

# Domain Adaptation 방법을 이용한 기계학습 기반의 미세먼지 농도 예측

강태천<sup>†</sup>, 강행봉<sup>\*\*</sup>

## Machine Learning-based Estimation of the Concentration of Fine Particulate Matter Using Domain Adaptation Method

Tae-Cheon Kang<sup>†</sup>, Hang-Bong Kang<sup>\*\*</sup>

### ABSTRACT

Recently, people's attention and worries about fine particulate matter have been increasing. Due to the construction and maintenance costs, there are insufficient air quality monitoring stations. As a result, people have limited information about the concentration of fine particulate matter, depending on the location. Studies have been undertaken to estimate the fine particle concentrations in areas without a measurement station. Yet there are limitations in that the estimate cannot take account of other factors that affect the concentration of fine particle. In order to solve these problems, we propose a framework for estimating the concentration of fine particulate matter of a specific area using meteorological data and traffic data. Since there are more grids without a monitor station than grids with a monitor station, we used a domain adversarial neural network based on the domain adaptation method. The features extracted from meteorological data and traffic data are learned in the network, and the air quality index of the corresponding area is then predicted by the generated model. Experimental results demonstrate that the proposed method performs better as the number of source data increases than the method using conditional random fields.

**Key words:** Fine Particulate Matter, Domain Adaptation, Big Data, Domain Adversarial Neural Network

### 1. 서 론

대기 오염은 넓은 지역에 누적적인 피해를 입히는 점에서 시급히 해결해야 할 환경 문제 중 하나이고, 그 위험성은 지난 1952년 런던 스모그 사태 이후 꾸준히 보고되어 왔다. 대기 오염 문제 중 하나인 미세먼지에 관한 연구 역시 활발히 이루어지고 있고

[1-4, 16], 미세먼지가 2013년 WHO에 의해 1급 발암 물질로 분류됨에 따라 사람들의 관심과 민감도가 급격하게 증가하고 있다. 미세먼지는 사람의 신체에 상당한 악영향을 끼치는데, 지난 2015년 중국 베이징에서 열린 마라톤대회에서 스모그로 인한 심장마비로 7명이 사망했으며 미세먼지 등의 대기오염으로 인해 우리나라 서울, 경기지역에서 한해 30세 이상 성인

※ Corresponding Author : Hang-Bong Kang, Address: 43 Jibong-ro, Bucheon-si, Gyeonggi-do, (14662), Republic of Korea, TEL : +82-10-3301-7298, FAX : +82-2-2164-4945, E-mail : hbkang@catholic.ac.kr

Receipt date : Jul. 19, 2017, Approval date : Jul. 27, 2017

<sup>†</sup> Dept. of Digital Media, Graduate School, Catholic University of Korea (E-mail : kstky@gmail.com)

<sup>\*\*</sup> Dept. of Digital Media, Graduate School, Catholic University of Korea

※ This research was supported by the National Research Foundation of Korea(NRF) Grant funded by the Ministry of Science, ICT & Future Planning for convergent research in Development program for convergence R&D over Science and Technology Liberal Arts (NRF-2016 M3C1B6929717), and also supported by Basic Science Research Program through the NRF of Korea funded by the Ministry of Science, ICT and future Planning (No. 2015R1A2A1A10056304).

15,000여명이 기대수명을 채우지 못한다는 연구결과가 나오기도 했다[5]. 실제로 사람들은 미세먼지 농도가 높은 날은 외출을 자제하고 미세먼지 농도가 낮은 지역으로 이사를 가기도 하는 등 미세먼지는 우리 삶에 직접적인 영향을 끼치고 있다.

미세먼지 자체가 갖는 사회, 경제적 문제에 더해 서 국민들의 미세먼지 측정치에 대한 불신 역시 해결해야 할 문제 중 하나이다. 이러한 문제를 해결하기 위한 방법 중 하나로 고속 자료동화 기법을 이용해 국내 대기질 예보의 성능과 속도를 향상한 연구[16]가 있으나, 측정소가 없는 장소의 미세먼지 농도를 추정하는 데에는 어려움이 따른다. 미세먼지 예방과 예보에 중요한 역할을 하는 미세먼지 측정소는 설치 및 유지비용으로 인해 그 수가 제한되어 있다. 면적 53.44 km<sup>2</sup>의 경기도 부천시는 5개의 미세먼지 측정소를 운용 중인데, 미세먼지 농도가 갖는 약간의 시간적, 공간적 종속성만을 이용해 측정소가 없는 구역의 미세먼지 농도를 추정하는 것에는 한계가 있다.

