퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 4, pp. 435-442

고농도 오존 예측을 위한 향상된 변환 기법과 예측 성능 평가

Modified Transformation and Evaluation for High Concentration Ozone Predictions

천성표* 김성신* 이종범**

Seong-Pyo Cheon, Sungshin Kim, and Chong-Bum Lee

* 부산대학교 전자전기통신공학부

** 강원대학교 환경과학과

요 약

대기중의 고농도 오존의 피해를 줄이기 위해서, 고농도 오존 발생 전에 미리 오존 농도를 예측하기 위한 연구가 진행 되었다. 하지만, 고농도 오존은 그 발생 빈도가 매우 희소하고, 대기 오존 생성 과정이 매우 비선형적이며 복잡한 특징이 있다. 이러한 특징을 극복하고 보다 정확한 예측 모델을 개발하기 위하여, 본 논문에서는 다양한 데이터 처리 기법을 도입하였다. 데이터 전처리과정에서 FCM(Fuzzy C-mean) 방법을 이용하여 오존 농도별 데이터 클러스터링을 시도하였으며, 결측 또는 비정상 데이터를 처리할 목적으로 Rejection 표본 추출법을 이용하였고, 모델의 입력과 출력의 상관관계를 향상시키기 위해서 로그 변환기법을 응용하였다. 오존 예측을 위한 모델링 기법은 DPNN(Dynamical Polynomial Neural Networks)을 이용하였으며, 최소 바이어스 판별법(Minimum Bias Criterion)으로 최적화된 모델을 선택하였다. 끝으로, 본 논문에서는 로그 변환기법이 예측 모델에 미치는 영향을 보이기 위해서 입력 데이터를 두 개의 집합으로 나누어 다양한 방법으로 예측 결과를 평가했다. 결과적으로 계절적 영향에 의해 특정 분포를 가지는 오존 관련 데이터에 있어서 로그 변환 방법이 모델의 성능을 향상 시킬 수 있다는 것을 보였다.

키워드 : 예측 모델, 퍼지 c-means. DPNN, 로그 변환법, 최소 바이어스 판별법.

Abstract

To reduce damage from high concentration ozone in the air, we have researched how to predict high concentration ozone before it occurs. High concentration ozone is a rare event and its reaction mechanism has nonlinearities and complexities. In this paper, we have tried to apply and consider as many methods as we could. We clustered the data using the fuzzy c-mean method and took a rejection sampling to fill in the missing and abnormal data. Next, correlations of the input component and output ozone concentration were calculated to transform more correlated components by modified log transformation. Then, we made the prediction models using Dynamic Polynomial Neural Networks. To select the optimal model, we adopted a minimum bias criterion. Finally, to evaluate suggested models, we compared the two models. One model was trained and tested by the transformed data and the other was not. We concluded that the modified transformation effected good to ideal performance in some evaluations. In particular, the data were related to seasonal characteristics or its variation trends.

Key Words : Prediction model, fuzzy c-means. DPNN, Log transformation, Minimum bias criterion.

1. 서 론

지구상에 인류가 출현한 이후로, 자연현상이나 인간의 삶과 운명에 대한 인과관계를 찾거나 예측하려는 시도는 끊임없이 계속되어 왔다. 인류 역사상으로 보더라도 수많은 과학자들이나 예지자들이 미래를 예측하였으며, 이들은 어떤 물리적 현상뿐만 아니라 화학적 반응 그리고 별자리 변화와 기상 현상을 통해서 예측이나 예언을 해 왔다. 수없이 많은 방법이 있지만, 아직도 모든 현상에 적용 가능한 절대적인 방법은 존재하지 않는다. 다만, 최근 IT 기술의 발전으로 인해서 데이터 마이닝이라는 분야가 새롭게 대두되었고, 이를 이용한 예측 모델 개발이나 의사결정 지원 시스템 개발 등이 활발히 이루지고 있으며, 점차 그 성능을 인정받고 있다[1-4]. 그러므로, 최근에는 예측 모델 개발이 곧 데이터 마이닝이라고 인식되어, 으레 예측 변수와 관련된 데이터를 모으고, 데이터의 특징을 추출할 때까지 분석하고, 발견한 특징을 근간으로 모델을 구성하는 것이 지극히 당연하게 받아들여지고 있다. 그런데 데이터 마이닝을 이용한 모델 개발과정에서 과정에서 가장 중요한 것은 특징을 찾아내는 직관력이다. 다행히도, 본 논문의 연구 대상인 고농도 오존 예측 시스템 개발은 앞서 언급한 과정들을 적용해 보고 시험해 볼 수 있는 훌륭한 예가 된다. 한 여름 대기상의 오존은 크게 네 가지 특징을 가진다. 첫째, 고농도 오존은 전 세계적으로 점차 증가한다고 하지만, 아직까지는 극히 드물게 발생하는 현상이다.

접수일자 : 2007년 4월 6일
완료일자 : 2007년 7월 27일

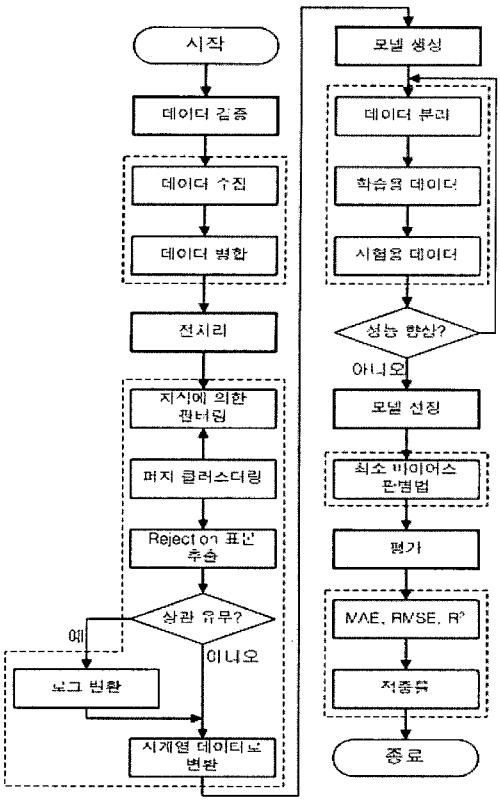


그림 1. 오존 예측 시스템의 과정 및 데이터 흐름도.
Fig. 1. The flow chart of the suggested Ozone prediction system.

