The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.1.083
28(1), 83–92

Particulate Matter Prediction using Quantile Boosting

Jun-Hyeon Kwon[a] · Yaeji Lim[b,1] · Hee-Seok Oh[a]

[a]Department of Statistics, Seoul National University; [b]Samsung Medical Center

(Received December 2, 2014; Revised January 19, 2015; Accepted February 4, 2015)

Abstract

Concerning the national health, it is important to develop an accurate prediction method of atmospheric particulate matter (PM) because being exposed to such fine dust can trigger not only respiratory diseases as well as dermatoses, ophthalmopathies and cardiovascular diseases. The National Institute of Environmental Research (NIER) employs a decision tree to predict bad weather days with a high PM concentration. However, the decision tree method (even with the inherent unstableness) cannot be a suitable model to predict bad weather days which represent only 4% of the entire data. In this paper, while presenting the inaccuracy and inappropriateness of the method used by the NIER, we present the utility of a new prediction model which adopts boosting with quantile loss functions. We evaluate the performance of the new method over various τ-value's and justify the proposed method through comparison.

Keywords: Boosting, particulate matter, prediction, quantile regression.

1. 서론

최근 언론에 의해 미세먼지에 대한 연구결과가 알려지면서 그에 대한 국민들의 경각심이 높아지고 있다. 분진(TPM; total particulate matter)이란 100μm 이하의 입경을 가진 대기 중 부유물질을 일컫는데, 그중 입경이 10μm 이하인 것을 미세먼지(PM_{10}), 2.5μm 이하인 것을 초미세먼지($PM_{2.5}$)로 분류한다. 알려진 바에 의하면 고농도의 미세먼지에 노출되었을 경우 호흡기는 물론이고, 심혈관계, 안구, 피부 등에도 질병이 생길 확률이 높아진다고 한다. 또한 세계보건기구(WHO; World Heath Organization)의 2006년 보고서 WHO (2006)에 따르면 미세먼지 농도가 10$\mu g/m^3$ 증가할 때마다 일일사망률이 0.5%씩 증가한다는 연구결과도 있다. 하지만 높아진 관심은 미세먼지 농도 예측을 담당하는 관계기관으로 이어졌고 잦은 오보에 따라 비난도 거세져왔다.

현재 미세먼지 예보는 환경부 소속의 국립환경과학원에서 담당하고 있는데, 예측 모형으로 Koo 등 (2010)에서 제시한 모형을 사용하고 있다. Koo 등 (2010)에서는 신경망 모형, 의사결정나무 모형, 중회귀분석 모형을 함께 사용하고 있는데 각각의 모형으로 예측치를 계산한 후 그에 따라 채택되는 통합대기

This work was supported by a grant from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP)(2012002717) and Basic Science Reserch Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning(2014R1A1A3049447).

[1]Corresponding author: Samsung Medical Center, 50 Irwon-dong, Gangnam-gu, Seoul, South Korea 135-710. E-mail: yaeji.lim@gmail.com

환경지수 중 가장 빈도가 높은 지수를 이용하여 최종 예보치를 결정한다. 신경망 모형은 인간의 두뇌 구조를 모방한 지도학습법으로 입력값과 출력값이 비선형적으로 복잡하게 연관되어 있는 것이 특징이다. 신경망 모형은 예측력은 좋지만 해석이 어렵다는 단점이 있다 (Park 등, 2013). 의사결정나무 모형 또한 비선형 구조를 가진 지도학습법이지만 신경망 모형과는 달리 해석이 용이한 반면, 예측력이 떨어지고 불안정하다는 단점이 있다 (Hastie 등, 2001). 마지막으로 중회귀분석 모형은 통계학에서 널리 쓰이고 있는 기법으로서 독립변수들과 종속변수 사이 관계를 선형적인 형태로 파악하여 조사하고 모형화하는 기법이다 (Montgomery 등, 2012).

현 예측모형은 세 기법을 이용하여 서로의 단점을 보완하고 있다. 하지만 각 기법의 기저에서는 미세먼지 농도의 평균(기댓값)에 초점을 맞추고 있어서 미세먼지 자료에서 약 4%정도의 비율로 나타나는 "나쁜날씨(미세먼지 농도가 $100\mu g/m^3$를 넘는 날씨)"를 예측하기에는 부적합하다고 할 수 있다. 더군다나 고농도 미세먼지의 경우 노약자나 호흡기질환자, 심질환자 등에게는 치명적인 영향을 주어 큰 사회적 비용을 발생시킬 가능성이 있으므로 미세먼지 농도의 과소예측은 과대예측에 비해 더욱 피해야 한다. 따라서 과소예측과 과대예측의 비용 차이를 고려하여 "나쁜 날씨"는 최대한 정확히 예측해서 미리 발표해야 할 필요가 있다.