본 논문에서는 각 기관으로부터 얻은 데이터를 이용해 측정소가 없는 지역의 미세먼지 농도에 따른 label을 추정하고자 한다. 미세먼지 측정소의 위치와 측정치를 이용해 다른 구역의 미세먼지 농도를 추정하는 방법으로는 보간법(interpolation)을 이용한 방법과 도시 영상을 이용한 방법[6]이 있는데, 이들은 미세먼지 농도에 영향을 끼치는 다른 여러 요소를 고려하지 못한다. 도시 대기에 영향을 끼치는 요소는 교통 흐름, 기상 데이터 등 여러 가지가 있으며, 모든 요소를 고려하여 특정 구역의 미세먼지 농도를 추정하는 데에는 많은 어려움이 따른다. 각종 도시 데이터를 이용해 미세먼지 농도를 측정하는 연구 역시 진행되어 왔으나 많은 경험적 가정을 기반으로 한다는 점에서 모든 도시 상황에 적용할 수 없다는 단점이 있다[7]. 도시 데이터의 공간적, 시간적 특성에 맞는 기계학습 알고리즘을 적용해 미세먼지 농도를 추정한 연구[2]는 추정한 결과를 다시 학습에 사용한다는 점에서 추정이 잘못되었을 때 오류가 누적될 가능성이 있다. 또한 획득할 수 있는 데이터가 다르기 때문에 우리나라의 다른 도시에 직접 적용하기에는 한계가 있다.

본 논문에서는 부천시를 작은 구역으로 나누고, 각 구역이 갖는 데이터로부터 특징을 추출하는 방법

을 제안한다. 추출한 특징을 domain adaptation[8-10] 기반의 Domain Adversarial Neural Network [11]를 이용해 학습함으로써 각 특징에 대해 실제 측정소로부터 측정된 값(label이 있는 데이터)이 부족한 문제 상황을 해결하고자 한다. 학습된 모델을 이용해 각 구역의 시간별 PM<sub>10</sub> 농도를 추정한 결과 학습에 사용한 다른 오염물질에 대한 데이터가 증가함에 따라 다른 방법보다 나은 성능을 보였으며, 제안한 방법은 특정 도시만이 갖는 특징을 사용하지 않기 때문에 동일한 추론 모델을 다른 여러 도시에 적용할 수 있는 장점이 있다.

본 논문의 구성은 다음과 같다. 2장은 제안하는 프레임워크의 개요를 설명하였고, 3장에서는 데이터를 획득하는 방법과 획득한 데이터로부터 학습에 사용할 특징을 추출하는 방법을 서술한다. 4장에서 domain adaptation 기반의 학습 알고리즘을 소개하고, 학습을 통해 생성된 모델을 이용해 미세먼지 농도를 추정하는 과정을 설명한다. 5장에서는 제안한 방법을 사용해 미세먼지 측정소가 있는 지역의 PM<sub>10</sub> 농도에 따른 label추정된 결과를 제시하고, CRFs를 이용한 결과와 비교, 분석한다. 6장에서 결론을 내리고 향후 과제를 제시한다.

## 2. 시스템 개요

부천시의 각 구역이 갖는 데이터를 획득하기 위해 구글 지도 api를 이용해 부천시를 0.5km × 0.5km의 크기를 갖는 360개 구역으로 나눈다. 각 구역 중 미세먼지 측정소가 위치한 곳의 대기 품질 지수(AQI, Air Quality Index)에 따른 label은 에어코리아에서 제공하는 Table 1를 따르고, 대기 오염물질 중 PM<sub>10</sub> (Particulate Matter Less than 10µm)을 기준으로 한다.