둘째로, 대기중 오존은 다양한 전구물질들의 반응에 의해 생성되는 2차 오염물질이므로 그 생성과정이 복잡하고 비선형적이다. 셋째, 대기중 오존은 어린이의 치사량을 높이고, 동·식물의 생장에 악영향을 주는 것으로 보고되고 있어 보다 정확한 고농도 오존 예측 모델의 개발이 절실하다[5-6]. 넷째, 고농도 오존은 지역적인 특색이 뚜렷하게 나타나기 때문에 한 가지 접근법으로 모든 나라와 지역의 문제를 해결 할 수 없다. 본 논문에서는 사례 연구를 통해서 실제 오존 관련 데이터를 수집하는 것에서부터 예측 결과를 평가하는 것까지 실용적이고 심층적으로 고찰해 보고자 한다. 본 연구에서 제안한 오존 예측 시스템에 대한 데이터 처리과정에 대해서 간략히 설명하겠다. 그림 1에 보이는 것처럼 첫째, 데이터 검증 단계에서 오존 발생과 관련된 기상학적 데이터와 오염물질 데이터를 수집한 후 필요한 경우 데이터를 병합한다. 전처리 과정에서는 지식에 의한 필터링과정을 통해서 예측 지역과 적절한 입·출력 변수를 선정한 후, 퍼지 클러스터링과정을 거친다[7]. 물론, 결측 및 비정상 데이터는 Rejection 표본 추출법으로 보간한다[8]. 더불어 각 변수의 상관관계를 계산하여 로그 변환 대상 변수를 선정하고 시계열 데이터로 변환한다. 이어지는 모델 생성과정에서는 동적다항식신경망(Dynamical Polynomial Neural Networks: DPNN)을 학습시키기 위해서 관련된 데이터를 임의 추출법을 이용해서 여러 데이터 쌍으로 분리한 후, 학습용 데이터와 시험용 데이터 집합으로 나눈다[9]. 모델 성능 평가 함수를 이용하여 학습에 따른 모델의 성능 향상 여부를 파악하여 과도학습을 방지한다. 여러 데이터 쌍을 이용하여 생성된 여러 모델 중에서는 최소 바이어스 판별법으로 하나의 모델을 선택한 후 예측 결과를 다양한 방법으로 평가하게 된다[10].

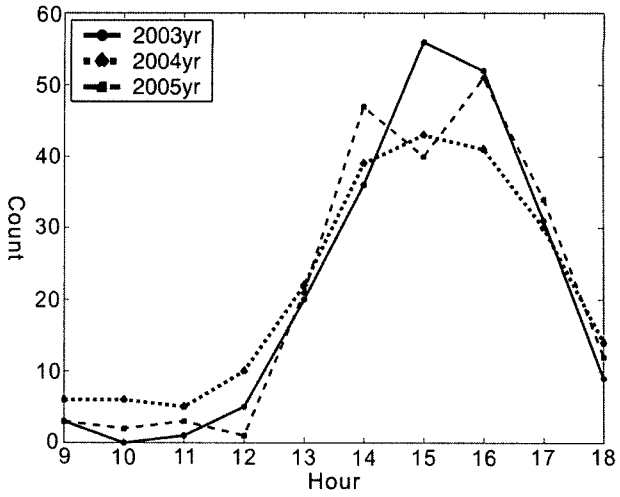


그림 2. 년도별 일별 최고 오존 농도 발생 시간 분포.
Fig. 2. Daily maximum Ozone measured hour(s).

2. 데이터 검증

대기 오염과 관련하여 최근에 가장 화제가 되고 있는 주제는 한 여름 대기권 고농도 오존 발생 현상이다. 오존은 강한 산화 물질이기 때문에 사람뿐 아니라 동·식물에게도 악영향을 주는 것으로 알려져 있다. 대기중 오존은 일반적으로 대기중 오염물질이 광화학적 작용에 의해서 생성되는 것으로 알려져 있다. 가장 흔하게 알려진 화학 반응은 대기중 이산화질소(NO_2)와 이산화탄소(CO_2)가 전구물질로 작용하고, 이들이 바람이 없고 높은 온도에서 일정량이상의 자외선(UV)에 노출 되었을 때 오존(O_3)이 생성된다고 한다. 따라서, 오존은 2차 오염물질이다[11]. 본 연구에서 사용된 데이터는 서울시 보건환경연구원이 서울시의 27개 관측소로부터 2002년부터 2005년 사이의 수집한 18개 성분의 기상 및 오염물질 자료이다.

3. 전처리 과정

좋은 데이터를 이용하였을 때 좋은 성능의 모델을 개발할 수 있다는 것은 당연하다. 따라서, 수집된 데이터에 적절한 처리과정을 거치게 해서 데이터의 특징이 쉽게 추출되도록 해주는 것이 전처리과정의 역할이다. 본 연구에 사용된 원천 데이터는 총 18개 성분으로 구성되어 있으며 그 성분은 오존, 이산화황(SO_2), 일산화질소(NO), NO_2, NO_X, 일산화탄소(CO), PM10, 풍속(Ws), 풍향(Wd), 온도(TMP), 상대습도(Rh), 자외선량, 일사량(Sr), 총부유먼지(TSP), PM2.5, 메탄(CH_4), 비메탄계 탄화수소(NMHC), 그리고 총탄화수소(THC)이다. 따라서, 이들이 본 연구에서 사용가능한 입력변수 후보가 된다. 전처리과정의 첫 번째는 경험이나 지식에 기반한 접근법을 적용하였다[12]. 첫째, 오존 예측 기간을 고농도 오존이 발생하는 5월에 10월 사이로 한정하였다. 둘째, 그림 2에 보이는 바와 같이 최고 오존 농도 발생 시간이 년

표 1. 일 최고 오존 농도와 각 성분별 R^2값 및 일별 오존 농도차와 각 성분별 R^2값
Table 1. R^2 values of (daily max ozone) vs. (daily max ozone - daily min ozone)

