퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 4, pp. 336-339

퍼지 클러스터링을 이용한 고농도오존예측

Forecasting High-Level Ozone Concentration with Fuzzy Clustering

김재용 · 김성신 · 왕보현*
Jae-yong Kim, Sung-shin Kim and Bo-hyeun Wang

부산대학교 전기전자통신공학부
* 강릉대학교 전자공학과

요 약

오존농도 메커니즘은 매우 복잡하고, 비선형성과 비정상성이 강하기 때문에 오존 예보시스템들은 많은 문제점을 가지고 있다. 특히 고농도 오존에 있어서 예측결과들이 성능이 좋지 않다. 본 논문은 뉴로-퍼지기법과 퍼지 클러스터링을 이용한 오존 예측시스템의 모델링 방법을 설명하고자 한다. GMDH의 전형적인 알고리즘에 기초한 동적 다항식 신경망은 데이터 분석, 비선형적이고 복잡한 시스템의 검증 그리고 동적 시스템의 예측을 위한 유용한 방법이다.

Abstract

The ozone forecasting systems have many problems because the mechanism of the ozone concentration is highly complex, nonlinear, and nonstationary. Especially, the performance of the prediction results in the high-level ozone concentration are not good. This paper describes the modeling method of the ozone prediction system using neuro-fuzzy approaches and fuzzy clustering methods. The dynamic polynomial neural network (DPNN) based upon a typical algorithm of GMDH (group method of data handling) is a useful method for data analysis, the identification of nonlinear complex systems, and prediction of dynamical systems.

Key Words : Clustering, polynomial neural network, GMDH, prediction decision support system

1. 서 론

20세기 들어 급속하게 발달된 산업화와 도시화로 인한 환경오염, 그 중에서도 대기오염은 공장 및 사업장의 산업활동으로 인한 오염물질의 배출로 인해서 점점 가중되어 왔다. 이런 산업의 발전과 기상 변화에 따른 대기중의 오존 농도 메커니즘은 질소산화물(NOx) 및 탄화수소(HC)류 등의 오염물질로 인한 광화학적인 작용과 일사량, 풍속, 기온 등의 기상학적인 변수들의 상호작용으로 생성되어 최근 국내외를 막론하고 하계중 6월부터 8월 사이에 집중적인 고농도 현상을 보이는 것에 관심을 가지고 있다. 이런 오존은 강력한 산화력으로 인해 적당량의 경우 인간에게 이로우나 기준농도 이상의 경우에는 인체의 건강이나 농작물의 수확 및 생태계에 피해를 주고 있다. 우리 나라에서는 산업화의 영향으로 각종 오염물질의 배출이 증가함에 따라 지표오존의 고농도현상의 발생이 해마다 빈번해지고 있으며, 이런 고농도 발생으로 인한 오존의 피해를 줄이기 위해 대기오염 예보제 실시 및 정확한 고농도 오존예측에 관한 연구가 진행되고 있다.

기존의 고농도 오존예측을 위한 방법으로 통계적 방법에 의한 선형 회귀 모델 그리고 다변량 통계분석에 의한 방법과 신경회로망을 이용한 예측모델에 의한 연구가 이루어져 왔으나, 기상학적인 변수와 오존 오염물질간의 강한 비선형성과 대류권내에서의 오존생성에 관한 매우 복잡한 반응기의 동작으로 인하여 고농도 오존 예측에 있어서는 정확한 모형화에 많은 문제들을 가지고 있으며, 대부분의 예측 결과가 실제 오존농도보다 좋은 성능을 보이지 못하고 있다 [1], [2]. 따라서, 본 논문에서는 퍼지 클러스터링과 동적 다항식 신경회로망 (dynamic polynomial neural network, DPNN) 구조를 이용하여 의사결정 구조에 의한 오존 농도 예측 시뮬레이션을 통하여 서울의 19개 지역에 대해서 수행하였다 [3]. 또한, PC (performance criterion)를 사용하여 모델링 과정 중에 발생할 수 있는 학습 데이터에 대한 외존 영향을 감소시키기 위해서 방법으로 사용되었다.