본 논문에서는 [2]의 구조에서 학습 알고리즘으로 오차역전파법을 사용하는 인공신경망(ANN, Artifi-

Table 1. PM10 concentration and corresponding label

PM10 Concentration (µg/m <sup>3</sup> )	Label	Color
0~30	Good	Blue
31~80	Normal	Green
81~150	Bad	Yellow
More than 150	Very Bad	Red

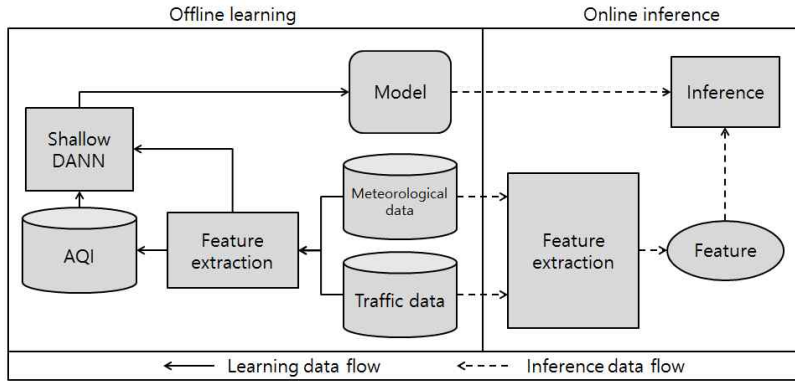


Fig. 1. system overview.

cial Neural Network)[12]과 조건부 무작위장(Conditional Random Field)[13] 대신 DANN (Domain Adversarial Neural Network)[11]을 사용함으로써 오류의 누적을 막고, 다른 여러 오염물질에 대한 데이터를 효과적으로 사용하여 학습에 도움이 되도록 한다.

전체 프레임워크는 데이터를 학습하는 부분과 추론하는 부분으로 이루어져 있다. 학습 부분에서 각 구역에서 추출한 특징과 레이블이 있는 데이터, 그리고 레이블이 없는 데이터를 DANN 에 넣어 학습시킨 모델을 생성하고, 추론 부분에서는 생성된 모델과 입력 데이터로부터 추출한 특징을 이용해 원하는 구역의 미세먼지 농도를 추정한다.

### 3. 특징 추출

#### 3.1 기상 데이터(Meteorological data)

시간적 데이터는 도시 데이터 중 시간의 흐름에 따라 변화하는 데이터를 말하며, 본 논문에서는 기상 데이터와 교통 데이터를 사용한다. 기상 데이터의 경우 부천시의 각 동의 기온, 습도, 강수량, 풍속 데이터를 사용한다. 기상청에서 부천시의 각 동별 데이터를 제공하는데 반해, 미세먼지 농도를 추정할 단위는 부천시의 각 구역이기 때문에 학습에 사용하기 전에 데이터를 가공할 필요가 있다. 이를 위해 획득한 부천시의 각 구역이 속한 동별 데이터  $D_N = \{d_1, d_2, \dots, d_n\}$  에 다음 식과 같이 연속균분포함수를 갖는 random noise를 생성해 데이터에 더한 값을 해당 동에 속한 구역에 할당한다.

$$f(x) = \begin{cases} \frac{5}{|\mu|} & \text{for } -\frac{|\mu|}{10} \leq x \leq \frac{|\mu|}{10} \\ 0 & \text{for } x < -\frac{|\mu|}{10} \text{ or } x > \frac{|\mu|}{10} \end{cases} \quad (1)$$

$$g_t^{ij} = d_t^i + x \quad (2)$$

여기서  $\mu$ 는 각 동의 기상 데이터의  $h$ 시간 단위 평균들의 집합  $M = \left\{ \frac{1}{h} \sum_{i=1}^h d_i, \frac{1}{h} \sum_{i=h+1}^{2h} d_i, \dots \right\}$ 의 원소이다.  $d_t^i$ 는  $i$ 번째 동의 시간  $t$ 에서의 기상 데이터를 나타내고,  $g_t^{ij}$ 는  $i$ 번째 동에 속한  $j$ 번째 구역의 시간  $t$ 에서의 요소별 기상 데이터를 나타낸다. 본 논문에서  $h$ 값은 24로 설정한다. noise를 생성해 각 구역에 적용한 결과를 그림으로 나타내면 Fig. 2와 같다.

#### 3.2 교통 데이터(Traffic data)

구역을 통과하는 도로의 평균통행속도를 해당 구역의 값으로 설정하고, 통과하는 도로가 없을 경우에는 그 구역과 인접한 구역에 있는 도로의 평균통행속도로 설정한다. 일반적으로 도로가 정체되었을 때 해당 구역의 자동차의 오염물질 배출량이 증가하므로, 두 개 이상의 도로가 구역을 통과할 경우 구역을 통과하는 모든 도로 구간의 평균통행속도 중 가장 작은 값으로 설정한다(Fig. 3. 참조).