	O_3	SO_2	NO	NO_2	NO_X	CO	PM10	Wd	Ws
일최고 O_3	N/A	0.428	0.747	0.426	0.429	0.426	0.406	0.398	0.410
일최고-일최저 O_3	0.747	0.426	0.429	0.426	0.406	0.398	0.410	0.355	0.290
	TMP	HUM	UV	SR	TSP	PM2.5	CH_4	NMHC	THC
일최고 O_3	0.355	0.290	0.387	0.468	0.252	0.451	0.500	0.499	0.500
일최고-일최저 O_3	0.387	0.468	0.252	0.451	0.500	0.499	0.500	0.500	0.500

도가 증가함에 따라 점차적으로 늦어지는 경향을 보이므로, 데이터의 시간 범위를 9시에서 18시까지로 제한하였다. 일중 최고 오존 농도는 주로 14시에서 15시 사이에 발생한다는 주장이 많았으나, 서울의 경우 15시 이후에도 종종 발생하는 경우가 나타났다. 셋째, 오존 전구물질 유효시간을 5시간미만으로 한정하였다. 이것은 지리적 기상학적 고려에 의한 판단이다. 기상학 또는 대기과학의 입장에서는 원천 데이터의 각 성분이 고농도 오존 발생에 미치는 영향도 매우 중요한 연구 주제지만, 본 연구에서는 원천 데이터를 이용한 고농도 오존 예측에 집중했다. 전처리 과정에서 고농도 오존과 관련된 몇 가지 특징을 발견하였다. 가장 고유한 특징은 고농도 오존 발생은 "온도가 섭씨 30도 이상이면 고농도 오존이 발생한다"와 같이 개별 전구물질이나 기상 자료의 절대조건에 의해 발생하지 않는다는 것이다. 결국 데이터 마이닝 관점에서는 "오전 10시의 온도가 섭씨 30도 이상이고 오전 11시의 온도와의 차이가 클수록 고농도 오존이 발생할 확률이 높다"와 같이 각 입력 성분의 변화폭과 절대값을 동시에 고려해야 한다는 결론을 내렸다. 이 가정을 뒷받침하기 위해서 표 1에 2004년도 데이터의 각 성분과 일 최고 오존 농도와의 상관관계 그리고 일 최고 농도와 최저 농도차와의 상관관계를 각각 정리했다. 본 연구의 전처리과정은 모델 성능 비교 및 평가에 사용될 학습용 데이터 집합과 시험용 데이터 집합 그리고 성능 평가용 데이터 집합으로 나누는 것으로 마무리 된다.

3.1 Rejection 표본 추출

데이터를 수집한 후, 전처리과정에서 가장 중요한 것은 결측 및 오류 데이터에 대한 보간이다. 이전 논문에서 보간법은 주로 선형보간법, 이동평균법 등이 적용되었다[13]. 그러나 본 연구에서는 Rejection 표본 추출법을 제안하였다. Rejection 표본 추출법은 데이터 분포를 파악하기 힘들지만 조건부 확률을 계산해야 할 경우 사용하는 방법으로 간단하면서도 강력한 표본 추출 기법이다[14]. Rejection 표본 추출 기법 과정을 간단히 기술하겠다. 첫째, 원천 데이터로 임의 추출을 통해 표본 집합을 만든다. 둘째, 표본 집합의 원소 중에서 원하는 조건을 만족시키는 원소만을 남기고, 그 외의 원소는 제외한다. 끝으로, 남겨진 원소를 이용하여 조건부 확률을 추정한다. 본 연구에서 Rejection 표본 추출법에 관심을 가진 이유는 이 표본 추출법이 가진 우연성과 휘발성 때문이다. 예를 들면, 특정일 13시의 NOx값이 결측이면, 원천 데이터 가운데 임의로 100일을 선택한 후, 결측일의 최고 오존 농도와 유사한 범위값을 가지는 날짜의 자료만을 남긴 후, 이 집합의 13시 평균 NOx농도 계산하여 보간하는 방법이다. 이전 연구에서 사용되었던, 선형보간법과 이동평균법의 경우 결측 변수가 선형 분포라는 가정하에서 보간을 시행했기 때문에 보간한 값은 다수 분포하는 값을 따르는 편향성을 가지게 되고 결국 예측 결과 역시 선형 분포를 가지게 되어 희소 현상인 고농도보다 데이터 수가 많은 중·저농도 모델로 되어 버리는 문제가 발생하였다.

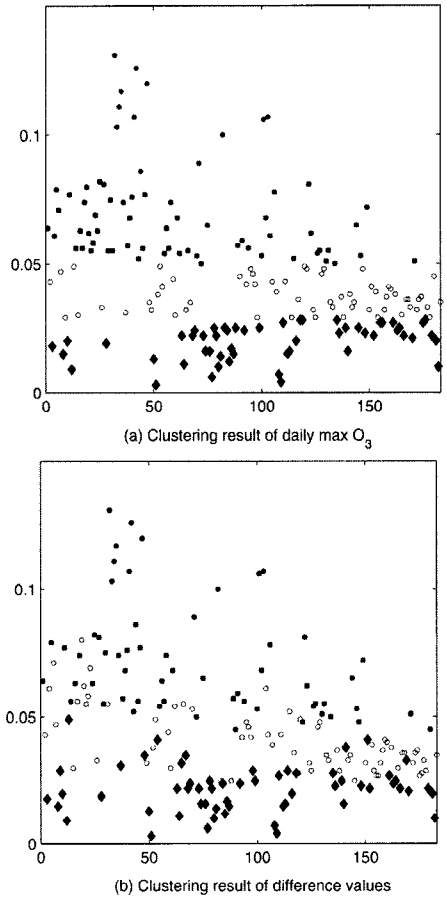


그림 3. 일별 오존에 관한 FCM 결과.
Fig. 3. FCM results of daily ozone.

