한국대기환경학회 2005 추계학술대회 논문집
Proceeding of the 40th Meeting of KOSAE(2005)
Korean Society for Atmospheric Environment

PA38) 미세먼지(PM10)의 통계적 예보모형에 관한 연구 - 포항지역을 중심으로

A Study on Statistical Forecasting Models of PM10

이영섭 · 박종석 · 김현구[1)]
동국대학교 통계학과, [1)]한국에너지기술연구원 풍력연구센터

1. 서 론

대기오염 물질 중에서 최근 주목 받고 있는 원인물질로 눈에 보이지 않는 미세먼지에 대한 관심이 갈수록 높아지고 있다. 먼지 중에서도 크기가 10μm 미만인 PM10은 건강에 치명적인 영향을 미치므로 이에 대한 연구가 활발하게 진행되고 있다. 현시점의 기상자료와 대기오염도 자료를 이용하여 미래의 PM10 최대농도를 예측 또는 예보 할 수 있는 모형구축이 본 연구의 목적으로, 우선 미세먼지 예보를 위한 다양한 통계적 기법을 서로 비교해 보고자 한다. 미세먼지 예측모형에 대한 국내 선행연구로는 구윤서(2004)의 서울, 수도권 지역을 대상으로 한 회귀분석과 신경망 분석을 통한 비교가 있었다. 본 연구에서는 포항시 세 곳의 환경측정소에서의 관측자료를 이용하여 회귀분석, 신경망 분석 뿐만 아니라 최근 각광 받고 있는 새로운 예측모형 기법인 SVR(Support Vector Regression)을 적용시켜 보고 이러한 세 모형의 결과들을 상호비교하여 보았다.

2. 연구 방법

본 연구에서는 2001년 1월 1일부터 2004년 12월 31일까지 4년간 포항시 환경측정소 세 곳(죽도동: KME112, 장흥동: KME113, 대도동: KME114)의 시간별 대기오염자료(SO_2, NO_2, CO, O_3, PM10)와 포항기상대의 기상자료(풍속, 온도, 습도)를 이용하였다. 미세먼지(PM10) 예측모형을 위한 독립(예측)변수로는 기상자료를 사용하였다. 대기오염도 항목은 SO_2, NO_2, CO, O_3 으로 당일 00:00부터 23:00까지의 관측값 중 최대값을, PM10-1(당일 00:00부터 11:00까지의 PM10 중 최대값), PM10-2(당일 12:00부터 23:00까지의 PM10 중 최대값)로 나누어 입력자료로 사용하며, 기상자료 항목은 당일 00:00부터 23:00까지의 풍속 평균값, 온도 최대값, 습도 최소값, 온도의 최대값과 최소값의 차이를 사용하였다.

통계 예측모형은 전통적인 방법인 선형 회귀분석 방법(Regression)과 인공지능기법을 사용한 비선형 방법이면서 정확도가 뛰어난 신경망 분석(Neural Networks, NN) 방법, 그리고 신경망 분석이 과도적합과 신경망 구조의 설계에 많은 시간과 노력이 필요한 단점을 해결하는 방법으로써 많이 쓰이는 SVR 모형을 적용하였다. 회귀모델은 변수들 간의 함수적 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계학적 분석방법 중 하나이다(Draper and Smith, 1998). SVR은 입력공간과 관련된 비선형문제를 고차원의 특징공간의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분석하는 것이 수월할 뿐만 아니라 조정해야 할 파라미터의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다. 그리고 구조적 위험을 최소화함으로써 과대적합 문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 국부적 최적해를 구할 수 있다는 점에서 신경망의 단점을 보완할 수 있는 학습기법으로 알려져 있다(Smola and Schölkopf, 1998). 신경망 모델은 복잡한 구조를 가진 자료에서의 예측(prediction)문제를 해결하기 위해서 사용되는 유연한 비선형 모델(nonlinear models)로, 입력 인자 간, 또는 입력과 출력간의 상호 인과 관계가 불분명한 경우, 서로의 관계를 효율적으로 찾는 특성이 있다. 신경망 분석의 장점은 선형회귀분석에서는 어려운 자료의 비선형적인 관계를 찾아 낼 수 있다는 것이다. 그러나 자료를 과대적합(overfit)하는 경향이 있기 때문에 새로운 자료가 주어졌을 때 적당하지 않을 수 있다는 단점이 있다. 또 다른 단점은 다른 분석과는 달리 모형의 결과를 해석하기가 어렵다는 것이다(Ripley, 1996).