부천시 내의 모든 도로의 데이터가 제공되지 않고, 포함하거나 가까운 도로가 없는 구역도 있기 때문에 모든 구역이 교통 데이터를 갖지는 않는다. 그러한 구역은 0을 할당해 학습을 할 때 통행속도를 제외한 다른 특징(기상 데이터)을 이용해 학습을 진행하도록 한다.

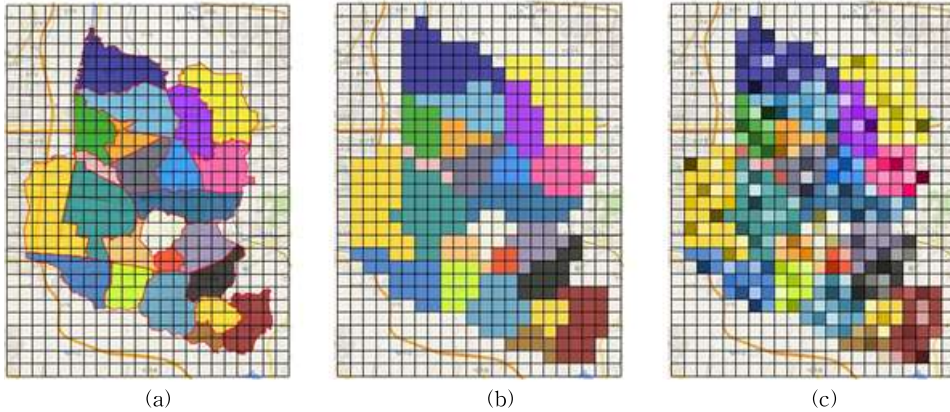


Fig. 2. The process of acquiring meteorological data for each grid. (a) meteorological data of each *dong* in Bucheon. (b) Assignment the data of *dong* to which each grid belongs. (c) the result of applying random noise to the (b).



Fig. 3. Traffic data for each grid. In case of B, there is no road section included, so it is set to the same value as the road (road of A) at a close distance.

$$g_{traffic} = \begin{cases} \min(V^R), & R \neq \emptyset \\ 0, & R = \emptyset \end{cases} \quad (3)$$

$$R = \begin{cases} R_g, & R_g \neq \emptyset \\ R_{g_c}, & R_g = \emptyset \text{ and } R_{g_c} \neq \emptyset \\ \emptyset, & R_g = \emptyset \text{ and } R_{g_c} = \emptyset \end{cases}$$

식 (3)에서,  $V^R$ 은 도로 구간 집합  $R$ 에 속한 구간들의 평균통행속도를 원소로 갖는 집합을 나타내고,  $R_g$ 와  $R_{g_c}$ 는 각각 구역  $g$ 와  $g$ 에 인접한 구역에 속해 있는 도로 구간 집합을 나타낸다.

교통 데이터는 2개 이상의 구역이 같은 도로를 공유할 경우 같은 값을 갖기 때문에 기상 데이터와 마찬가지로 식 (1)을 따르는 random noise를 생성시켜 적용한 값을 정규화하여 특징으로 사용한다.

#### 4. 학습 및 추론

기계학습 알고리즘을 이용해 문제를 해결하고자 할 때 직면하는 대부분의 문제는 학습 데이터 중 label이 있는 데이터에 비해 label이 없는 데이터가 훨씬 많다는 것이다. 이 문제를 해결하기 위한 많은 연구가 진행되어 왔다[14,15]. 본 논문에서는 Ganin 등이 제안한 Domain Adversarial Neural Network (DANN)[11]를 이용해 학습을 진행하고 학습한 모델과 추출한 특징을 이용해 측정소가 없는 구역의 미세먼지 농도에 따른 label을 추정한다.