3.2 퍼지 C-Means 클러스터링

고농도 오존은 연중 수회 발생하는 희소 현상이기 때문에 단일 또는 통계 모델로 오존 발생 전 범위에 대해 예측하는 것은 무리가 있다. 오존 예측에서 기존에 연구 되었던 통계 모델을 적용하였을 경우 평균 농도 근처에서 다수를 차지하는 중·저농도의 경우 예측 가능한 것으로 판명되었으나, 고농도 예측의 경우 데이터 자체의 희소성 문제와 더불어 고농도 오존이 발생했을 경우 원인을 분석해 보면 대부분 경우에 따라 다른 원인에 의해 발생하는 것으로 파악되는 실정이라 모델의 성능을 보장하기 힘들었다[15]. 본 연구에서는 이러한 단점을 극복하기 위해서 그림 3에 보이는 것과 같이 일별 최고 오존 농도값에 퍼지 C-Mean 클러스터링 기법을 적용하여 데이터를 나눈 후, 각 영역별 개별 모델을 구성하였다. 경험적으로 볼 때, 희소 현상 예측의 경우 변수들에 대한 최소 경계와 최대 경계를 정하는 것이 바람직하다. 그 이유는 희소 현상 예측 모델을 학습시키는 데이터 집합 또한 희소하기 때문에 특정한 한 개 혹은 두 개 변수 조건의 영향에 의해서 예측값 변동폭이 매우 크기 때문이다. 본 연구에서는 FCM을 이용하여 그림 4와 표 2를 구한 다음, 이웃한 클러스

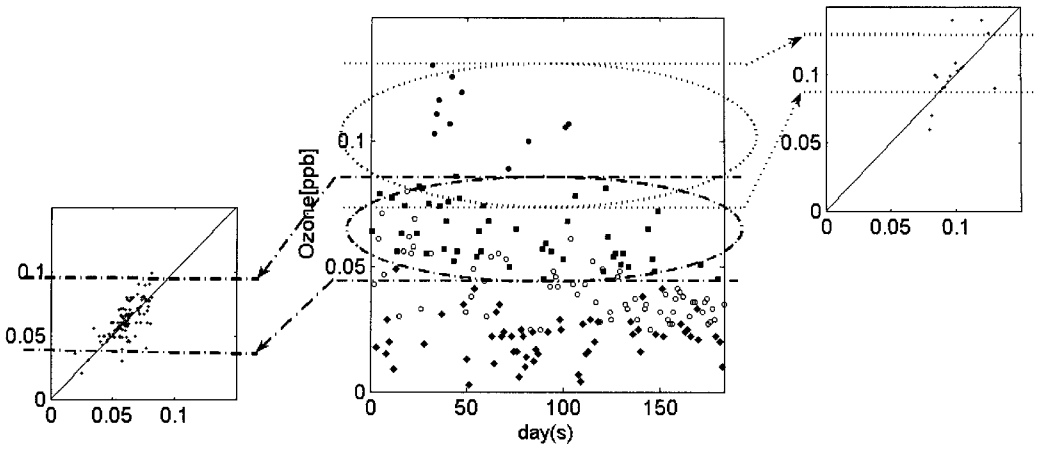


그림 4. FCM을 이용한 예측 경계값 설정.
Fig. 4. Setting of prediction limits with FCM.

터의 중심값을 이용하여 모델 입력 변수의 최적 경계값을 구했다. 이 과정은 모델링과정의 전과 후에 각각 수행되는데, 희소 현상 예측을 위해서 본 연구에서 제안한 독창적인 과정이다. FCM은 HCM(Hard C-Mean) 알고리즘을 보다 일반화 시킨 방법으로 특정한 점이 갖는 모든 클러스터에 대한 소속도를 계산하여 클러스터링하기 때문에 클러스터의 경계 영역에 속한 점도 자연스럽게 나누어 주는 장점이 있다 [14,15]. 본 논문에서는 일별 최고 오존 농도 데이터에 FCM을 적용해서 퍼지 공간 분할을 한 후, 각 클러스터에 속한 일별 최고 농도의 기상 및 오염물질 데이터를 가져와서 모델을 학습시키는 방법을 사용한다.

표 2. 일 최고 오존 농도와 일별 오존 농도차에 대한 FCM 결과
Table 2. FCM results of (daily max ozone) vs. (daily max ozone - daily min ozone)

		클러스터 1	클러스터 2	클러스터 3	클러스터 4
일최고 O_3	중심값	0.109	0.062	0.038	0.020
	평균값	0.108	0.063	0.038	0.020
	최대값	0.131	0.082	0.049	0.028
	최소값	0.086	0.050	0.029	0.003
일최고 -일최저 O_3	중심값	0.093	0.049	0.030	0.014
	평균값	0.110	0.063	0.041	0.022
	최대값	0.131	0.086	0.080	0.049
	최소값	0.089	0.045	0.024	0.003

3.3 로그 변환의 적용

원천 데이터는 전체 18개 성분으로 구성되었으며, 표 1에 나타난 바와 같이 오존, PM10, TMP, HUM, TSP, PM2.5 그리고 NMHC 성분의 경우에는 표본 추출 시간에서의 성분값과 당일 최고 오존 농도의 상관관계보다는 당일 오존 농도의 일교차와 상관관계가 더 높은 것을 알 수 있다. 따라서 본 연구에서는 이러한 특징을 예측 모델에 명확히 반영하기 위해서 위의 성분 값에 식 (1)과 같은 로그 변환을 취하였다.

$$y = 1 - \log_{10}(A_{\max} / A_{current}) \quad (1)$$

여기서, $A_{\max}$는 해당 성분의 일 최고값이고, $A_{current}$는 현재 상태 값이다. 제안한 변환 기법의 유효성을 보이기 위해서 그림 5에 제시된 것처럼 흔히 사용하는 가우시안 정규화 방법과 시계열 분석 과정에서 대표적인 변환 방법인 Box-Cox변환과 비교해 보았다. 본 논문에서는 제안된 로그 변환법을 가장 유명한 변환 방법군인 Box-Cox 변환과 비교해 보았다. Box-Cox 변환식은 식 (2)에 나타내었다.