본 논문에서는 미세먼지 농도의 예보 적중률을 높이기 위해 주어진 자료를 보다 효과적인 통계 모형을 이용하여 개선하는 데 초점을 맞추었다. 특히 내일예보모형에 사용되는 신경망, 의사결정나무, 중회귀분석 모형 중 의사결정나무의 단점을 보완하고 "나쁜날씨"에 대한 예측력을 높이기 위해 분위수 부스팅 모형을 제시했다. 논문의 전개는 다음과 같다. 우선 2절에서는 분석에 사용한 자료에 대한 설명 및 분위수 부스팅의 이론과 구현 방법 등에 대해 다루었다. 3절에서는 기존 나무모형과 분위수 부스팅 모형을 실제 자료에 적용하여 비교하였고, 4절에서는 결론 및 발전 방안을 제시하였다.

2. 분석 방법

2.1. 자료 기술

본 논문에서는 2011년 1월부터 2014년 6월까지 약 1300일 동안 서울시 종로구에서 관측한 미세먼지 자료를 사용하였다. 이 중 결측치를 제외한 처음 1092일은 학습집합으로, 나중 180일은 테스트집합으로 사용하였다. 예측변수로는 대기오염 물질 중에선 PM_{10} 농도와 O_3, NO_2, CO, SO_2 등을, 기상 요소 중에선 평균기온, 최고기온, 최저기온, 평균풍속, 최대풍속, 순간최대풍속, 기압, 수증기압, 일조량, 운량, 강우, 습도, 황사 여부 등을 사용했다. 반응변수는 익일의 PM_{10} 농도에 따라 "나쁜날씨"면 1로, 그렇지 않으면 0으로 설정했다. 그리고 테스트집합을 이용해 국립환경과학원에서 미세먼지 예보에 사용하는 방법 중 하나인 의사결정나무 모형과 본 논문에서 제안하고자 하는 방법인 cost-sensitive stochastic gradient boosting의 성능을 확인하고 서로 비교하여 기존 모형 개선의 타당성을 확인하였다.

2.2. 반응변수 설정

최근 가장 이슈가 되는 것은 $PM_{2.5}$이지만 현재 대부분의 역학 연구에 쓰이는 분진 노출 지표는 전세계적으로 더 많은 관측 자료가 있는 PM_{10} 농도다. PM_{10}을 반응변수로 설정해도 크게 문제되지 않는 이유는 PM_{10}은 $PM_{2.5}$을 포함한 부유물질을 일컫는 용어이고, PM_{10}과 $PM_{2.5}$의 농도비는 지역마다 차이는 있지만 대략 2 : 1을 유지하기 때문이다. 우리나라의 경우도 PM_{10} 농도 관측 자료를 보다 많이 구할 수 있기 때문에 국립환경과학원에서는 PM_{10} 농도를 반응변수로 사용한다. 따라서 기존 모형과의 비교 연구가 필요한 본 논문에서는 반응 변수로 PM_{10} 농도를 사용하기로 한다.

Table 2.1. Contingency table of predicted and observed values

		predicted class	
		$< 100\mu g/m^3$	$\geq 100\mu g/m^3$
observed class	$< 100\mu g/m^3$	a	b
	$\geq 100\mu g/m^3$	c	d

2.3. 예측모형 평가 기준

본 논문에서는 기존 방법 중에 하나인 의사결정나무 모형과의 비교를 위해 PM_{10}의 농도를 $100\mu g/m^3$을 기준으로 분류했다.

예측모형의 성능을 평가하기 위한 측도로 Table 2.1에서 정의된 a, b, c, d를 이용해 오분류율, 민감도, 특이도를 다음과 같이 정의한다.

- 오분류율(misclassification rate): $(b+c)/(a+b+c+d)$, 전체 측정일 중 예측 분류가 틀린 날의 비율.
- 민감도(sensitivity): $d/(c+d)$, 미세먼지 농도가 $100\mu g/m^3$이상으로 관측된 날 중에서 예측 농도가 $100\mu g/m^3$이상인 날의 비율.
- 특이도(specificity): $a/(a+b)$, 미세먼지 농도가 $100\mu g/m^3$미만으로 관측된 날 중에서 예측 농도가 $100\mu g/m^3$미만인 날의 비율.

