Proceedings of KFIS Spring Conference 2004 Volume 14, Number 1

퍼지 클러스터링과 결정 트리를 이용한 모델기반 오존 예보 시스템

Model-based Ozone Forecasting System using Fuzzy Clustering and Decision tree

천성표, 이미희, 이상혁, 김성신
부산대학교 전기공학과

Seong-Pyo Cheon, MiHee Lee, Sang-Hyuk Lee and Sung-Shin Kim
School of Electrical Engineering, Pusan National University
30 Jangjeon-dong, Geumjeong-gu, Busan 609-735, Korea
E-mail : buzz74@pusan.ac.kr

요 약

오존 반응 메카니즘은 상당히 복잡하고 비선형적이기 때문에 오존 농도를 예측하는 것은 상당한 어려움을 안고 있다. 따라서, 신뢰성 높은 오존 예측값을 구하는데 단일 예측모델만으로는 한계가 있으며, 이를 개선하기 위하여 다중 모델을 제안하였다. 입력데이터에 퍼지 클러스터링을 사용하여 고, 중, 저농도별로 그룹핑한 후, 그룹핑된 오존농도에 대해서 의사결정 트리를 사용하여 그룹핑된 오존데이터가 어느 정도 분류능력을 갖는지 파악하여, 오차가 가장 적은 분류특성을 갖는 그룹을 설정하여, 다중모델의 입력 데이터로 사용하여 모델을 형성하였다. 의사결정 트리를 이용하여 모델의 입력 데이터를 설정하는 것은 어떤 오존농도까지의 범위를 클래스로 설정하느냐에 따라서 모델의 성능과 고, 중, 저농도의 오존을 분류하는 성능이 달라지므로 본 논문에서는 퍼지 클러스터링을 이용하여 의사결정 트리의 클래스의 범위를 설정하여 예측 시스템을 구현하였다.

1. 서론

산업화에 따른 지표면의 오존 발생량이 증가하면서, 고농도 오존발생에 대한 관심이 증가하고 있다. 지표면의 오존 농도가 기준치를 초과하게 되면 인체에 해로운 영향을 미치기 때문에 고농도 오존 발생으로 인한 피해를 줄이기 위해 오존 예보의 필요성이 대두되고 있다. 하지만, 기상학적인 변수와 오존 오염물질간의 강한 비선형성과 대류권내에서의 오존생성에 관한 매우 복잡한 반응기 동작으로 인하여 고농도 오존 예측에 있어서는 많은 어려움이 따른다.

오존예측에 있어서 통계적인 선형회귀방법 등은 학습자료로의 편중과 과도한 계산시간 등의 문제점을 가지고 있으며, 비선형성을 내포하고 있는 복잡한 시스템의 모델을 구성하고, 시스템의 출력을 예측하기에는 여러 가지 제한들이 존재하며, 예측오차도 크다.

본 논문에서는 정확하고 신뢰할 수 있는 오존농도 예측을 위해 다중 모델을 형성하였다. 다중모델은 입력데이터의 특성을 판단하여, 데이터의 특성에 맞는 모델을 통해서 예측하였을 때 만족스러운 결과를 얻을 수 있다. 입력데이터의 특성 판단을 위해 퍼지 c-Means 클러스터링(Fuzzy c-Means Clustering)[8]방법과 의사결정 트리(Decision tree)[4-6]를 이용하였고, 모델은 입력과 출력사이에 존재하는 비선형성을 적극 반영하고 모델링에 적절한 입력요소를 선택하기 위해 동적 다항식 신경회로망(DPNN: Dynamic Polynomial Neural Networks)[1-3]을 구성하여 예측을 수행하는 시스템을 구현하고, 이를 평가했다.