표3. 원천데이터와 식(1)과 (2)를 이용한 변환 그리고 정규화 결과
Table 3. Raw data and transformed data of Eqs.(1),(2) and normalization

	점 1	점 2	점 3	점 4	점 5
원천 데이터	0.02	0.03	0.04	0.06	0.07
식 (1)	0.59	0.65	0.81	0.96	1.06
식 (2)	0.93	0.82	0.45	0.09	-0.43
정규화 결과	1.72	0.89	0.48	0.63	0.43

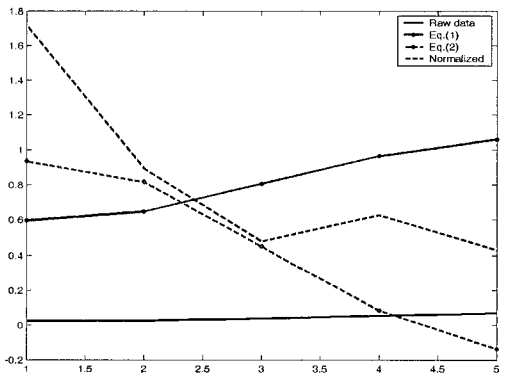


그림 5. 원천 데이터와 변환 데이터의 경향 비교.
Fig. 5. Plot of raw data and transformed data trends.

$$g(X_t)=\begin{cases}\dfrac{X_t^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln(X_t), & \lambda=0,\end{cases} \tag{2}$$

여기서, $\{X_t\}$는 $X_t>0$인 시계열 관측 데이터를 나타내며, $\ln(\cdot)$는 자연로그를, 그리고 λ는 실수값을 가지는 상수이다. $\lambda=0$인 경우 변환은 $\lim_{\lambda\to 0}\dfrac{X_t^{\lambda}-1}{\lambda}=\ln(X_t)$을 만족하게 된다. Box & Cox 참고[18,19].

변환 결과의 이해를 돕기 위해서 그림 5와 표 3을 제시하였다. 그림 5를 보면 식(1)을 이용한 변환 결과는 원천 데이터 분포와 유사하면서 동시에 각 점의 변화량을 확대시켜 나타낸다는 것을 쉽게 알 수 있다. 식 (2)를 이용한 경우는 분포 경향성이 역으로 나타나며, 단순한 정규화 역시 식 (2)를 이용한 변환과 유사한 경향을 보인다. 이로써 본 연구에서 제안한 식 (1)을 이용한 로그 변환은 표본 추출 시간 사이의 변화량을 강조하는 효과가 있음을 알 수 있다.

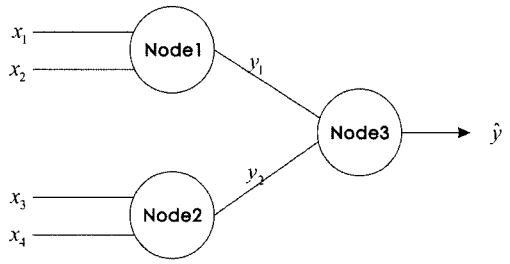


그림 6. 네 개의 입력 변수와 하나의 출력 변수를 가진 기본적인 DPNN 구성.
Fig. 6. Basic structure of 4-input and 1-output DPNN.

4. 동적 다항식 신경망을 이용한 모델링

4.1 DPNN의 기본 구성

DPNN은 GMDH(Group Method Data Handling) 기법을 사용하여 측정 데이터나 변수를 기초로 하여 입·출력 모델을 구성하는 방법이다[18]. 이 방법은 모델링, 예측, 그리고 지능형 제어기에 광범위하게 사용된다. 입력 변수 4개와 출력 변수 1개로 구성된 간단한 DPNN구성의 경우를 식 (3)과 식 (4)로 표현하였다. 입력 변수 x_1, x_2가 출력 변수 y_1을 x_3, x_4가 y_2를 결정하고, 이들 값으로부터 최종 출력 값 $\hat{y}$이 계산되어 진다.

$$\begin{aligned}y_1&=\omega_{01}+\omega_{11}x_1+\omega_{21}x_2+\omega_{31}x_1x_2+\omega_{41}x_1^2+\omega_{51}x_2^2\\ y_2&=\omega_{02}+\omega_{12}x_3+\omega_{22}x_4+\omega_{32}x_3x_4+\omega_{42}x_3^2+\omega_{52}x_4^2\end{aligned} \tag{3}$$

그리고,

$$\hat{y}=\omega_{03}+\omega_{13}y_1+\omega_{23}y_2+\omega_{33}y_1y_2+\omega_{43}y_1^2+\omega_{53}y_2^2 \tag{4}$$

여기서, $\omega_{ij}(i=0,1,2,\cdots,n,\ j=0,1,2,\cdots,k)$는 계수이다. 만약, 입력 변수가 세 개 이상이라면, 위 식 (3)과 식 (4)는 더욱 복잡하게 될 것이다. DPNN에서 각 노드의 계수는 최소자승법으로 추정한다. 본 연구에서 최종 출력단의 예측값과 실제값 사이의 오차를 이용한 목적함수를 식 (5)와 같이 정의했으며, 최적의 목적함수를 만족하는 계수는 식 (6)을 이용해서 계산한다.

$$J=\sum_{k=1}^{\nu m\,of\,data}(y(k)-\hat{y}(k))^2=\|\,y-\omega A\,\|^2 \tag{5}$$

1.
$$\omega=(A^TA)^{-1}A^Ty \tag{6}$$

식 (5)와 (6)을 이용한 계수 추정과정을 반복하므로 인해서 결국 DPNN은 최고 성능의 다항식 구조와 계수를 결정하게 된다.

4.2 자기 조직화

DPNN의 다른 특징 중 하나는 자기 조직화이다[20]. GMDH 기법을 근간으로 한 DPNN은 데이터를 학습용과 시험용으로 나누어 학습을 수행함으로써, 과도 학습을 방지하고 입·출력 값의 안정성을 동시에 추구한다[21]. 단순히 데이터를 나누는 것만으로 과도 학습이 방지되는 것이 아니고, 성능 판별식(Performance Criterion; PC)을 정의하여 이것으로부터 학습데이터를 이용하여 계수를 향상시킨 후, 시험용 데이터로 출력값을 계산하여 성능을 평가한다. 특히, DPNN의 경우 각 단에서 다음 단으로 네트워크를 확장시킬지 여부와 이전 단에서 특정 노드를 선택하고 제외하는 과정에도 PC를 적용함으로써, 최종적인 DPNN의 구조는 완전히 데이터에 의해서 결정된다고 할 수 있다. 본 연구에서는 식(7)과 같이 성능 반별식을 정의하였다.

$$\begin{aligned}e_1^2&=\sum_{i=1}^{n_A}(y_i^A-f_A(x_i^A))^2/n_A,\\ e_1^2&=\sum_{i=1}^{n_B}(y_i^B-f_B(x_i^B))^2/n_B,\\ PC&=e_1^2+e_2^2+\eta(e_1^2-e_2^2)^2\end{aligned} \tag{7}$$