접수일자 : 2001년 05월 19일
완료일자 : 2001년 07월 31일

2. 퍼지 클러스터링

본 논문에서는 오존농도를 저농도와 고농도로 분류하고 퍼지 분할공간을 통하여 입력 변수들에 미치는 분류 정도를 알기 위해 Bezdek의 FCM(fuzzy c-means)을 사용하였다 [3]. 일반적으로 퍼지 클러스터링은 목적 함수 J_m을 최소화하는데 목적이 있다.

$$J_m(U,v)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m\|x_k-v_i\|^2 \qquad (1)$$

우선 목적 함수에서 $m(1<m<\infty)$와 클러스터의 개수 c(2≤c<n)를 결정한다. 본 논문에서는 m=1, c=2로 선정하였다. 클러스터의 중심벡터 $v_i^{(l)}$을 $U^{(l)}$과 식 (2)으로 계산한다.

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n}(\mu_{ik})^m \mathbf{x}_k}{\sum_{k=1}^{n}(\mu_{ik})^m} \quad (2)$$

다음으로 클러스터의 중심과 데이터와의 거리를 계산한다. 식 (3)와 $v_i^{(l)}$ 을 사용하여 $U^{(l)}$ 을 업데이트시킨다.

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{||x_k - v_i||^2}{||x_{k-v_j}||^2}\right)^{\frac{1}{m-1}}} \quad (3)$$

$||U^{(l+1)} - U^{(l)}|| \le \varepsilon_L$의 조건이 만족하면 클러스터링을 종결하고, 조건을 만족하지 않으면 반복하게 된다.

3. 다항식 신경회로망

3.1 다항식 신경회로망의 기본구조

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 [4], [5]. 각 노드에 대해서 두 개의 입력변수로부터 하나의 출력을 생성하는 일반적인 형태의 다항식 신경회로망의 구조를 그림 1에 나타내었으며 다항식 신경회로망의 각 노드에서 입력변수와 출력과의 관계는 다음과 같은 함수식으로 표현할 수 있다.

각 노드에서의 출력 y_1, y_2 는

$$y_1 = w_{01} + w_{11}x_1 + w_{21}x_2 + w_{31}x_1x_2 + w_{41}x_1^2 + w_{51}x_2^2 \quad (4)$$

$$y_2 = w_{02} + w_{12}x_3 + w_{22}x_4 + w_{32}x_3x_4 + w_{42}x_3^2 + w_{52}x_4^2 \quad (5)$$

와 같은 형태의 다항식으로 나타내어지며, 최종 출력 z는 각 노드의 출력 값 y_1, y_2 의 다항식으로 나타내어진다.

$$z = w_{03} + w_{13}y_1 + w_{23}y_2 + w_{33}y_1y_2 + w_{43}y_1^2 + w_{53}y_2^2 \quad (6)$$

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

각 노드의 파라미터를 결정하기 위해 최소자승법을 사용하며 목적함수인 실제 측정된 값과 학습된 출력된 값과의 차이를 최소화하는 파라미터를 구하게 된다.

$$J = \sum_{k=1}^{n}(z(k) - \hat{z}(k))^2 = ||\mathbf{z} - \boldsymbol{\Phi}\boldsymbol{w}||^2 \quad (7)$$

$$\boldsymbol{w} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{z} \quad (8)$$

이렇게 해서 구해진 파라미터가 현재노드의 출력함수를 구성하고, 다음 층 노드의 입력 값으로 들어가게 되며 이러한 과정을 반복하여 최종적으로 우수한 성능을 보이는 출력 값을 나타내는 함수를 얻을 수 있다.

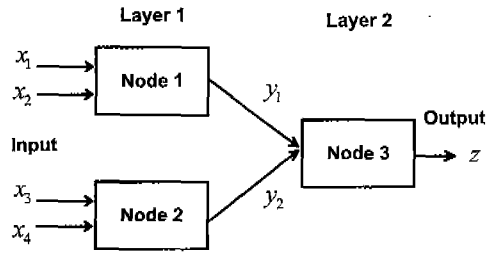

그림 1. 다항식 신경회로망의 기본구조.
Fig 1. The basic structure of DPNN.