2. 퍼지 클러스터링과 의사 결정 트리 그리고 동적 다항식 신경망

2.1 퍼지 클러스터링

퍼지 c-means 클러스터링은 입출력 공간상의 데이터(x_j)에 대한 미지의 중심(v_i)를 구하기 위해 퍼지 이론을 적용하여 각각의 데이터에 대한 소속도(u_{ij})를 나타내고, 이를 통해서 식(1) 같은 목적 함수(J)를 최소화하여 효율적으로 시스템의 특성을 알 수 있는 기법이다.

$$J(u_{ij},v_k,x_j)=\sum_{i=1}^{C}\sum_{j=1}^{n}(u_{ij}^{m})\,|x_j-v_i|^2,\ m>1 \qquad (1)$$

이때, m은 소속도 함수(u_{ij})의 퍼지화 정도를 나타내는 지수형 계수이다. 목적 함수의 최소화 문제를 풀기 위한 미지의 중심(v_i)과 소속도 함수값은(u_{ij})는 각각 식(2), 식(3)과 같다.

$$v_i=\frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m},\ i=1,2,\cdots,c \qquad (2)$$

$$u_{ij}=\frac{(1/|x_j-v_i|^2)^{1/m-1}}{\sum_{k=1}^{c}(1/|x_j-v_k|^2)^{1/m-1}} \qquad (3)$$

$$i=1,2,\cdots,c;\quad j=1,2,\cdots,n.$$

퍼지 c-Means 클러스터링은 아래 4단계의 과정을 거쳐 이루어진다.

1)선택할 군(Cluster)의 개수($2\le c\le n$)와 계수$m(1<m<\infty)$을 정하여 퍼지 멤버쉽 함수$U^{(l)}$의 분할 행렬을 구한다.

2) 구해진$U^{(l)}$와 식(2)을 이용하여 새로운 중심을 구한다.

3) 구해진 새로운 중심과 식(3)을 이용하여 새로운 $U^{(l+1)}$ 행렬을 구성한다.

4)$\triangle=|U^{(l+1)}-U^{(l)}|=max\,u_{ij}^{(l+1)}-u_{ij}^{(l)}$를 계산하여 만약$\triangle\le\epsilon$(한계 범위)이면, l을 1 증가시켜 2단계과정을 반복 수행하고, $\triangle\le\epsilon$이면 클러스터링 과정을 멈춘다.

2.2 의사결정 트리

의사결정 트리(Decision Tree)는 분류 또는 예측을 목적으로 귀납적 추론방법에 있어서 널리 사용되고 있는 학습방법론이다. 본 논문에서는 의사결정나무 알고리즘 중 하나인 C4.5 의사결정 알고리즘을 사용하여 의사결정 트리를 형성하여 저, 중, 고농도 모델의 입력데이터를 분류를 형성하고 있다. C4.5알고리즘은 ID3를 통해서 발전되었으며, Quinlan에 의해 주창되었다. 각 마디(node)에서의 불순도(Impurity)를 재는 척도인 엔트로피 지수(Entropy Index)를 분리 기준으로 사용하고 있다. 어떤 class에 속하는지의 예상정보(expected information)는 식(4)로 표현되고 이는 집합 S에 대한 엔트로피(Entropy)를 나타내는 것이다.

$$info(S)=-\sum_{j=1}^{k}\frac{freq(C_j,S)}{|S|}\times log_2\left(\frac{freq(C_j,S)}{|S|}\right)bits \qquad (4)$$

위의 Expected Information에 대해 training case에 적용시키면 $info(T)$로 놓을 수 있다. T node에서 test X를 node로 분할하려고 할 때, 예상되는 information은 식(5)로 표현된다.

$$info_X(T)=-\sum_{i=1}^{n}\frac{|T_i|}{|T|}\times info(T_i) \qquad (5)$$

따라서, T에서 분할하면서 얻어지는 이득(Gain)은 식(6)과 같다.

$$gain(X)=info(T)-info_X(T) \qquad (6)$$

information gain을 최대화하는 test를 선택한다. ID3에서의 test의 선택은 위의 gain criterion을 기초로 하여 만들어졌다. 하지만 gain criterion은 결과(outcomes)가 많으면 강한 바이어스(bias)가 존재한다. 그리고 gain criterion에 존재하는 bais는 식(7)과 같이 정규화를 사용하여 gain ratio로 수정될 수 있다.

