KIPS Tr. Software and Data Eng.
Vol.6, No.9 pp.457~464 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.9.457

Indoor Air Condition Measurement and Regression Analysis System Through Sensor Measurement Device and Gated Recurrent Unit

Jaehyun Ahn† · Dongil Shin†† · Kyuho Kim††† · Jihoon Yang††††

ABSTRACT

Indoor air quality analysis is conducted to understand abnormal atmospheric phenomena and the external factor affecting indoor air quality. By recording indoor air quality measurements periodically, we are able to observe patterns in air quality. However, it difficult to predict the number of potential parameters, set parameters for a given observation and find the coefficients. Moreover, the results are time-dependent. Thus to address these issues, we introduce a microchip capable of periodically recording indoor air quality and a model that estimates atmospheric changes based on time series data.

Keywords : Atmospheric Observation System, Time Series Prediction, Long-Short Term Memory (LSTM), Circuit Type Circulation Unit (GRU)

센서 측정기와 회로형 순환 유닛(GRU)을 이용한 실내 공기 품질 측정 및 추세 예측 시스템

안 재 현† · 신 동 일†† · 김 규 호††† · 양 지 훈††††

요 약

실내 공기 품질 측정은 측정 대상 공간의 대기 상태 유지, 외부 변인으로 인한 대기 이상 현상을 검출하려는 방법이다. 실내 공기 품질을 주기적으로 기록하면서 변인에 따른 공기 변화에 특정 패턴이 발생함을 관측할 수 있었으나, 파라미터를 설정하고 계수를 찾아 나가기엔 파라미터의 개수나 그 영향력을 추산하기 어렵고 결과가 시간에 의존적이라는 문제가 있다. 따라서 본 실험은 이것을 공식화하는 대신, 측정 주기마다 추이를 예측하는 관측치 중심의 기계 학습 모델을 개발하는 것을 목표로 한다. 본 논문은 실내 대기 품질을 주기적으로 전송 및 저장하는 측정기의 기록 데이터로 공기 품질 변화를 예측하는 모델을 설명하고 시계열 분석 모델을 구축한다.

키워드 : 대기 관측 시스템, 시계열 예측, 장기-단기 주기 메모리 네트워크(LSTM), 회로형 순환 유닛(GRU)

1. 서 론

실내 공기 품질 측정은 측정 대상 공간의 대기 상태 변화를 검출하려는 방법이다. 이러한 공간의 공기 품질은 미시적으로는 사람의 출입, 냉·난방기의 사용과 같은 변인에 의해 급격히 변할 뿐 아니라, 변인 제거 시 변인 통제 이전의 상태로 복원되는 속도 역시 빠르다. 그러므로 실내 공기 품질의 변화를 예측하는 공식을 설정하는 것은 많은 변수의 영향력을 광범위하게 파악하고 있어야 하며, 그것이 어떤 부피의 공간에 설치되어 있는지, 공간 안에 어떤 열전도율을 가진 물체들이 배치되어있는가 등을 계산할 수 있어야 한다는 것을 의미한다. 그러나 이것은 불특정 다수의 물체에 대하여 모두 수행되어야 하는 작업이며, 실내 공기 품질 예측의 보편적인 공식 유도를 불가능에 가깝게 만든다.

위와 같은 어려움 때문에 최근까지도 많은 실내 공기 품질 제어 시스템은 임계점 초과 방식으로 변인을 통제해 왔

※ 이 논문은 2016년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구(NO: R7117-16-0098) 및 2017년도 산업통상자원부 및 산업기술평가관리원의 지원을 받아 수행된 연구임(NO: 10076752)
† 비 회 원 : 버즈니(주) 연구원
†† 비 회 원 : 서강대학교 컴퓨터공학과 석사과정
††† 비 회 원 : 서강대학교 산학협력중점 교수
†††† 종신회원 : 서강대학교 컴퓨터공학과 교수
Manuscript Received : May 30, 2017
First Revision : July 5, 2017
Accepted : July 25, 2017
* Corresponding Author : Jihoon Yang(yangjh@sogang.ac.kr)

다. 임계점 초과 방식은 공간 안에 얼마나 많은 변인이나 방해 요소가 있든 간에 상관없이, 미리 설정해둔 수치를 넘어서면 강력한 자체 변인을 작동시키는 방식이다. 임계점 초과 방식의 실내 공기 품질 제어의 경우, 변인의 특징, 영향력과 같은 변수나 실험 공간이 위치한 환경 상황을 고려하지 않아도 된다는 장점이 있다. 그러나 위 방식은 미세한 공기 품질 변화에도 큰 영향을 받는 정밀 실험 기구에 적합하지 않다. 또한 측정 대상의 주변 환경이 공간과 어떻게 상호작용하는가를 고려하지 않는다. 임계점 측정 방식의 경우, 특정 공기 품질 하나에 대하여서만 작동하는 방식이기 때문에 변인의 원인을 복합적으로 파악한다거나 분별하는데 어려움이 있다.