2.4. 기존 예측방법

의사결정나무 모형은 변수들의 영역을 격자 형태로 나눈 후 각 구획마다 간단한 적합을 하는 방법이다 (Hastie 등, 2001). 나무 모형은 성장, 가지치기 등의 과정을 거쳐 만들어진다. If-then 형식의 규칙을 사용하기 때문에 앞서 언급했듯이 해석력이 좋고, 분류 문제에 적합하고, 모형에 대한 가정이 필요 없는 방법이라는 것이 장점이다 (Park 등, 2013). 의사결정나무 모형의 아이디어가 제시된 이후 CART, C4.5, CHAID 등 여러 종류의 나무모형들이 개발되었다. 현재 국립환경과학원에서 사용하고 있는 모형은 CART(classification and regresstion tree) 모형인데 이는 회귀와 분류 문제를 모두 다룰 수 있는 모형이다.

하지만 의사결정나무는 모형 자체의 불안정성, 예측함수가 매끄럽지 못하여 분류경계가 사각형이 아닌 경우엔 좋지 않은 결과를 줄 수 도 있는 점, 가법적 구조를 찾아내기 어려운 점 등의 단점 때문에 점차 외면 받아 왔다 (Hastie 등, 2001). 또한 미세먼지 농도 문제에 관해서는 과소예측과 과대예측의 비용 차이까지 고려해야 하기 때문에 나무모형은 더욱 적절치 못하다.

Figure 2.1은 국립환경과학원에서 사용하고 있는 미세먼지 예보를 위한 나무모형이다. 수집할 수 있는 자료가 허용되는 한 R 패키지를 이용해 Figure 2.1의 나무 모형과 최대한 비슷하게 재구현 했다. 이 때 나무모형의 끝마디에 있는 "모델로 예측"은 어떤 모형을 이용했는지가 명확하지 않아 R 패키지 ridge의 능형회귀모형을 이용했다. 능형회귀의 제한모수인 λ 값은 패키지 내부에서 자동으로 설정하게 했고, 반응변수인 y의 왜도 문제를 해소하기 위해 로그변환을 했다. 이와 같은 과정을 통해 테스트 집합에서 나무모형의 오분류율과 민감도를 계산하였다.

2.5. 분위수 부스팅

Gradient boosting이란 기존의 부스팅 알고리즘에 최급강하법(steepest descent)을 적용하여 최적화를 한 일반화가법모형이다. 그리고 gradient boosting 알고리즘의 매 반복에서 잔차를 적합하기 위해