3.2 Self Organization

다항식 신경회로망의 또 다른 특징으로서 자기 조직화하는 능력을 들 수 있는 데, 다항식 신경회로망은 GMDH방법을 사용하여 입력 자료를 분산에 따라 학습 데이터와 테스터 데이터로 나누어서 시스템을 모델링한다.

두 개로 나누어진 데이터집합을 이용하여 모델의 구조와 각 노드에서의 파라미터를 구하게 되는데, 일반적으로 학습 데이터를 이용하여 파라미터를 구하고, 테스터 데이터를 이용하여 모델의 성능을 평가하여 학습 오차와 테스터 오차의 관계를 이용하여 모델의 최종적인 구조를 결정하게 된다. 따라서 다항식 신경회로망은 각 층에서 선택기준에 의해 다음 층의 입력을 선택하고, 그림 2에 나타낸 것처럼 학습 오차와 테스터 오차를 이용한 성능기준에 따라서 과도학습을 방지하고, 모델의 입력변수와 구조를 스스로 결정한다.

식 (9)에 학습 오차와 테스터 오차를 이용하여 PC를 결정하기 위한 식을 나타내었으며, η는 0 에서 1 사이의 값을 사용하였다. 식 (9)에 의한 성능평가방법은 구성된 모델이 학습 데이터 뿐만 아니라 테스터 데이터에도 잘 적응하는가를 평가하며, 준비되지 않은 새로운 데이터에도 적용되는 모델평가방법이다.

$$\begin{aligned} e_1^2 &= \sum_{i=1}^{n_A}(y_i^A - f_A(x_i^A))^2/n_A, \\ e_2^2 &= \sum_{i=1}^{n_B}(y_i^B - f_A(x_i^B))^2/n_B, \\ PC &= e_1^2 + e_2^2 + \eta(e_1^2 - e_2^2)^2 \end{aligned} \quad (9)$$

여기서 e1은 학습오차, e2는 테스트 오차, nA는 학습 데이터의 개수, nB는 테스트 데이터의 개수, y_i는 실제 측정된 값, $f_A(x_i^A)$와 $f_B(x_i^B)$는 학습에 의한 결과 값과 테스트에 의한 결과 값이다. 그리고 전체 데이터의 개수는 n=nA+nB이다. 따라서 최적의 모델은 학습 오차와 테스터 오차에 대해 PC가 최소화되는 곳에서 얻어진다.

4. 시뮬레이션

일반적으로 오존의 농도예측을 위해 사용되는 대기오염자료로서는 O_3, CO. NO_2, SO_2, 부유먼지등이 사용되며, 기상자료로서 풍속, 풍향, 기온, 일사량, 강수량, 습도 및 운량 등이 사용된다. 이러한 자료 중에서 100PPB 이상의 고농도 오존에서의 강수량은 대부분 0mm이기 때문에 고농도 오존예측에 거의 영향을 미치지 못하며, 또한 풍향과 운량은 수치화의 어려움, 그리고 아황산가스 와 부유먼지는 대기오염규제로 인해 그 배출량이 줄어들고 있기 때문에 입력자료에서

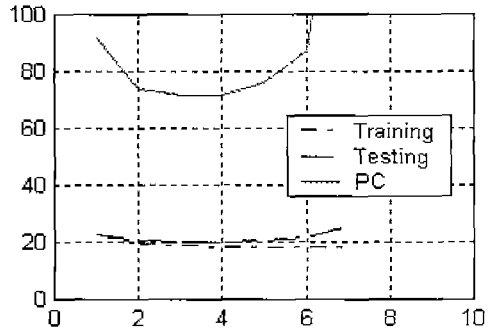

그림 2. 층수의 증가에 따른 모델성능의 변화.
Fig 2. The variation of PC according to the change of layer.

표 1. 입력 변수들과 데이터의 구성
Table 1. Prospective input variables and data structure for forecasting system.