본 연구는 위에서 언급한 기존의 센서 측정 방식의 한계를 인식하고, 기계 학습으로 시계열 데이터를 예측하는 모델을 구상한다. 기계 학습 모델의 경우 기존의 실험이 가지고 있는 한계를 극복할 수 있는 근거가 있다.

첫 번째로, 기계 학습 모델은 데이터 중심적이다. 기계 학습 모델은 대기 품질에 영향을 미칠 수 있는 변인의 영향력을 고려하지 않는다. 대신에 모델은 결과로 계측되는 내용으로부터 패턴을 찾아내고, 특정 패턴이 어떠한 추세 선을 그리는가를 파악한다. 결과적으로 데이터의 크기가 충분히 크고, 패턴의 조합이 복합적이라면, 기계 학습 모델은 복잡하고 가변적인 공식의 유도가 없어도 대기 품질을 예측할 수 있다.

또한, 기계 학습은 모델에 따라 복합적인 변인 요소 고려가 가능하다. 기계 학습으로 단순하게 한 변인의 추세를 예측하는 모델을 구축하는 것도 가능하지만, 조금 더 복잡한 시계열 분석 모델을 사용할 경우 앞서 설명한 태양과 같은 복합적 대기 품질 변인 패턴 검출이 가능하다. 이것은 기존에 사용되던 단순한 공기 품질 계측기 유닛들을 조합하는 것으로 기존의 센서들이 검출하지 못했던 복합적인 변인의 검출이 가능하다는 것을 의미한다.

본 논문은 실내의 대기 품질을 주기적으로 기록하는 대기 품질 측정 장치로 실내 공기 품질 추세를 예측하는 모델을 소개한다. 위 과정으로 구성된 기계 학습 모델은 과거 폐쇄 환경 공기 품질을 예측하기 위해 공식을 적용하는 방식과 유니 모달 기계학습 모델(uni-modal machine learning model)과 그 성능을 비교했을 때, 더욱 높은 정확도의 분류 성능을 보임을 확인할 수 있다.

2. 관련 연구

2.1 기계 학습을 사용한 공기 품질 예측

기계 학습을 사용하여 공기 품질을 예측하려는 시도는 지속해서 이루어졌다. 실내 미립자 측정 모델[1]은 기계 학습 기법인 선형 회귀(Linear Regression) 모델을 실험 검증 비교군으로 사용한다.

이외에도 유럽의 여섯 국가를 바탕으로 시행한 실내 이산화질소, 배기가스 밀집도 공기 품질 연구는 해당 결과를 시각화하는 과정에서 추세선을 보여주는 방식으로 시계열 초미세먼지(PM2.5) 지수를 예측한다[2].

이뿐 아니라 회기 분석은 각종 실내 공기 품질 요소 추세를 예측하는 방법으로 사용된다. 온실 내부 온도는 식물 재배 환경 모니터링에 주요한 변수다. 이에 대한 수학적 모델이 이미 존재하지만, 극지방의 건조하고 추운 환경 변수는 위의 공식을 적용하기에 문제가 있다. 따라서 위 공식에 회귀 모델을 결합하고, 외부 계절 상황을 고려하여 새로운 관계 모델을 구성하는 연구가 진행된 바 있다[5].

