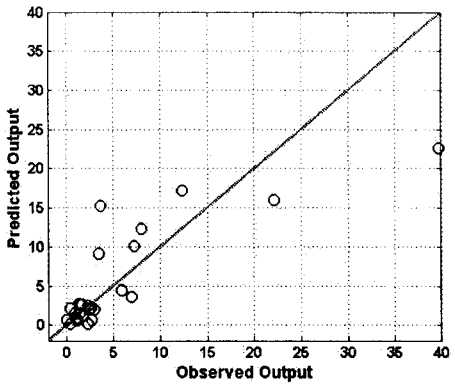


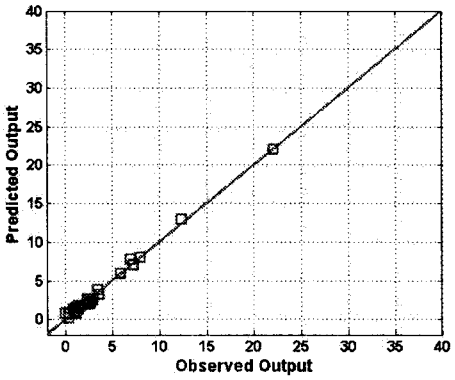
그림 4. 측정지점 1에서의 클로로필-a 농도 예측

4. 결 론

본 논문은 모델트리 알고리즘인 M5에 부분 최소법을 적용하여 클로로필-a 농도의 예측 모델을 제안하였다. 제안된 방법은 M5을 이용하여 모델트리를 구축한 후 잎노드에서 PLS를 적용하여 지역모델을 구축하였다. 제안된 방법의 우수성을 보이기 위해 수질 데이터를 대상으로 실험한 결과 기존의 M5 방식에 비하여 향상된 성능을 보임을 알 수 있었다.



(a) M5 with LSE



(b) M5 with PLS

그림 5. 측정지점 1에서의 예측 오차의 분포도

참 고 문 헌

[1] Kass, Chi-squard Automatic Interaction Detection, Magidson and SPSS inc., 1980.

[2] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., "Classification and regression tree", Belmont CA:Wadsworth, 1984

[3] Quilan, J.R., "C4.5: Programs for machine learning", Morgan Kaufmann, 1993.

[4] 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.

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