##### 4.1 Domain Adversarial Neural Network의 학습

DANN의 주요 아이디어는 학습을 진행하며 특징을 추출할 때, label이 있는 데이터의 클래스는 잘 분류하게 하고 동시에 label이 있는 도메인과 label이 없는 도메인을 잘 구분하지 못하게 하는(도메인에 강건한) 특징을 추출하려는 것이다. DANN에서 사용하는 목적함수는 다음과 같이 정의된다.

$$E(\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}, \mathbf{u}, z) = \frac{1}{n} \sum_{i=1}^n L_y^i(\mathbf{W}, \mathbf{b}, \mathbf{V}, \mathbf{c}) \quad (4)$$

$$- \lambda \left( \frac{1}{n} \sum_{i=1}^n L_d^i(\mathbf{W}, \mathbf{b}, \mathbf{u}, z) + \frac{1}{n'} \sum_{i=n+1}^{N'} L_d^i(\mathbf{W}, \mathbf{b}, \mathbf{u}, z) \right)$$

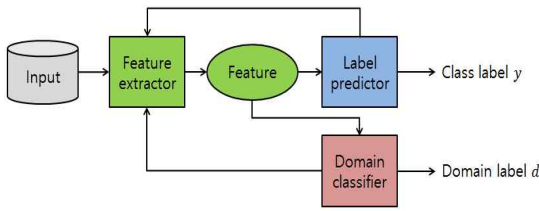


Fig. 4. DANN architecture.

위 식에서  $L_y$ 와  $L_d$ 는 각각 클래스 분류기와 도메인 분류기의 손실(Loss)값을 나타낸다.  $\mathbf{W}, \mathbf{b}$ 는 특징을 추출하는 네트워크의 매개변수이고  $\mathbf{V}$ 와  $\mathbf{c}$ , 그리고  $\mathbf{u}$ 와  $\mathbf{z}$ 는 각각 클래스 분류기, 도메인 분류기의 매개변수를 나타낸다. 학습 과정에서 획득해야 할 각각의 파라미터  $\hat{\mathbf{W}}, \hat{\mathbf{V}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\mathbf{u}}, \hat{\mathbf{z}}$ 는 다음 식을 따른다.

$$(\hat{\mathbf{W}}, \hat{\mathbf{V}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}) = \underset{\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}}{\operatorname{argmin}} E(\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}, \hat{\mathbf{u}}, \hat{\mathbf{z}}) \quad (5)$$

$$(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = \underset{\mathbf{u}, \mathbf{z}}{\operatorname{argmin}} E(\hat{\mathbf{W}}, \hat{\mathbf{V}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}, \mathbf{u}, \mathbf{z}) \quad (6)$$

식 (5)와 식(6)에 따르면 DANN의 최적화 문제는 하나의 목적함수에 대하여 일부 파라미터는 목적함수를 최소화하고, 일부 파라미터는 목적함수를 최대화해야 하는 문제가 있다. 이 문제를 stochastic gradient procedure 방법을 이용해 해결하기 위해 최소화 매개변수에 대해서는 식 (4)의 기울기와 반대 방향으로 갱신을 진행하고, 최대화 매개변수에 대해서는 기울기 방향으로 갱신이 이루어지게 한다. 즉, 학습은 식 (4)에 대해  $\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}$ 를 매개변수로 갖는 클래스 분류기와 매개변수  $\mathbf{u}, \mathbf{z}$ 를 갖는 도메인 분류기로서 적대적인 방향으로 이루어진다.

4.2 추론 과정

먼저 3장에서 설명한 방법을 이용하여 기상 데이

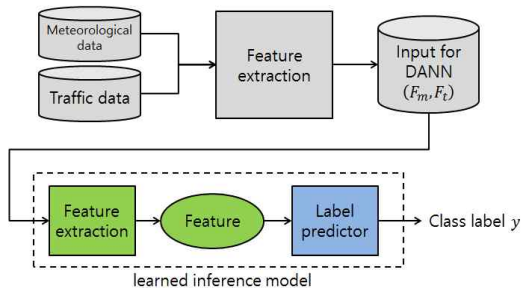


Fig. 5. Inference process.

터와 교통 데이터로부터 각각 기상 데이터 특징( $F_m$ )과 교통 데이터 특징( $F_t$ )을 추출한다. 추출한 특징은 DANN의 학습 결과로 생성된 추론 모델의 입력으로 사용된다. 추론 모델은  $\mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}$ 를 매개변수로 갖고, 입력 데이터에서 특징을 추출해 데이터에 대응하는 클래스를 추정한다.