2.2 시계열 데이터 기반 기계 학습 연구

시계열 데이터를 예측하기 위한 기계 학습 연구는 음향 신호를 인지하고 환경을 인식하기 위한 시도로 처음 등장하였다. 초기 음향 신호 구분 연구는 조건부 확률을 사용한 은닉 마르코프 모델(Hidden Markov Model: HMM)을 사용하는 방식과 지지 벡터 기계(Support Vector Machine: SVM)을 사용하여 음성 인식 모델을 구성하였다[4]. 이후 딥 러닝이 등장하면서 신경망 구조에 기반을 둔 시계열 분석 기법이 등장한다. 시계열 분석 기법은 깊은 신경망에서 시계열 정보를 사용하기 위한 순환 신경망 모델이 있고, 순환 신경망의 장기 기억 의존성 문제를 해결하기 위해 만들어진 장기-단기 주기 메모리 네트워크(Long-Short Term Memory: LSTM) 모델이 있다[3]. 기본적으로 순환 신경망 구조의 기계 학습 모델은 시간적인 정보를 다시 한 번 은닉 신경망에 활용하기 위한 구조로 되어 있다. 그래서 순환 신경망 구조는 과거 정보를 현재 정보와 취합하여 분류를 진행할 수 있다는 장점을 가진다[6]. 그리고 LSTM 네트워크 모델과 함께 다양한 응용 모델이 등장하고 있다. 응용 모델 중 성능이 가장 좋다고 알려진 모델은 LSTM 신경망과 회로형 순환 유닛(Gated Recurrent Units: GRU)이다[7].

GRU는 LSTM과 동일한 필터 방식을 사용하지만 게이트의 수를 줄이고 파라미터 숫자를 효과적으로 감소시킨 네트워크 모델이다. 리셋 게이트와 업데이트 게이트 두 개로 구성되어 있으며, 두 게이트의 시차에 따른 학습 비중 조정 방식(Backpropagation Through Time: BPTT)으로 모델을 학습한다. GRU는 모체 신경망 구조인 LSTM보다 적은 파라미터를 사용하기 때문에 기존 신경망에 비해 학습 속도가 빠르고, LSTM 신경망보다 사용되는 데이터의 수가 적다는 장점이 있다. 그렇지만 GRU가 LSTM 신경망보다 분류 성능이 더 좋은가에 대해서는 논의의 소지가 있다[8]. 이 소지는 두 모델 중 어떤 모델이 특정 데이터를 처리하는 것에 적합한지 함께 비교하는 것이 좋다.

3. 데이터 구성

3.1 데이터 수집

본 실험에서 측정하는 대기 측정 요소는 이산화탄소, 미세먼지, 온도, 습도, 광량, 휘발성 유기 화합물(Volatile Organic Compounds: VOC)의 6종류이다. 센서 측정 모듈은 HTTP 리퀘스트 방식으로 수집한 데이터를 서버에 주기적으로 전송한다. 서버 전송 주기는 1분이며, 한 주기 동안 모듈은 6종의 공기 품질 측정 기준에 맞추어 공기 함량을 측정하고 리눅스 서버에 그 기록을 전송한다. Fig. 1은 데이터를 주기적으로 측정 및 전송하는 모듈을 도식한 것이다.

Fig. 1의 센서 노드 체계로 수집된 데이터는 MySQL의 형태로 주기적으로 수집했으며, 실험에 사용한 데이터는 서버에 기록되어 있는 약 2천 1백만 레코드를 CSV(comma-separated values) 파일로 저장한 뒤 학습 및 테스트 모델을 구성하였다. Table 1은 수집한 데이터의 요약된 정보를 나타낸다.

Table 1. Summary of Metadata for Building Learning Models

Data	Detail
Measurement site	SK Corporation Jongro Building (Euljiro 65, Jung-gu, Seoul)
Number of all records	21,781,467
Data size	1.36 GB (1,426,063 Bytes)
Data collection period	60,504 hours (16/2/22 - 16/9/20)
Data type	MySQL
Data components	6 air quality variables (CO2, Dust, Temperature, Humidity, Light, VoC) etc.

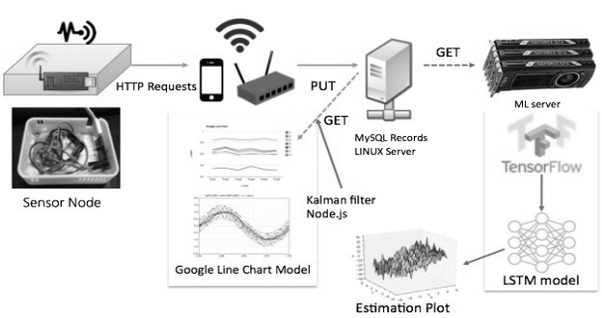

Fig. 1. Module Diagram for Periodic Measurement and Transfer of Air Quality Data

3.2 데이터 전처리

수집한 데이터는 공기 성분의 구분 없이 매 1분당 여섯 개의 센서 노드에서 레코드가 기록되어 있다. 이렇게 직렬로 쌓이는 데이터를 시계열 분석 기계 학습 모델에 적용하기 위해서는 근접 시간 내에 수집된 여섯 개의 서로 다른 센서 노드끼리 한 특징 벡터(feature vector)로 묶고, 병렬화하는 데이터 전처리(preprocessing) 과정이 필요하다.