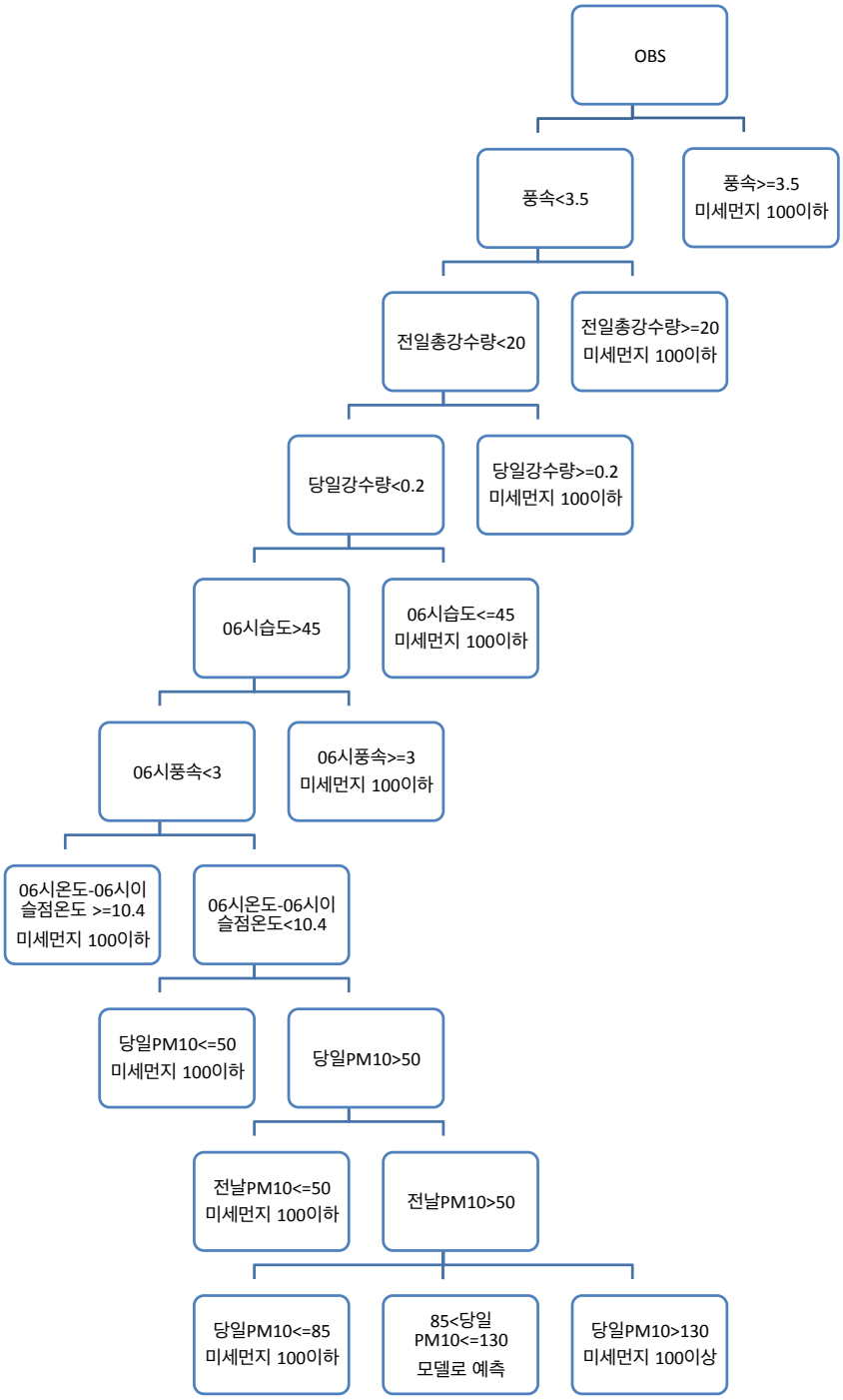


Figure 2.1. Decision tree model used in NIER

약한학습기를 만들 때마다 전체 표본에서 비복원추출로 부표본을 얻어 계산을 한다면 이는 stochastic gradient boosting이 된다 (Friedman, 2002). Stochastic gradient boosting의 구체적인 알고리즘은 다음과 같다.

1. 주어진 손실함수 L에 대해 $\gamma_0 = \arg\min_\gamma \sum_{i=1}^N \mathrm{L}(y_i, \gamma)$를 초기값으로 나무모형 f_0를 설정한다.
2. 반복횟수 m $(m = 1, \ldots, M)$에 대해
   (a) $i = 1, 2, \ldots, N$에 대하여 다음을 계산한다.
   $$r_{im} = -\left[\frac{\partial \mathrm{L}(y_i, \mathrm{f}(x_i))}{\partial \mathrm{f}(x_i)}\right]_{\mathrm{f}=\mathrm{f}_{m-1}}.$$
   (b) N개의 관측값 중에서 비복원추출로 N'개의 부표본을 얻는다.
   (c) 부표본을 이용해 r_{im}을 끝마디 영역 R_{jm} $(j = 1, 2, \ldots, J_m)$을 갖는 나무모형으로 적합한다.
   (d) $j = 1, 2, \ldots, J_m$에 대해 다음을 계산한다.
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in \mathrm{R}_{jm}} \mathrm{L}(y_i, \mathrm{f}_{m-1}(x_i) + \gamma).$$
   (e) f를 다음과 같이 갱신한다. 여기서 λ는 학습률이다.
   $$\mathrm{f}_m(x) = \mathrm{f}_{m-1}(x) + \lambda \sum_{j=1}^{J_m} \gamma_{jm} \mathrm{I}(x \in \mathrm{R}_{jm}).$$
3. 결과값으로 $\hat{\mathrm{f}}(x) = \mathrm{f}_M(x)$를 출력한다.

Kriegler와 Berk (2010)은 미국 로스앤젤레스 내의 노숙자 수를 추정하기 위해 cost-sensitive stochastic gradient boosting을 제안하였다. 그들은 노숙자 쉼터를 지을 때 쉼터의 수용인원을 수요인원보다 적게 설계하는 것은 많게 설계하는 것보다 더 커다란 문제를 야기시킨다는 점에 착안해서 이 방법을 고안하였다. 비용의존(cost-sensitve)이란 예측오차에 수반되는 비용의 부호에 따른 비대칭성을 고려함을 의미한다. Cost-sensitive stochastic gradient boosting의 손실함수는 기존의 라플라스 손실함수 $(1/\sum w_i)\sum w_i|y_i - f(x_i)|$에 비용의존을 조절하는 모수 $\tau \in (0, 1)$을 결합하여 다음과 같이 표현한다.

$$\mathrm{L}(y_i, \mathrm{f}(x_i)) = \frac{1}{\sum w_i}\left(\tau \sum_{y_i > \mathrm{f}(x_i)} w_i|y_i - \mathrm{f}(x_i)| + (1-\tau)\sum_{y_i \le \mathrm{f}(x_i)} w_i|\mathrm{f}(x_i) - y_i|\right).$$

즉, $y > f\,(x)$인 부분과 $y \le f(x)$인 부분으로 나누고 각각 τ와 $1-\tau$의 가중을 주어 새롭게 손실함수를 정의한다.