$$splitinfo(X)=-\sum_{i=1}^{n}\frac{|T_i|}{|T|}\times\log_2\left(\frac{|T_i|}{|T|}\right) \qquad (7)$$

정리하면, 식(8)과 같이 분리에 의해 생성되는 information의 비율로 나타나게 되며, 이 gain ratio의 값이 가장 크게 하는 test를 선택하여 분류를 실시하게 된다.

$$gain\,ratio(X)=gain(X)/splitinfo(X) \qquad (8)$$

2.3 동적 다항식 신경망

다량의 관측자료와 변수들로부터 시스템의 모델을 구성하기 위해 GMDH(Group Method of Data Handling)를 이용한 다항식 신경회로망은 비선형적이고 복잡한 동적인 시스템의 모델링과 예측 및 지능제어에 응용되어지고 있다. 각 노드에 대해서 두 개의 입력변수로부터 하나의 출력을 생성하는 일반적인 형태의 다항식 신경회로망의 구조를 그림 1에 나타내었으며 다항식 신경회로망의 각 노드에서 입력변수와 출력과의 관계는 식(9), 식(10)으로 표현할 수 있다.

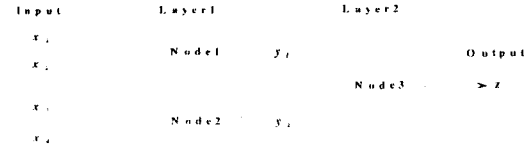

그림 1. 동적 다항식 신경망의 기본 구조

$$y_1 = f_1(x_1, x_2) = w_{01} + w_{11}x_1 + w_{21}x_2 + w_{31}x_1x_2 \quad w_{41}x_1^2 + w_{51}x_2^2 \tag{9}$$

$$y_2 = f_1(x_3, x_4) = w_{02} + w_{12}x_3 + w_{22}x_3 + w_{32}x_3x_4 \quad w_{42}x_3^2 + w_{52}x_4^2 \tag{10}$$

최종 출력 z는 각 노드의 출력 값 y_1, y_2의 다항식으로 식(11)로 나타내어진다.

$$z_1 = f_1(y_1, y_2) = w_{03} + w_{13}y_1 + w_{23}y_2 + w_{33}y_1y_2 \quad w_{42}y_1^2 + w_{52}y_2^2 \tag{11}$$

단, $w_{ij}(i = 0, 1, 2, \cdots, n; j = 0, 1, 2, \cdots, k)$이다.

각 노드의 파라미터를 결정하기 위해 최소 자승법을 사용하며, 목적함수인 실제 측정된 값과 학습된 출력된 값과의 차이를 최소화하는 파라미터를 식(12)로 구하게 된다.

$$J = \sum_{k=1}^{\#ofdata} (z(k) - \hat{z}(k))^2 = \| z - \Phi w \|^2 \tag{12}$$

$$w = (\Phi^T\Phi)^{-1}\Phi^T z$$

즉, J를 최소화하는 w를 구하는 것이 목적이다. 과도 학습 방지와 입·출력의 안정도 평가, 학습 종결 시점을 정하기 위해 식(13)의 PC(Performance Criterion)를 사용하였다.

$$PC = e_1^2 + e_2^2 + \eta(e_1^2 - e_2^2)^2 \tag{13}$$

여기서, e_1^2는 학습오차, e_2^2는 테스트 오차를 나타내며, 최적 모델은 PC가 최소인 곳이다.

3. 오존 예보시스템

3.1 다중모델 구성

오존예보시스템의 다중모델 구성은 그림 2과 같이 구성되어 있다. 서울지역 27개 대기 측정소, 기상청, MM5 모델로부터 데이터를 받아서 퍼지 클러스터링을 이용하여 클래스를 형성한 후, C4.5 의사결정 트리과정을 통하여 선정된 클래스에 대한 입력데이터를 분류하여 다중모델을 형성한다.