데이터 전처리 순서는 다음과 같다. Table 1에서 수집된 약 2천 1백 7십만 개의 레코드를 시간 순으로 정렬한 후, 각 센서 노드를 종류별로 리스트에 구분 지어 넣는다. 이후 수집된 센서 노드들을 시간 중심으로 병렬로 엮은 후, 여섯 개의 특징 벡터가 모두 수집되면, 그 데이터를 하나의 데이터 프레임으로 간주한다. 이렇게 모인 데이터 세트 집합은 단위 시간당 센서 노드 특징들이 기록되어 있다. 이렇게 누적된 시계열 데이터 세트는 데이터 종류, 데이터 크기, 시간 값을 가지는 삼차원 텐서(3D Tensor)가 된다.

3.3 시계열 학습 데이터의 구성

3.2에서 여섯 종류의 센서 노드를 하나의 벡터로 산정한 뒤 시간 순서대로 정렬한 리스트는 2차원 텐서의 모습을 가진다. Fig. 2은 3.2에서 형성한 2차원 텐서에 시계열 데이터를 축적하여 3차원 텐서로 데이터를 축적하는 것을 시각화한 그림이다. 2차원 텐서는 여섯 센서 노드가 단위 시간 하나의 간격으로 일렬로 늘어서 있다. 본 실험의 3차원 텐서는 이러한 2차원 텐서를 경과 시간(time step) 단위로 묶은 것으로, 실험의 측정 주기가 1분이라는 것을 고려했을 때, $1\times$timestep 크기의 부피를 갖는다. 이러한 3차원 텐서 하나의 경과 시간(time step)을 t라고 할 때, 시계열 예측 모델은 $t+1$ 시간이 지났을 때, 벡터들의 모습을 예측한다.

3차원 텐서의 학습/실험 데이터 세트는 10-fold cross validation을 사용하여 분리 마련하였다. 총 21,781,467개 상당의 10종 공기 품질이 계측된 CSV 레코드는 그중 큰 의미가 있는 6종의 공기 품질 계측 데이터(미세먼지, 광량, 휘발성 유기화합물, 이산화탄소, 온도, 습도: 6종)를 선별하여 2,173,790(전체 레코드의 약 9.98%)개의 직렬 레코드를 남기고, 다시 3차원 텐서로 재가공하는 과정을 거친다. 본 과정이 끝났을 때, 예측 모델을 구성하기 위한 학습 데이터는 $[t\times 6\times 1]\times 299596$의 3차원 텐서가 되고, 테스트 데이터는 $[t\times 6\times 1]\times 33289$ 모양의 3차원 텐서를 구성할 수 있다. Table 2는 3차원 텐서를 구현하기 위한 알고리즘이다.

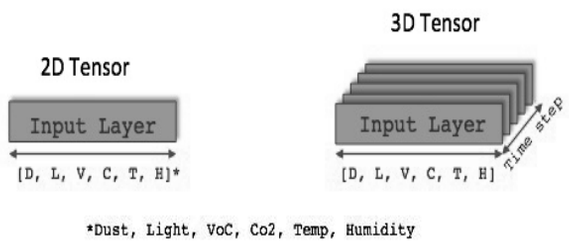

Fig. 2. 2D Tensor(left) and 3D Tensor(right)

Table 2. 3D Tensor Implementation Algorithm

Algorithm 1. 3D tensor implementation

Input : CSV record $\{v_{1,2,3,\cdots,9,10}\}_{record}$,
Fold rate r
Time step t

Initialize Model parameter fold rate is used for K-fold cross validation

1. Select_fecature() ← Select $\{v_{1,2,3,\cdots,5,6}\}_{record}$
2. Set_feacutre_vector() ← Sort $\{v_{1,2,3,\cdots,5,6}\}_{record}$through time
3. $\{x_{v\times t}\}_{record}$← Pile $\{x_{v\times 1}\}_{record}$with t times in order
4. $\{x_{v\times t}, y_{v\times 1}\}_{test}/\{x_{v\times t}, y_{v\times 1}\}_{train}$
← Separate $\{x_{v\times t}\}_{record}$by fold rate r