포항지역의 각 측정소에서 측정된 자료를 이용 데이터의 특성을 잘 반영하기 위해 단순임의추출을 하였으

며 데이터의 60%를 모형 구축을 위한 훈련용 데이터(train data)로 사용하였고 나머지 40%를 구축된 모형의 평가를 위한 검증용 데이터(test data)로 사용하였으며 세 기법의 예측값의 정량적인 평가를 위하여 RMSE, CORR, IOA등의 지수를 구하였다. 각 정량 분석 항목의 계산식은 다음과 같으며 P_i: 예측값, O_i: 관측값을 의미한다.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}\ ,\quad IOA = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - O_{mean}| + |O_i - O_{mean}|)^2}\ ,\quad CORR = \frac{N\left(\sum_{i=1}^{N}O_iP_i\right) - \left(\sum_{i=1}^{N}O_i\right)\left(\sum_{i=1}^{N}P_i\right)}{\sqrt{\left[N\left(\sum_{i=1}^{N}O_i^2\right) - \left(\sum_{i=1}^{N}O_i\right)^2\right]\left[N\left(\sum_{i=1}^{N}P_i^2\right) - \left(\sum_{i=1}^{N}P_i\right)^2\right]}}$$

한편 RMSE 값은 적을수록, CORR과 IOA는 1에 가까울수록 좋은 예측력을 가짐을 의미한다.

3. 결과 및 고찰

포항지역의 각 환경측정소별로 세 개의 통계 예측모형을 구축하였다. 구축된 모형을 검증용 데이터에 적용하여 예측값을 구한 후, 이 예측값과 원래의 PM10 값을 이용하여 RMSE, CORR, IOA의 측정지수 값을 구하여 그 결과를 표 1에 나타내었다. 표 1에서 보는 바와 같이 KME112, KME113 측정소에서는 회귀 모델(Regression)의 예측력이 가장 뛰어났으며, KME114에서는 신경망 모델(NN)의 예측력이 가장 뛰어났음을 알 수 있었다. 예상했던 것과는 달리 SVR이 근소한 차이로 예측력이 떨어짐을 알 수 있었다. 어떠한 데이터에서나 항상 우수한 통계 모형은 있을 수 없으며 데이터의 특성에 따라 적용결과가 다를 수 있으며, 독립변수의 개수나 속성을 변화시켜 봄에 따라 예측력의 변동이 있을 수 있다. 따라서 본 연구는 지금까지 많이 사용하지 않은 SVR 모형을 대기오염 예측에 적용하여 그 성능을 비교 분석하는데 의의를 두고자 한다. 또한 본 연구에서 사용한 통계적 예측모형이 아니라 시계열 분석 기법을 사용하여 PM10 예측모형을 구축하고 그 정확도를 본 연구에서 사용한 통계적 기법들과 비교하여 보는 것을 향후과제로 남겨두고자 한다.

Table 1. 각 기법에 따른 예측값의 정량적 평가

모형기법	평가방법	환경측정소		
		KME112	KME113	KME114
Regression	RMSE	72.6846	63.2651	64.7749
	CORR	0.5053	0.6538	0.3862
	IOA	0.6074	0.7289	0.4690
SVR	RMSE	73.5792	72.8580	65.9821
	CORR	0.4866	0.5165	0.3645
	IOA	0.4834	0.6092	0.3997
NN	RMSE	72.8694	70.2838	63.5025
	CORR	0.5019	0.5347	0.4261
	IOA	0.6132	0.6241	0.4886

참 고 문 헌

구윤서 (2004) 미세먼지 예보시스템의 도입방안.

Draper, N. and Smith, H. (1998) Applied Regression Analysis(3rd ed) New York, Wiley.

Smola, A.J. and Schölkopf, B. (1998) A Tutorial on Support Vector Regression, Royal Holloway College, U.K, Neuro COLT Tech. Rep. TR-1998-030.

Ripley, B.D. (1996), Pattern Recognition and Neural Networks, Cam-bridge University Press.