		O_3	NO_2	CO	Rh	Sr	Ws	O_3 Max	Tair Max	O_3
		6	6	6	14	14	14	전날	전날	14
	245	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		7	7	7	15	15	15	전날	전날	15
	245	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
980		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		8	8	8	16	16	16	전날	전날	16
	245	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		9	9	9	17	17	17	전날	전날	17
	245	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

*Tair: 대기온도, Rh:상대습도, Sr:일사량, Ws:풍속, O_3 Max: 전일 최고오존농도, Tair Max:전일 최고온도

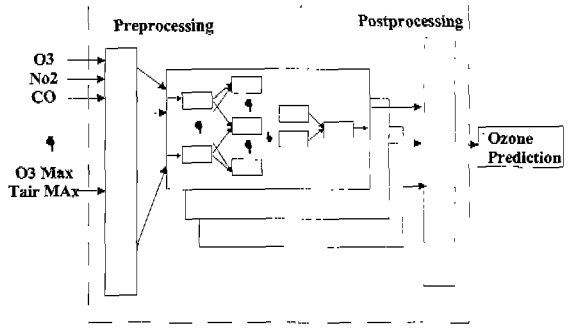

그림 3. 오존 예측시스템의 전체 구조.
Fig 3. The total structure of the ozone prediction system.

삭제하였다. 따라서 학습될 변수로 O_3, CO. NO_2, 기온, 일사량, 습도, 풍속, 전일 최고오존농도, 전일 최고 기온을 사용하였다. 본 논문에서는 하루 중 14~16시경에 주로 최고농도를 나타내기 때문에 이 시간대의 오존농도를 예측하는 것을 목표로 하였다. 특히, 서울지역의 하계에 발생하는 고농도 오존을 8시간 전에 예측(기간: 1997년 8월~9월. 14시 예측)하기 위해서 기본 학습자료를 1996년 5월~9월, 1997년 5월~7월(980개)을 사용하였고, 결측 데이터는 spline 보간 하여 사용하였다. 표1은 입력변수와 실제 학습 데이터들의 구성을 나타냈었다. 그림3은 오존예측시스템의 전체적인 구조를 보여주고 있다. 본 시뮬레이션에서는 클러스터의 개수를 2에서부터 시작하여 4일 때까지를 기준으로 하였다. 기본적으로는 고농도를 최고로 높은 것을 사용하고, 저농도를 나머지의 집합으로 구성하였다.

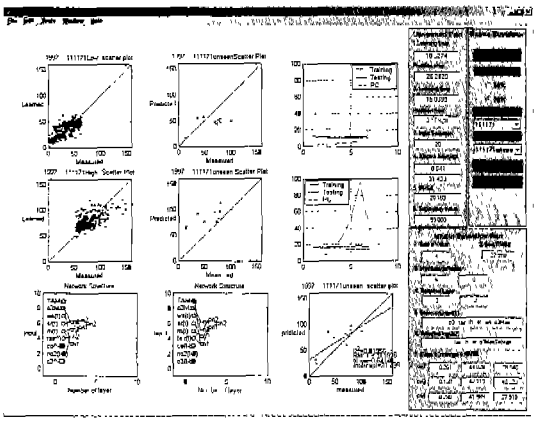

그림 4. 의사결정시스템에 의한 예측 결과.
Fig 4. The prediction result using the decision support system (BankHak-Dong).

표 2. 19개 지역의 결과.
Table 2. Prediction result of 19 areas in Seoul.

지역 이름	RMSE	Slope	Intercept
KwangHwaMoon	21.222	0.514	30.500
HanNam	21.844	0.282	35.899
GooEui	13.040	0.896	2.990
SeongSoo	20.827	0.425	30.928
MyeonMok	18.823	0.351	30.349
SinSeol	15.059	0.560	21.886
GilEum	22.859	0.300	38.405
BangHak	20.182	0.641	31.439
BoolKwang	22.801	0.265	45.388
MaPo	17.415	0.442	21.944
HwaGok	27.707	0.300	57.169
GooRo	23.329	0.133	35.489
O-Lyu	16.880	0.442	29.744
MoonRae	21.464	0.286	33.801
ShinRim	20.688	0.554	33.339
DaeChi	21.893	0.599	36.726
BanPo	14.394	0.331	36.079
JamSil	34.287	0.420	45.723
BangIi	25.400	0.734	30.756