5. 실험 결과 및 고찰

5.1 데이터 구성

기상 데이터는 기상청 홈페이지(<https://data.kma.go.kr>)에서 제공하는 데이터를 사용했다. 기상청에서는 각 지역의 기상요소별 측정치를 제공하는데, 본 논문에서는 강수량, 기온, 습도, 풍속 데이터를 사용하였다.

교통 데이터는 부천시교통정보센터(<http://www.bcits.go.kr/>)에서 얻을 수 있지만 해당 홈페이지에서는 도로별 각 구간의 하루 동안의 평균값을 제공한다. 본 연구에 필요한 데이터는 한 시간 동안의 평균 통행속도이므로 부천시시설관리공단(<http://www.bcs.or.kr/>)에 2016년 3월, 6월, 9월, 12월 데이터를 요청해 사용하였다. 두 기관에서 제공하는 데이터의 도로 구간은 같으나 시설관리공단에서 제공하는 데이터의 도로 구간이 좀 더 세분화되어 있다.

미세먼지 측정치는 에어코리아(<https://www.airkorea.or.kr/index>)에서 제공하는 데이터를 사용하였다. 부천시의 경우 오정동, 내동, 계남공원, 중2동, 소사본동의 5군데에서 정각을 기준으로 한 시간 단위로 대기 질을 측정한다.

5.2 실험 결과

실험은 측정소가 있는 구역 중 한 군데의 label을 제거하고, 생성된 모델을 이용해 label을 제거한 구역의 label을 추정하는 방식으로 진행하였다. 사용한 DANN의 은닉층의 크기는 25, 학습률은 0.05,  $\lambda$ 값은 6으로 설정하였고, 각 네트워크가 하나의 은닉층을 갖는다. source domain은  $PM_{2.5}$ 와 오존 데이터를 사용했고 target domain은  $PM_{10}$  데이터를 사용하였다.

source domain의 데이터의 수가 성능에 영향을 끼치는 정도를 확인하기 위해 source domain의 데이터의 수를 증가시키며 학습을 진행하고, 2016년 3월1일부터 3월31일까지 미세먼지 측정소가 있는 구역 293

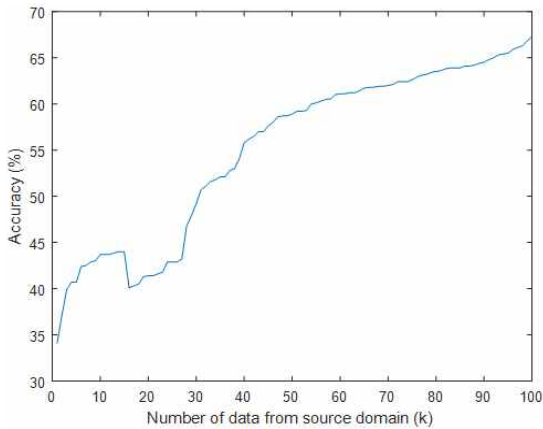


Fig. 6. Accuracy changes with number of data from source domain.

의  $PM_{10}$  농도에 따른 label (Table 1 참조)을 추정된 결과를 비교하였다. Fig. 6의 그래프는 학습에 사용한 source domain의 데이터 수의 증가에 따른 정확도의 변화를 나타내는데, source domain의 데이터의 수가 증가함에 따라 정확도 역시 증가하고 있음을 확인할 수 있다.

또한 미세먼지 측정소가 있는 구역 71, 97, 166, 221과 293의 2016년 3월 1일과 2016년 3월 31일 사이에 측정된 미세먼지 농도에 따른 label과 각 추론 모델(CRFs, DANN)의 추정 결과를 비교해 정확도를 계산하였다. 전체 시간 흐름  $N(H)$ , 구역  $g$ 에 대한 모

델  $f$ 의 정확도는 다음 식을 이용해 계산한다.

$$Accuracy(f) = \frac{N(f_g^H(F_m, F_t) = I_g^H)}{N(H)} \quad (7)$$

여기서  $f_g^H$ 와  $I_g^H$ 는 각각 구역  $g$ , 시간 집합  $H$ 에서의 모델  $f$ 의 추론 결과와 실제 label을 나타낸다.