여기서, e_1, e_2, n_A, n_B, y_i는 각각 학습오차, 시험오차, 학습데이터 수, 시험데이터 수, 그리고 실제 출력값이다. 그리고, $f_A(x_i^A)$와 $f_B(x_i^B)$ 는 각각 학습데이터를 입력했을 때 예측값과 시험데이터를 입력 했을 때의 예측값을 나타낸다. 전체 데이터 수는 $n=n_A+n_B$가 되며, 성능 판별식은 PC를 최소화 할 때 가장 좋은 결과를 나타내게 된다.

$$\eta_{bs}^2\equiv\sum_{p\in W}(\widehat{y^A}-\widehat{y^B})_p^2/\sum_{p\in W}y_p^2 \tag{8}$$

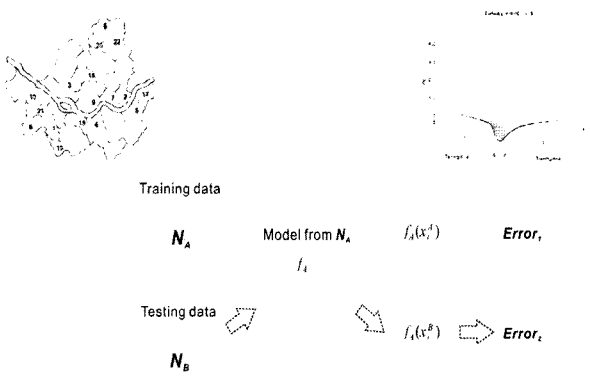


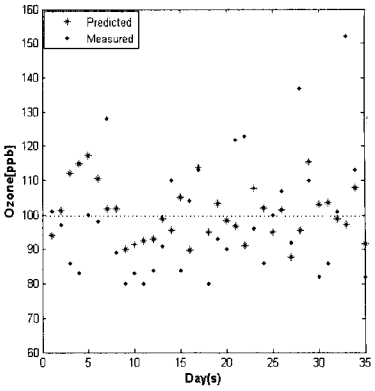
그림 7. 데이터 수집, 분류, 그리고 모델 평가 모식도.
Fig. 7. Schematic diagram of data gathering, splitting and evaluation.

4.3 모델 선정: 최소 바이어스 판별법

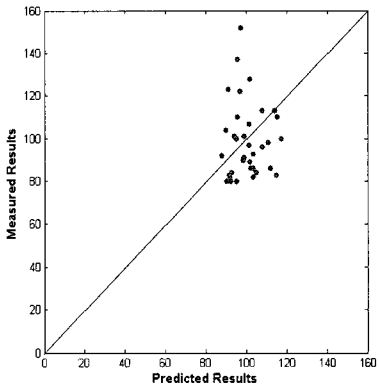
모델 선정과정은 민감한 과정이다. 이것은 사전에 연구목적과 모델링 방법 및 개발과정 그리고 연구과정에서 발견된 여러 가지 제약 조건들까지 고려해서 몇 개의 모델 후보들 중에서 결정해야 한다. 본 연구의 경우 가장 큰 고려 사항은 희소 현상의 예측 모델이라는 점이다. 그래서, 입력 변수의 편향성과 모델의 과도 학습을 극복하는 데 초점을 맞추어 모델 선정 기법을 최소 바이어스 판별법으로 결정했다. 이 방법은 예측을 위해서 사용할 N개의 입력 자료를 A, B, C 세 개의 부분집합으로 p번 나눈다. 여기서, $A \cap B = W$이고 $C = N - W$이다. W는 학습 및 시험 데이터이고, C는 평가용 데이터이다. 따라서, p번째 부분집합 A, B로 모델을 학습 및 시험하는데 각 집합의 역할을 교환하여 두 개의 모델을 생성한 후 식 (8)를 이용하여 선정한다. p개의 부분집합 A, B, C 중에서 가장 η_{bs}가 0에 가까운 모델을 선정하기 때문에 최소 바이어스 판별법이라고 한다. 여기서, $\hat{y}^A$는 집합 A로 학습된 예측값이고, $\hat{y}^B$는 집합 B로 학습된 예측값이다. 만약, 집합 A와 집합 B로 학습 시킨 모델의 예측값 $\hat{y}^A = \hat{y}^B$이면, 두 모델은 바이어스 되지 않았고 따라서 $\eta_{bs} \rightarrow 0$ 이다.

4. 시뮬레이션

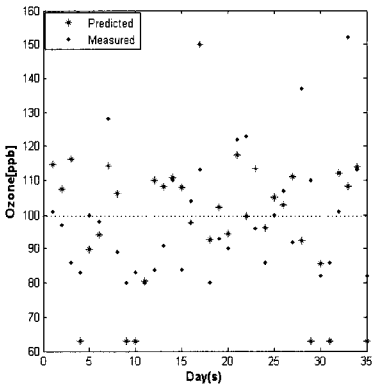
본 연구에서는 일 최고 농도가 100ppb이상 인 날을 고농도 오존 발생일이라고 정의했으며, 서울 지역의 2002년에서 2005년 사이의 기상 및 오염물질 데이터를 분석하여, 가장 고농도 오존 발생 빈도가 높은 관측 지점인 관악구를 예측 지점으로 선정했다. 관악구 지역에 대한 시뮬레이션은 2004년 데이터를 학습용 및 시험용으로 활용하였으며, 2005년 5월부터 10월 사이의 184일간의 오존 농도를 예측하도록 했다. 먼저, FCM을 이용하여 일 최고 오존 농도를 4개의 클러스터로 나누었다. 그 결과 89ppb이상 인 날을 고농도 발생일로 결정했다. 고농도 발생 관련 데이터만을 학습용 데이터로 삼고, 최고 농도 발생 5시간 이전까지의 자료를 활용하여 시계열 형태로 구성하였다. 결측 및 오류 데이터 보간은 Rejection 표본 추출 기법을 적용하였으며, 표본 추출은 희소 현상임을 고려하여 고농도 영역과 바로 아래 클러스터까지 포함하여 시행하였다. 로그 변환의 성능을 파악 할 목적으로 하나의 DPNN모델은 정규화 과정만 거친 입력 데이터를 사용하였고, 다른 모델은 로그 변환을 거친 변수를 포함한 입력 데이터를 사용하였다. 모델 생성과정에서 학습용 및 시험용 데이터의 분포나 편향성에 의한 예측 성능의 차이를 제거하기 위해서 임의 표본 선택을 100번 수행하였으며, 각 데이터 쌍의 조합을 통해서 모델을 학습 시켰다. DPNN 모델의 레이어 확장과 멈춤은 PC를 이용하였다. 이렇게 생성된 모델들 중에서 최종 모델 선정은 최소 바이어스 판별법을 이용했다. 2005년 5월부터 10월까지의 관악구 지점에 대한 예측 결과는 표 4에 정리해 두었다. 객관적인 모델의 성능 평가를 위해서 희소 예측 결과 평가에 사용되는 성능 평가 지수를 정의하여 사용하였다. 표 4에서 N은 오존 농도를 예측하고 측정한 전체 일수를 나타내며, F는 모델이 고농도라고 예측한 전체 횟수를, M은 고농도가 측정된 전체 횟수를 나타낸다. 따라서, N은 184일, M은 15일, 그리고 F는 각 모델별로 16회와 20회가 된다. A는 고농도 발생을 예측하고 실제로 고농도 오존에 발생한 횟수이므로, 각 6회와 11회이다.