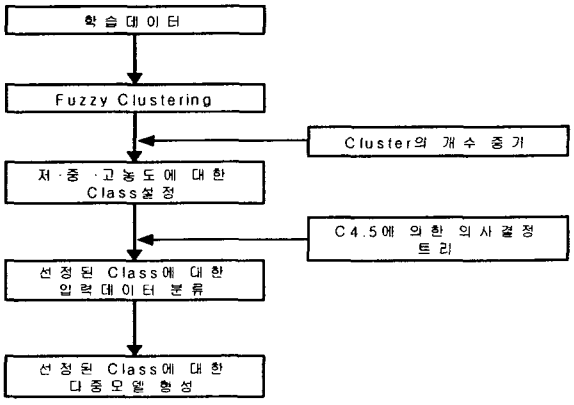

그림 2. 다중모델 형성 과정

4. 권역별 오존 예측 및 평가

4.1 권역별 고농도 오존 예측 및 평가

실험은 2003년 5월 1일부터 19일(5월 8일 데이터 누락)중에서 하루를 제외한 18일간 예측하였으며, 다중 모델과 중회귀모델을 이용하여 서울시 각 권역별 오존의 최고 농도 예측결과이다.

그림 3에서부터 그림 6은 5월1일부터 19일까지의 실제 오존농도값과 각 권역별로 예측모델을 이용하여 예측결과와 중회귀모델을 이용하여 예측한 결과를 비교한 그림이다. 표 1은 실제 오존농도값과 예측값의 RMSE값을 산출한 결과값이다. 결과에서 보듯이 예측모델은 기존에 개발되었던 중회귀모델보다는 각 권역의 예측결과의 RMSE값이 작고, 전체지역에 대해서도 RMSE가 예측모델은 15.0074이고 중회귀 모델은 23.7095로 예측모델의 예측성능이 더 우수함을 알 수 있었다. 표 1에서 북서지역에 대한 예측모델의 성능이 다른 지역에 비하여 성능 낮음을 알 수 있다. 이는 오존농도 측정소 27개지역 가운데 본 모델에서는 17개 지역에 대한 예측을 실행하고 있다. 특히, 북서지역에 대한 오존농도 측정소 6개 지역 중에서 최고 오존농도 빈도수가 많은 지역은 불광동 지역인데 본 예측모델은 불광동이 제외된 3개 지역에 대해 예측을 실행하므로 불광동에 대한 예측오차를 극복하지 못함이기 때문이다.

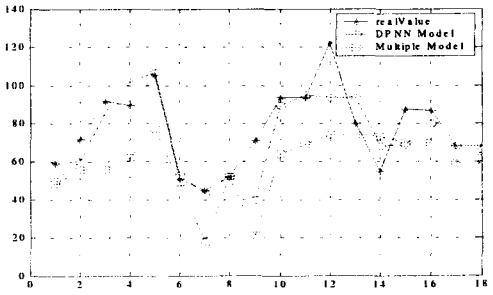

그림 3 북서지역 실측과 예측결과 비교

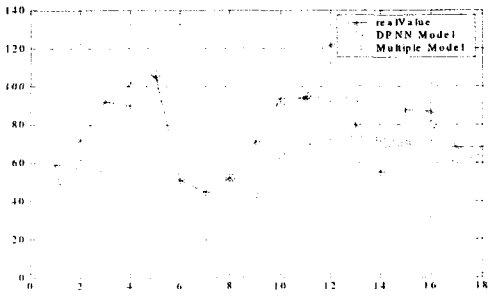

그림 4 북동지역 실측과 예측결과 비교

표 1 예측 모델과 중회귀 모델의 RMSE

지 역	RMSE	
	예측 모델	중회귀 모델
북서지역	18.6905	27.1968
북동지역	13.0192	25.1915
남서지역	15.3406	24.7117
남동지역	12.1129	16.2361
전체지역	15.0074	23.7095