본 논문에서 제안한 방법과 CRFs를 이용한 방법의 성능을 비교했을 때 (Table 2) 제안한 방법을 사용했을 때 정확도가 다소 더 높은 것을 확인할 수 있다. 제시한 추론 모델(DANN)을 학습하기 위해서는 추가적인 domain을 필요로 함에도 불구하고 월등한 성능 차이를 보이지는 못했는데, DANN은 CRFs와 달리 연속적인 데이터가 갖는 특징을 충분히 활용하지 못하기 때문인 것으로 보인다.

Fig. 7은 각 추론 모델을 이용해 2016 3월 4일 20시 부천시 일부 지역의 미세먼지 농도에 따른 label을 추정된 결과를 나타낸 것이다. 색깔은 미세먼지 농도에 따른 label을 나타내며, 측정소가 있는 구역은 스타(별)로 표시하였다.

선형 보간법을 이용해 미세먼지 농도를 추정했을 때에는 다른 요소를 고려하지 않고 측정소의 미세먼지 수치와 측정소 사이의 공간적 관계만 고려하기 때문에 공간적으로 두 측정소 사이에 있는 구역은 측정소의 값에 따라 선형적으로 변할 수밖에 없고, 좋지 않은 추론 결과를 보인다. Fig. 7(a)에서 측정소의 label 값이 주변 영역에 영향을 미치는 것을 확인

Table 2. Performance comparison between methods

	Method	Tested grid				
		grid 71	grid 97	grid 166	grid 221	grid 293
Accuracy	CRFs	59.72%	57.56%	58.46%	59.28%	61.63%
	DANN	61.35%	59.12%	62.19%	60.84%	67.3%

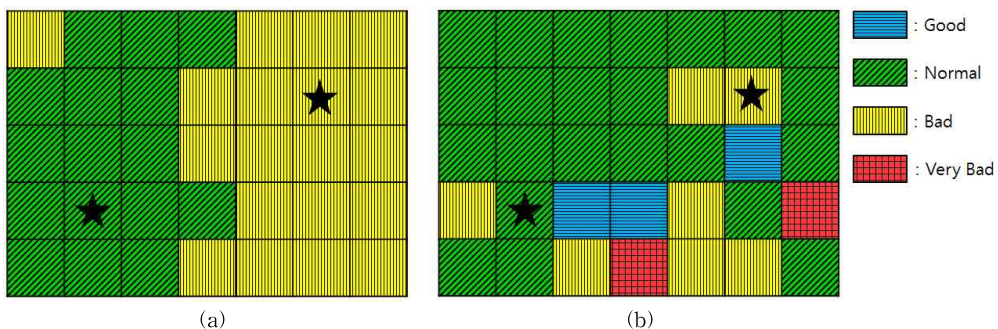


Fig. 7. Classification result of inference model, (a) Linear interpolation, (b) Our inference model(DANN).

할 수 있다. 이에 반해 DANN을 이용한 추론 모델은 측정소의 label 값과 측정소가 있는 구역과 없는 구역 모두의 기상데이터, 교통데이터를 고려하기 때문에 보다 더 정밀한 추론이 가능하다(Fig. 7(b))

## 6. 결 론

본 논문에서는 측정소가 없는 지역의 미세먼지 농도를 기계학습 알고리즘을 이용해 추정하고자 하였다. 이를 위해, 일부 측정소가 있는 구역의 미세먼지 농도와 두 개의 도시 데이터(기상 데이터, 교통 데이터)로부터 기계학습 알고리즘을 적용하기 위한 특징을 추출하는 방법을 제안하였다. 추출한 특징을 DANN의 입력으로 넣어 네트워크를 학습하고, 학습된 모델을 이용하여 해당 구역의 미세먼지 농도에 따른 label을 추정하였다. 실험 결과에서 source domain의 데이터 수가 많아짐에 따라 CRFs를 사용한 방법보다 성능이 좋아짐을 확인할 수 있었고, 네트워크의 구조를 변경하거나 유동인구 등의 추가적인 데이터를 활용함으로써 성능을 개선할 수 있을 것으로 보인다. 본 연구의 한계로는 미세먼지가 농도가 갖는 시간, 위치에 따른 변화를 반영하지 못한다는 것이다. 추후 연구에서 연속적인 데이터를 처리하는데 적합한 RNN(Recurrent Neural Network)등의 네트워크를 추가적으로 활용하여 성능을 개선하고자 한다.