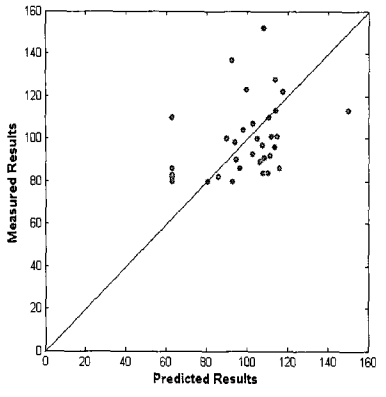
(a) 로그변환 거친 입력 변수를 이용한 예측결과와 측정결과 비교.
(a) Prediction and measured results of log transformed input model.



(b) 로그변환 입력 변수를 이용한 예측결과와 측정결과 산포도.
(b) Scatter plot of log transformed input model.



(c) 정규화만 거친 입력 변수를 이용한 예측결과와 측정결과 비교.
(c) Prediction and measured results of untransformed input model.



(d) 정규화만 거친 입력 변수를 이용한 예측결과와 측정결과 산포도.
(d) Scatter plot of untransformed input model.

그림 8. 관악구에 대한 두 모델의 예측 및 측정 결과 비교.
Fig. 8. Comparison results of different two-type models in Gwanak-gu.

이 값을 이용해서 본 연구에서는 모델이 고농도를 예측하고 실제로 고농도를 측정한 T.P.R. (=A/M, True Positive Rate)와 고농도가 아니라고 예측했으나 실제로 고농도가 관측된 경우의 비율을 계산한 F.P.R (=(F-A)/(N-M), False Positive Rate)를 모두 계산하였으며, 위의 두 비율 지수와 더불어 잘못된 경보 횟수를 나타내는 F.A. (=(F-A)/F, False Alarms)와, 성공 지수 S.I. (=T.P.R - F.P.R, Success Index)를 정의하여 종합적으로 평가하였다. 평가 결과 표 4에 나타난 대로 로그 변환을 이용하여 매 추출 시간별 변화량을 강조하여 입력한 로그 변환 입력 모델의 성능이 모든 지수에서 앞서는 것으로 나타났다.

표 4. 희소 예측 모델 평가 지표를 이용한 예측 결과 평가.
Table 4. Evaluation results of Rare event prediction model evaluation methods.

	A	F	M	N	T.P.R	F.P.R	F.A.	S.I
정규화 변환 입력 모델	6	16	15	184	0.400	0.059	0.625	0.341
로그 변환 입력 모델	11	20	15	184	0.733	0.053	0.450	0.680

그림 8과 표 5에서는 제안한 모델의 고농도 오존 예측 성능 좀 더 살펴보기 위해서 2005년에 실제 80ppb이상 고농도 오존이 발생한 35일에 대한 모델별 예측값과 측정값을 (a)와 (c)에 그리고, 산포도를 각각 (b)와 (d)에 나타냈었다. 산포도의 경우 대각선 가까이 존재하는 점이 예측치와 실측치의 오차가 유사한 것이다. 표 5에서는 예측 모델 성능을 가장 흔히 사용하는 R^2, MAE, RMSE를 이용해서 정리해 보았다. 이 경우에는 MAE와 RMSE가 정규화 변환 입력 모델이 더욱 작게 나타나고, R^2값 만 로그 변환 입력 모델이 우수한 것으로 나타나는데 이것은 실측치와 예측치의 오차값을 이용해서 계산하였기 때문이다. 따라서, 기본적인 통계 모델 평가 방법을 희소 데이터 예측에 적용할 경우, 다양한 평가 지수를 도입해서 평가할 필요가 있다. 본 연구에서는 개별 예측치와 실측치의 오차를 줄이는 것을 목적으로 하는 것이 아니라, 희소 데이터인 고농도 오존의 발생을 예측하는 것이기 때문에 표 4에서와 같이 희소 데이터 예측 평가 지수를 이용해서 제안한 방법을 이용한 모델이 더욱 우수한 성능을 나타낸다고 판단하였다.

표 5. 기본적인 모델 평가 지표를 이용한 예측 결과 평가.
Table 5. Evaluation results of basic evaluation methods.