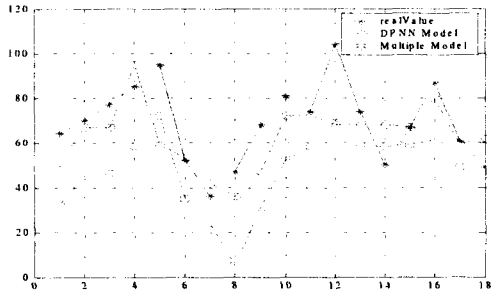

그림 5 남서지역 실측과 예측결과 비교

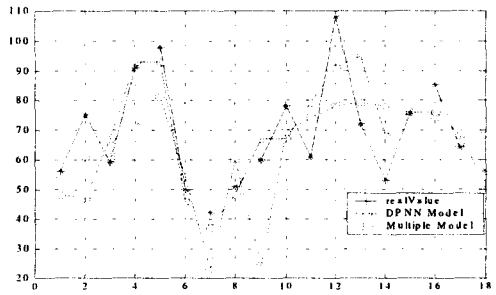

그림 6 남동지역 실측과 예측결과 비교

5. 결론

복잡하고 비선형적인 거동을 예측하는 것은 결코 쉬운일이 아니다. 고농도 오존 예측 역시 기존의 통계적 방법의 한계를 극복하기 위한 시도로 보다 비선형성을 잘 반영해 줄 수 있는 예측 모델로 동적 다항식 신경회로망으로 구현하였으며, 예측의 정확도와 신뢰성을 향상하기 위해, 퍼지 클러스터링과 의사결정 트리기법을 도입해서 다중 모델로 구현하고, 이를 평가해 보았다. 하지만, 이러한 시도 역시 여러 가지 해결해야할 문제점을 안고 있다. 대표적인 것을 지적하면, 예측의 근간이 되는 데이터의 양과 신뢰성을 들 수 있다. 아직까지 고농도의 오존 발생 빈도가 적고, 발생전후의 상관관계를 파악할 수 있는 데이터 역시 부족하다. 또한, 데이터에 대한 신뢰성 평가 역시 이루어지지 않고 있으므로, 향후 보다 정확한 예측을 위해서는 고농도 데이터의 축적 및 신뢰성 평가가 기본적으로 필요하다고 생각된다. 본 논문의 결과 역시 기존의 단편적인 중회귀 분석법보다 성능이 향상된 모델을 제안하였다는 것에 만족하지 않고, 더욱 향상된 성능을 나타내는 모델 및 예측 시스템을 구현하는데 노력하겠다.

6. 참고문헌

[1] A. G. Ivahnenko. "Polynomial theory of complex system," IEEE trans. Syst. Man and Cytern, vol. SMC-12, pp.364-378, 1971.

[2] Duc Trung Pham and Liu Xing, *Neural Networks for Identification, Prediction and control*, Springer-Verlag Inc., 1995.

[3] A. G. Ivakhnenko, "The Group Method of Data Handling in Prediction Problem," *Soviet Automatic Control*, vol. 9, no. 6, pp. 21-30, 1976.

[4] J.R. Quinlan, "Introduction of Decision Trees," *Machine Learning*, 1, pp.81-106, 1986.

[5] Tom M. Mitchell, *Machine Learning*, WCB McGraw-Hill, 1997.

[6] Michael J. A. Berry, Gordon Linoff, *Data Mining Techniques*, Wiley-Interscience, 1997.

[7] 김재용, 김성신, 황보현. "퍼지 클러스터링을 이용한 고농도 오존 예측", 퍼지 및 지능시스템학회 논문지. vol. 11, No. 4, pp. 336-339, 2001.

[8] J. S. R. Jang, C. T. Sun and E. Mizutani, *Neuro-Fuzzy and Soft Computing*, Prentice-Hall Inc., 1997.