## REFERENCE

- [ 1 ] M.K. Shin, C.D. Lee, H.S. Ha, C.S. Choe, and Y.H. Kim, "The Influence of Meteorological Factors on PM 10 Concentration in Incheon," *Journal of Korean Society for Atmospheric Environment*, Vol. 23, pp. 322-331, 2007.
- [ 2 ] Y. Zheng, F. Liu, and H. Hsieh, "U-Air: When Urban Air Quality Inference Meets Big Data," *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1436-1444, 2013.
- [ 3 ] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, et al. "Forecasting Fine-grained Air Quality Based on Big Data," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2267-2276, 2015.
- [ 4 ] J.Y. Zhu, C. Zhang, S. Zhi, V.O.K. Li, J. Han, and Y. Zheng, "P-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data," *arXiv Preprint arXiv:1610.07045*, 2016.
- [ 5 ] J.H. Leem, S.T. Kim, and H.C. Kim, "Public-health Impact of Outdoor Air Pollution for 2nd Air Pollution Management Policy in Seoul Metropolitan Area, Korea," *Annals of Occupational and Environmental Medicine*, Vol. 27, pp. 1-9, 2015.
- [ 6 ] H. Wang, X. Yuan, X. Wang, Y. Zhang, and Q. Dai, "Real-time Air Quality Estimation Based on Color Image Processing," *Proceeding of Visual Communications and Image Processing Conference*, pp. 326-329, 2014.
- [ 7 ] S. Vardoulakis, B.E.A. Fisher, K. Pericleous, and N.G. Flesca, "Modelling Air Quality in Street Canyons: A Review," *Atmospheric Environment*, Vol. 37, Issue 2, pp. 155-182, 2003.
- [ 8 ] J. Blitzer, R. McDonald, and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 120-128, 2006.
- [ 9 ] P. Germain, A. Habrard, F. Laviolette, and E. Morvant, "A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers," *Proceeding of International Conference on Machine Learning*, pp. 738-746, 2013.
- [ 10 ] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized Denoising Autoencoders for Domain Adaptation," *arXiv Preprint arXiv:1206.4683*, 2012.
- [ 11 ] Y. Ganin, H. Ajakan, H. Larochelle, F. Laviolette, and V. Lempitsky, "Domain-adversarial Training of Neural Networks," *Journal of Machine Learning Research*, Vol. 17, pp. 1-

- 35, 2016.
- [12] D.M.S. Arsa, G. Jati, A.J. Mantau, and I. Wasito, "Dimensionality reduction using deep belief network in big data case study: Hyperspectral image classification," *Big Data and Information Security International Workshop*, pp. 71-76, 2016
- [13] J. Lafferty, A. McCallu, and F.C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proceedings of the 18th International Conference on Machine Learning*, pp. 282-289, 2001.
- [14] Semi-supervised Learning Literature Survey (2008), [http://pages.cs.wisc.edu/~jerryzhu/pub/ssl\\_survey.pdf](http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf) (accessed Jul., 15, 2017)
- [15] M. Baktashmotlagh, M.T. Harandi, B.C. Lovell, and M. Salzmann, "Unsupervised Domain Adaptation by Domain Invariant Projection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 769-776, 2013.
- [16] H.S. Bae, S.H. Yu, and H.Y. Kwon, "Fast Data Assimilation using Kernel Tridiagonal Sparse Matrix for Performance Improvement of Air Quality Forecasting," *Journal of Korea Multimedia Society*, Vol.20, No. 2, pp. 363-370, 2017



강 태 천

2017년 가톨릭대학교 컴퓨터정보공학부 컴퓨터공학전공 (학사)  
 2017년~현재 가톨릭대학교 디지털미디어학과 석사과정  
 관심분야: 컴퓨터비전, 기계학습, 인공지능



강 행 봉

1980년 한양대학교 전자공학과 (학사)  
 1986년 한양대학교 전자공학과 (석사)  
 1989년 Ohio State Univ. 컴퓨터공학(석사)

1993년 Rensselaer Polytechnic Institute 컴퓨터공학(박사)  
 1993년~1997년 삼성종합기술원 수석연구원  
 1997년~현재 가톨릭대학교 디지털미디어학과 교수  
 2005년 UC Santa Barbara, Visiting Professor  
 관심분야: 컴퓨터비전, 기계학습, HCI, 컴퓨터그래픽스, 인공지능, 빅데이터