본 논문에서는 미세먼지 농도를 예측하기 위해 cost-sensitive stochastic gradient boosting을 이용할 것을 제안하고 추후 연구에서 이 모형의 확장적 사용을 고려하여 이를 분위수 부스팅이라 새로 명명하고자 한다. Cost-sensitive stochastic gradient boosting은 매 반복마다 정해진 모수 τ 값에 따라 나무모형이 만들어진다. 하지만 이는 직전 모형의 잔차에 초점을 맞추었을 뿐, 과대예측과 과소예측의 실제적인 비용차를 반영하지 못한다. 모수 τ 값은, 1에 가까워지면 기댓값보다 점점 작아지는 값을, 0에 가까워지면 기댓값보다 점점 더 커지는 값을 예측할 수 있다는 점에서 그 역할이 분위수 분석에서 분위수의

Table 3.1. Misclassification rates and sensitivities of decision tree and quantile boosting

Decision tree		Misclassification rate	Sensitivity
		5.0%	38.5%
Quantile boosting	$\tau = 0.50$	5.6%	23.1%
	$\tau = 0.55$	5.6%	23.1%
	$\tau = 0.60$	4.5%	38.5%
	$\tau = 0.65$	4.5%	46.2%
	$\tau = 0.70$	5.6%	53.9%
	$\tau = 0.75$	6.7%	69.2%
	$\tau = 0.80$	7.8%	69.2%
	$\tau = 0.85$	11.1%	69.2%
	$\tau = 0.90$	12.8%	69.2%
	$\tau = 0.95$	13.3%	76.9%

역할과 유사하다. 따라서 본 논문에서는 cost-sensitive stochastic gradient boosting이라는 이름 대신 분위수 부스팅이라는 새로운 이름을 쓰고 τ 값에 따른 추정치의 변화를 분석하였다.

분위수 부스팅의 알고리즘을 실제 자료에 적용하기 위해 R 패키지 gbm을 사용했는데, 이를 위해 아래의 요인들에 대한 부가적인 설정이 필요하다.

- τ 값: 우선 τ 값을 과소예측과 과대예측에 따르는 실제 비용의 비를 나타내기 위해 사용하는 대신 주어진 예측 모형의 오분류율과 민감도를 조정하기 위한 척도로 사용한다. τ 값을 0.05단위로 0.50부터 0.95까지 모두 10개로 나누어 서로 비교한다. 이 때, τ 값이 커지면서 예측값도 함께 커지는 경향을 보이기 때문에 앞서 언급했듯이 여기서 비용의존 추정 방법과 분위수 추정 방법 사이의 접점을 찾을 수 있다.
- 교차검증의 반복수: 최적의 반복수는 교차검증(cross-validation)의 평균오차를 이용해서 패키지 내의 함수로 얻는다. 본 논문의 분석에서는 교차검증의 적절한 분할수를 5로 정하였다. 분할수를 10으로 정하더라도 결과엔 큰 차이가 없었다.
- 학습률: 경험적으로 학습률은 작을수록 더 작은 오차를 제공한다고 알려져 있다 (Friedman, 2002). 본 분석에서는 알고리즘의 수렴속도를 고려해서 학습률을 0.01로 설정했다. 참고로 이 값보다 작은 0.001을 고려했을 때 그 성능의 차이가 없었다.
- 교호작용: 본 분석에서는 교호작용은 고려하지 않았다. 만약 교호작용을 고려한다면 과적합의 문제도 생길 수 있을뿐더러 실제 테스트집합에 적용했을 때 성능에서 차이가 없거나 어떤 경우는 오히려 민감도를 낮추었기 때문이다.
- 결과물 산출: stochastic gradient boosting은 그 특성상 나무모형을 만들 때마다 부분표본을 뽑기 때문에 매번 약간씩 다른 결과를 보여준다. 따라서 예측을 각 τ 값마다 10개씩 만든 후 예측값들을 평균 내어 $100\mu g/m^3$을 기준으로 0과 1로 분류하였다.