	R^2	MAE	RMSE
정규화 변환 입력 모델	0.155	0.015	0.112
로그 변환 입력 모델	0.483	0.016	0.121

5. 결론 및 고찰

본 연구는 희소 현상 중의 하나인 대기중 고농도 오존 예측 모델 개발이 목표이다. 희소 현상 예측 분야는 데이터 마이닝이 지금껏 다양한 접근법으로 도전해 온 분야 이다. 신용 카드 사기 여부 결정, 산업 분야에서 진단 및 검출, 화학 반응에서의 결과 분석 등 이와 유사한 현상에 대한 연구가 있었다. 하지만, 이런 연구들이 직면하게 되는 여러 문제점이 있는데, 그 중에서 몇 가지를 짚어 보겠다. 우선, 대체적으로 희소 현상은 각 사건들이 서로 연관성이 적은 독립적인 움직임을 가지므로 일반화시키기가 힘들다. 둘째, 현상을 나타내는 데이터 자체의 희소성을 들 수 있다. 셋째, 데이터 마이닝을 통한 예측 결과를 검증 할 수 있는 방법 역시 부족하다는 점이다. 본 연구에서 우리는 데이터 마이닝의 시작에서부터 예측 모델을 구성하여 예측 결과를 이끌어 내기까지 다양하면서도 실제적인 방법들을 적용해 보았으며, 그 결과를 고찰하였다. 전처리 과정에서 지식과 경험적 접근법, 데이터 클러스터링, 결측 데이터 보간을 위한 Rejection 표본 추출법을 사용 하였다. 특히, 시계열 모델의 입력 변수가 가진 특징을 부각 시켜주는 로그 변환을 응용해 보고, 그 결과를 확인했다는 점 역시 연구 성과 중의 하나이다. 자기 조직화 능력을 가진 DPNN의 성능을 최적화시키며, 데이터 편향성과 과도 학습을 막는 성능 판별식의 제안과 후보 모델군 중에서 가장 적절한 모델을 선정하는 최소 바이어스 판별법의 적용 역시 본 연구의 결과이다. 이러한 다양한 과정을 거친 결과, 희소 현상에 예측 모델이 여러 성능 평가 기준을 넘어서는 좋은 결과를 가지게 되었다.

참 고 문 헌

[1] Wang, Z. Xue, Data Mining and Knowledge Discovery for Process Monitoring and Control, Springer-Verlag , Berlin Heidelberg London, 1999.
[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery: An Overview. In : Fayyad, U.M. et al: Advances in

Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, California, 1996.
[3] J. F. Elder IV, D. Pregibon, A Statistical Perspective on Knowledge Discovery in Databases. In U. M. Fayyad et al, Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, Menlo Park, California, 1996.
[4] B. Devlin, Data Warehouse from Architecture to Implementation, Addison-Wesley, 1997.
[5] R. T. Burnett, M. Smith-Doiron, D. Stieb, M. E. Raizenne, J. R. Brook, R. E. Dales, J. A. Leech, S. Cakmak, and D. Krewski, "Association between ozone and hospitalization for acute respiratory diseases in children less than 2 years of age," Am. J. Epidemiol. Vol.153, pp.444-452, 2001.
[6] R. Matyssek, W. M. Harvranek, G. Wieser, J. L. Innes, Forest Decline and Ozone, A Comparison of Controlled Chamber and Field Experiments, Springer, Berlin, 1997.
[7] J. Kim, S. Kim, and B. Wang, "Forecasting High-Level Ozone Concentration with Fuzzy Clustering," KFIS, Vol. 11, No.4, pp.336-339, 2001.
[8] Stuart Russell, Peter Norvig, Artificial Intelligence - A Modern Approach, Prentice Hall, 2003.
[9] D. T. Pham, and L. Xing, Neural Networks for Identification, Prediction and Control, Springer-Verlag, London, 1995.
[10] H. R. Madala, A. G. Ivakhnenko, Inductive Learning Algorithms for Complex Systems Modeling, CRC Press, 1994.
[11] M. Millan, R. Salvador, E. Mautilla, "Meteorology and photochemical air pollution in southern Europe: experimental results from EC research projects," Atmospheric Environment Vol.30, pp.1909-1924, 1996.
[12] T. Dasu, T. Johnson, Exploratory Data Mining and Data Cleaning, John Wiley & Sons, 2003.
[13] S. Cheon, S. Kim, "Directed Knowledge Discovery Methodology for the Prediction of Ozone Concentration," Lecture Notes in Computer Science, Vol. 3613, pp.772-781 , 2005.
[14] C. Yuan, and M. J. Druzdzel, "Importance sampling algorithms for Bayesian networks," Principles and performance, Mathematical and Computer Modelling Vol. 43, pp.1189-1207, 2006.
[15] U. Schlink, S. Dorling, E. Pelikan, G. Nunnari, G. Cawley, H. Junninen, A. Greig, R. Foxall, K. Eben, T. Chatterton, J. Vondracek, M. Richter, M. Dostal, L. Bertucco, M. Kolehmainen, and M. Doyle, "A rigorous inter-comparison of ground-level ozone predictions," Atmospheric Environment Vol.37, 3237-3253, 2003.
[16] S. Kim, "A Neuro-Fuzzy Approach to Integration and Control of Industrial Processes: Part I," KFIS, pp.58-69, 1998.
[17] J. C Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, 1981.
[18] G. E. P. Box, and D. R Cox, "An analysis of transformations," Journal of the Royal Statis-tical Society, Vol. B-26, pp.211-243, 1964.
[19] G. E. P. Box, and G. M. Jenkins, "Time series analysis: Forecasting and control," Holden-Day, San Francisco, 1976.
[20] A. G. Ivakhnenko, "The Group Method of Data Handling in Prediction Problem," Soviet Automatic Control, Vol. 9, No. 6, pp.21-30, 1976.
[21] S. Farlow, Self-Organizing Method in Modeling: GMDH-Type Algorithms, Marcel Deck-ker, New York, 1984.

저 자 소 개

천성표(Seong-Pyo Cheon)
1999년 : 부산대 전기공학과 공학사.
2001년 : 동 대학원 공학석사.
2004년 : LG CNS 근무.
2004년~현재 : 동 대학원 박사과정.

관심분야 : 신경회로망, 베이지안 네트워크, 머신 러닝.
Phone : +82-51-510-2367
Fax : +82-51-513-0212
E-mail : buzz74@pusan.ac.kr

김성신(Sungshin Kim)
1986년 : 연세대 전기공학과 공학석사.
1996년 : Geogia Institute of Tech. 전기공학과 공학박사.
1998년~현재 : 부산대 전자전기통신공학부 부교수.

관심분야 : 지능시스템, 데이터마이닝.
Phone : +82-51-510-2374
Fax : +82-51-513-0212
E-mail : sskim@pusan.ac.kr

이종범(Chong-Bum Lee)
1973년 : 서울대 천문기상학과 이학사.
1976년 : 서울대 기상학과 이학석사.
1985년 : 일본 쓰꾸바대 대기환경학 이학박사.
현재 : 강원대 환경과학과 교수.

관심분야 : 대기 오염 예측, 기상장 모델링.
Phone : +82-33-250-8571
E-mail : cbl@kangwon.ac.kr