그림 4는 학습데이터를 클러스터링후 만들어진 모델에서 나온 각각의 RMSE중 최저인 값을 가지는 모델을 이용하여 1997년 8월 1일부터 8월 10일까지의 오존예측을 나타내고 있다. 그림 4에서 학습 RMSE가 최저인 경우는 클러스터의

수가 4일 때 27.918이고 예측시의 RMSE는 20.183이었다. 성능의 다른 평가방법으로써 예측결과에 대해서 기울기와 절편을 구한다. 기울기는 1, 절편은 0이 되는 경우가 예측이 잘되었다고 할 수 있다. 본 실험에서 기울기가 0.641이고, 절편은 31.439이다. 위와 방법으로 서울의 19개 지역에 대해서 실험한 결과를 표2에 나타내었다.

5. 결 론

본 논문에서 고농도 오존 농도 예측에 있어서 우수한 예측을 위해서는 모델의 구조를 고정시키는 것보다는 매일의 새로운 자료를 모델의 새로운 입력변수로 사용하여 예측모델의 구조를 조정해 나가는 것이 더 우수한 예측 결과를 낼 것으로 본다. 퍼지 클러스터링에서의 클러스터의 개수 선택에 의한 고농도모델의 선정이 중요할 것으로 생각된다. 오존 예측에 있어서 풍향의 성분도 고려되어야 하지만 본 논문에서는 풍향의 성분을 고려하지 않았다. 이런 풍향과 풍속을 같이 고려한 입력성분을 고려한 모델이 예측의 결과를 낼 것으로 본다.

참 고 문 헌

[1] 김용국, 이종범. "하계의 일 최고 오존농도 예측을 위한 신경망 모델의 개발," *한국 대기보전학회지*, vol. 10, no. 4, pp. 224-232, 1994.

[2] 허정숙, 김동술. "다변량 통계분석을 이용한 서울시 고농도 오존의 예측에 관한 연구," *한국대기보전학회지*, vol. 9, no. 2, 1993.

[3] James C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, 1981.

[4] A. G. Ivahnenko. "Polynomial theory of complex system," *IEEE Trans. Syst. Man. Cybern*, vol. SMC-12, pp. 364-378, 1971.

[5] S. Farlow, ed., *Self Organizing Method in Modeling: GMDH-Type Algorithms*, Marcel Deckker, Inc., New York, 1984.

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

[7] 김성신, "A Neuro-Fuzzy Approach to Integration and Control of Industrial Processes : Part I," 퍼지 및 지능시스템 학회 논문지, vol. 8, no. 6, pp. 58-69, 1998.

저 자 소 개

김 재 용 (Jaeyong Kim)

1999년 : 경성대학교 전기공학과 공학사
1999년~현재 : 부산대학교 전기공학과 공학 석사과정

관심분야 : 데이터 마이닝, 신경회로망
E-mail : ktommy@hanmail.net

김 성 신 (Sungshin Kim)

1984년 : 연세대학교 전기공학과 공학사
1986년 : 연세대학교 전기공학과 공학 석사
1996년 : Georgia Inst. of Tech. 공학박사
1996년~1998년 : Appalachian Electronic Instruments Inc. 연구원
1998년~현재 : 부산대학교 전기공학과 조교수

관심분야 : 데이터 마이닝, 퍼지, 예측 시스템
Phone : +82-51-510-4774
Fax : +82-51-512-0212
E-mail : sskim007@pusan.ac.kr

왕 보 현 (Bo-Hyeun Wang)

1987년 : 연세대학교 전기공학과 공학사
1990년 : Georgia Institute of Tech. 공학 석사
1991년 : Georgia Institute of Tech. 공학 박사
1991년~1998년 : LG 종합기술원 책임연구원
1998년~현재 : 강릉대학교 전자공학과 조교수

관심분야 : 기계 학습, 데이터 마이닝, 예측 시스템
Phone : +82-33-640-2384
Fax : +82-33-640-2244
E-mail : bhw@kangnung.ac.kr