3. 결과 분석

Table 3.1은 의사결정나무 모형과 제안한 분위수 부스팅의 오분류율과 민감도 결과를 제공한다. 먼저 기존 의사결정나무 모형의 오분류율은 5% (9/180)로 작지만 "나쁜 날씨"에 대한 민감도는 38.5% (5/13)였다. 이 때 주목해야 할 점은 오분류된 날짜는 전체 9일이었는데 이 중 8일이 실제 "나쁜 날

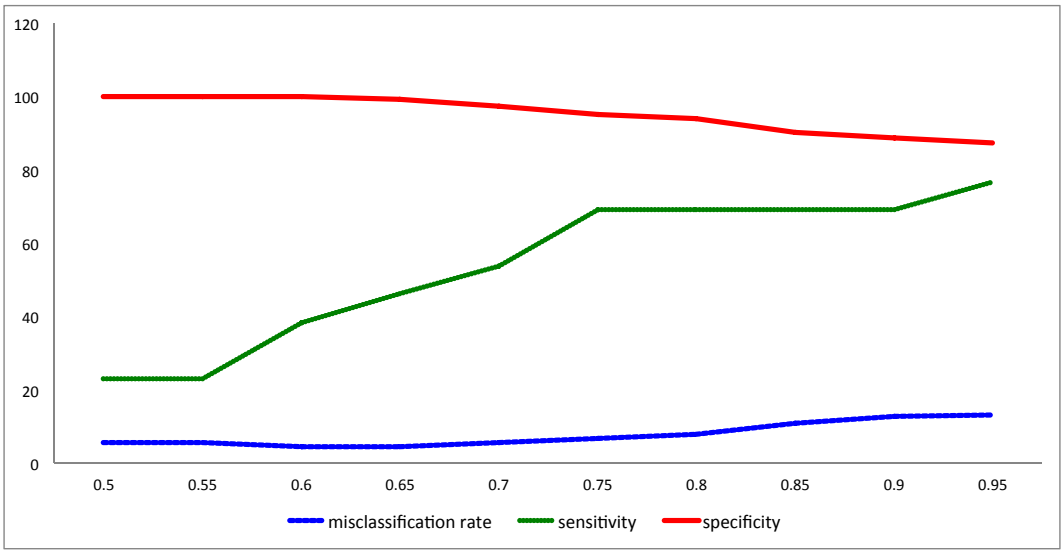


Figure 3.1. Misclassification rates and sensitivities over τ

씨"에 해당된다는 것이다. 한편 분위수 부스팅을 테스트집합에 적용해 본 결과 τ 값이 높아짐에 따라 전체 오분류율은 높아지는 대신 민감도가 높아져 "나쁜날씨"를 잘못 분류하는 비율 또한 낮아짐을 알 수 있다.

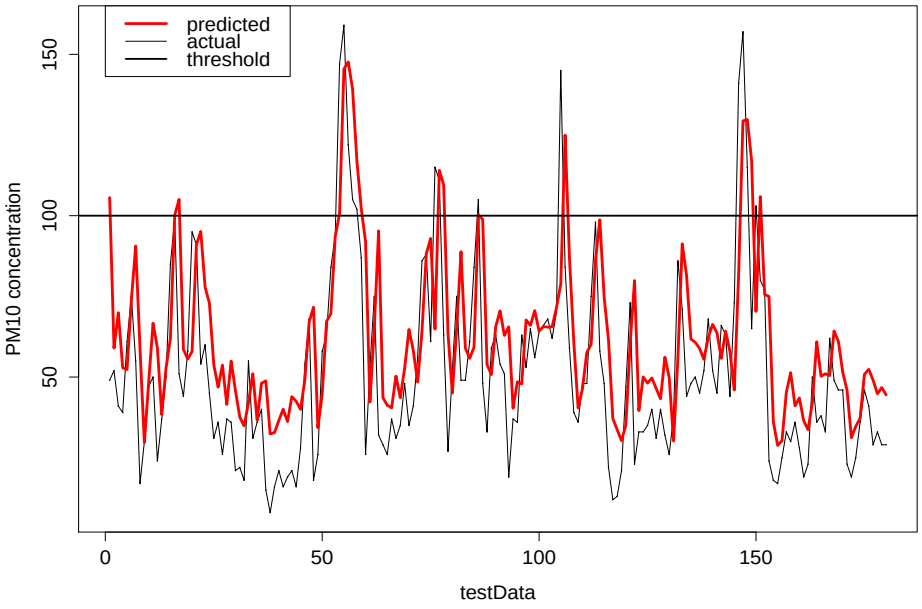
Figure 3.1은 τ 값에 따른 분위수 부스팅의 오분류율, 민감도 그리고 특이도의 추세를 나타내고 있다. Table 3.1과 Figure 3.1로부터 알 수 있듯이 $\tau = 0.75$일 때 분위수 부스팅은 오분류율을 기존 모형과 비슷하게 유지하면서 민감도를 약 두 배 가량 증가시킨다.

위의 결과로 볼 때 $\tau = 0.75$로 설정된 분위수 부스팅을 사용하는 것이 좋은 선택임을 알 수 있다. 이 때의 상황을 보다 자세히 살펴보기 위해 Figure 3.2에서 실제 관측값과 예측값의 시계열도와 산점도를 그려보았다. Figure 3.2(a)를 보면 분위수 부스팅에 의한 예측결과는 전반적으로 실제값보다 높음을 알 수 있다. 또한 Figure 3.2(b)에서 볼 수 있듯이 대부분의 관측값은 "나쁜날씨"의 기준치인 $100\mu g/m^3$ 보다 다소 낮은 곳에서 형성되어 오분류율 증가에 기여하지 않은 반면, 미세먼지 농도가 특이하게 높은 날에 대해선 약간만 과소된 값을 알려주어 민감도 증가에 기여하고 있다. 마지막으로 "나쁜날씨" 예측 실패는 실제 관측값이 기준치 부근에 있을 경우에 일어남을 알 수 있다. 본 논문에 쓰인 자료의 수가 적어 직접 확인할 수 없었지만, 예측결과값에 따른 "나쁜날씨" 판단 기준치를 적절히 조정함으로써 이를 해결할 수 있을 것으로 예상한다.

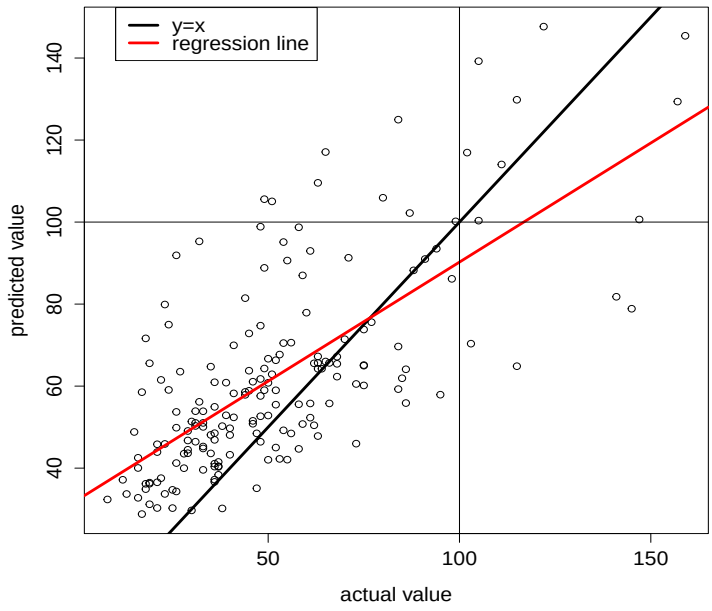
4. 결론

본 논문에서는 고농도 미세먼지를 효과적으로 예측하기 위해 분위수 부스팅 방법을 그 알고리즘과 함께 소개하였다. 또한 위의 방법론을 실제 서울시 종로구의 미세먼지 자료에 적용하고, 기존에 현업에서 사용하고 있는 모형 중 하나인 의사결정나무모형과 비교분석하였다. 예상한대로 제안한 방법론은 고농도 미세먼지 예측에 우수한 성능을 보임을 확인할 수 있었다.

한편 Koo 등 (2010)에서 제시한 평가치를 따르더라도 분위수 부스팅은 현업에서 사용하고 있는 모형



(a) Time series plot



(b) observed value vs. predicted value plot

Figure 3.2. Comparison between predicted and observed values when $\tau = 0.75$

에 비해 우수한 결과를 보여주었다. Koo 등 (2010)에서는 PM_{10} 농도를 구간에 따라 "좋음", "보통", "민감군 영향", "나쁨", "아주 나쁨"으로 분류하여 평가치를 정의하였다. 그중 예측정확도는 관측된 통합대기환경지수와 예측된 통합대기환경지수가 일치하는 날이 전체 관측된 날에서 차지하는 비율로 계산되었다. 감지확률은 본 논문의 민감도와 동일하게 계산되었는데 분류기준을 $80\mu g/m^3$으로 했다는 점에서 차이가 있다. 이러한 차이 뿐 아니라 테스트 집합이 지나치게 적어 분위수 부스팅과의 직접적인 비교가 어려웠지만, 이 기준에 맞춰 분위수 부스팅의 예측정확도와 감지확률을 계산한 결과, τ 값이 증가함에 따라 예측정확도는 감소하는 경향을 보였고 감지확률은 증가하는 경향을 보였다. 특히 $\tau = 0.60$일 때 예측정확도가 68.9%, 감지확률이 42.3%였고, $\tau = 0.80$일 때는 예측정확도가 58.9%, 감지확률이 69.2%였다. 이를 통해 분위수 부스팅은 예측 용도에 따라 예측정확도와 감지확률의 절충이 가능함을 확인할 수도 있었다. 또한 현업에서 사용하는 내일예보모형의 의사결정나무 모형과 $\tau = 0.60$일 때의 분위수 부스팅을 비교하였을 때, 제안한 모형은 예측정확도와 감지확률을 각각 2.6%p, 68.9%p 높일 수 있음을 확인할 수도 있었다.

자료와 관련되어서 결측치 문제 때문에 수집할 수 있는 자료가 한정적이어서 연구를 하면서 많은 어려움이 있었다. 보다 많은 자료를 축적할 수 있게 된다면 교차검증을 할 때 계절성이 서로 상쇄되어 믿을 수 있는 교차검증 오차를 알 수 있게 될 뿐만 아니라 분위수 부스팅을 적용할 때 필요한 세부설정도 보다 최적화된 값으로 조정할 수 있게 될 것이다. 또한 테스트 집합도 크게 잡을 수 있게 되어 오분류율과 민감도가 τ 값에 따라 어떤 변화를 하는지 정밀하게 관찰할 수 있으리라 기대한다.

마지막으로 분위수 부스팅을 활용한 모형은 τ 값이 증가함에 따라 미세먼지 농도가 높은 날을 더 잘 예측하게 되는 대신 전체 테스트 집합에 대해선 오분류를 더 많이 하였다. 그러나 비교적 0.5에 가까운 τ에 대해선 오분류율을 의사결정나무 모형과 비슷하게 유지하면서도 민감도를 높여 성능을 높일 수 있었다는 점에서 본 연구의 의의를 찾을 수 있었다. 전체 4%에 불과한 "나쁜날씨"를 예측해야 함에도 불구하고 적절한 τ 값을 설정한다면 보다 효율적인 예보가 가능해지는 것이다. 이는 분위수 모형이 기존 모형을 대체하기에 충분한 근거가 되리라 판단한다.

References

Friedman, J. H. (2002). Stochastic gradient boosting, *Computational Statistics and Data Analysis*, **38**, 367–378.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction : with 200 Full-Color Illustrations*, Springer, New York.

Koo, Y. S., Yun, H. Y., Kwon, H. Y. and Yu, S. H. (2010). A development of pm10 forecasting system, *Journal of Korean Society for Atmospheric Environment*, **26**, 666–682.

Kriegler, B. and Berk, R. (2010). Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting, *The Annals of Applied Statistics*, **4**, 1234–1255.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, John Wiley & Sons, Hoboken, New Jersey.

Park, C. Y., Kim, Y. D., Kim, J. S., Song, J. W. and Choi, H. S. (2013). *Data Mining with R*, KyoWooSa, Seoul.

WHO (2006). *Air Quality Guidelines-2005 Global Updates*, World Health Organization.

분위수 부스팅을 이용한 미세먼지 농도 예측

권준현[a] · 임예지[b,1] · 오희석[a]

[a]서울대학교 통계학과, [b]삼성의료원

(2014년 12월 2일 접수, 2015년 1월 19일 수정, 2015년 2월 4일 채택)

요 약

고농도 미세먼지(PM_{10})에 노출되는 것은 호흡기 질환 뿐만 아니라 피부, 안구, 심혈관계 질환 등을 야기한다. 따라서 미세먼지 농도를 정확히 예측하는 방법을 개발하는 것은 국민건강과도 깊은 관련이 있다. 현재 국립환경과학원에서는 미세먼지 농도가 높은 "나쁜날씨"를 예측하기 위해 의사결정나무 모형을 사용하고 있다. 그러나 모형 자체의 불안정성은 차치하더라도 의사결정나무는 전체 데이터의 9%밖에 차지하지 않는 "나쁜날씨"를 예측하기에 적합하지 못하다. 본 논문에서는 국립환경과학원에서 사용하는 모형의 부정확성과 부적절성을 제시하는 한편, 분위수 손실 함수를 적용한 새로운 모형의 유용성을 제시한다. 그리고 새로운 모형의 성능을 여러 τ 값에 대해 평가하고 비교를 통해 기존 모형 교체의 타당성을 보인다.

주요용어: 미세먼지, 부스팅, 분위 회귀분석, 예측.

[1]교신저자: (135-710) 서울시 강남구 일원동 50, 삼성의료원. E-mail: yaeji.lim@gmail.com