Output : Training Dataset/Test Dataset ,
$\{x_{v\times t}, y_{v\times 1}\}_{train}/\{x_{v\times t}, y_{v\times 1}\}_{test}$

4. 기계 학습 모델에 기반 한 시계열 예측 시스템

시계열 데이터를 예측하기 위한 모델로 현재까지 가장 널리 사용되는 세 가지 모델을 살펴본다. 세부적으로는 하나의 선형 모델과 두 가지 딥 러닝 모델 구조에 대하여 살펴볼 것이며, 각각 이름은 선형 회귀 기계 학습 모델(Linear Regression Machine Learning Model)과 장기-단기 주기 메모리 네트워크(LSTM), 그리고 회로형 순환 유닛(GRU)이다.

4.1 선형 회귀 분석 모델

선형 회귀 모델은 데이터 지점 사이에 선형 관계가 있다고 간주하고, 하나의 독립 변수를 이용하여 종속 변수의 값을 설명하거나 예측할 수 있는 모델을 구성하는 분석을 말한다. 다중 회귀 분석은 k개의 독립 변수가 존재할 때, Equation (1)과 같은 선형 방정식이 유효함을 확인하는 모델 구성 방식이다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \quad (1)$$

여기서 $X_1, X_2, \cdots, X_k$는 독립 변수, Y는 종속 변수, $\beta_0, \beta_1, \cdots, \beta_k$는 회귀 계수로서 미지의 상수이며, ϵ은 오차 항으로서, 통계적 추론을 위해 기댓값 0, 분산이 σ^2인 정규 분포를 따른다고 가정한다[9].

4.2 장기-단기 주기 메모리 네트워크

장기-단기 주기 메모리 네트워크(Long-Short Term Memory Network: LSTM Network)는 순환 신경망(Recurrent Neural Network: RNN)의 구조를 응용한 네트워크이다. LSTM 구조는 순환 신경망이 장기 과거 정보를 저장하지 못하는 단점을 극복하기 위해 등장했다. LSTM 구조는 3개의 게이트를 이용하여 이전 정보의 저장을 조절한다[6].

LSTM 네트워크는 총 3개의 게이트로 구성되어 있으며, 입력 게이트, 출력 게이트, 그리고 망각 게이트의 조합이다. LSTM 네트워크 모델은 이 3가지 게이트 정보를 조합하여 과거 정보를 얼마나 저장할 것이며 어떤 정보를 추가로 다음 게이트에 전달할 것인가를 판단한다.

Equation (2)에서 확인할 수 있듯, 입력 게이트에서는 현재 시각에서의 입력 데이터를 제어한다. 이때 x_i^t 는 t시점에 i번째 노드로부터 받은 입력 값이다. b_h^{t-1}은 $t-1$시점에 h번째 노드의 결과 값을 의미한다. s_c^{t-1}은 $t-1$시점에 c번째 노드의 셀 상태(cell state)를 의미한다. w는 weight로, 각 노드와 노드 사이를 잇는 가중치 값이다. 함수 f는 활성 함수(activation function)이다.

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{hl} b_h^{t-1} + \sum_{c=1}^{C} w_{cl} s_c^{t-1} \quad (2)$$
$$b_l^t = f(a_l^t)$$

Equation (3)의 출력 게이트는 현재 시점에서의 값을 출력 노드로 전달하는 역할을 한다.

$$a_w^t = \sum_{i=1}^{I} w_{iw} x_i^t + \sum_{h=1}^{H} w_{hw} b_h^{t-1} + \sum_{c=1}^{C} w_{cw} s_c^{t-1} \quad (3)$$
$$b_w^t = f(a_w^t)$$

마지막으로 망각 게이트에서는 Equation (4)와 같이 현재의 값을 셀 상태(cell state)에 저장하는 역할을 한다.

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1} \quad (4)$$

위 수식들을 조합한 연결망의 사용으로 LSTM 모델은 음향 신호와 같은 복잡한 시계열 데이터 정보를 현재의 분류 정보에도 영향을 미치도록 설계된다. LSTM 모델은 RNN이 가지고 있던 장기 기억 의존성 문제에 더욱 향상된 성능을 보인다.

4.3 회로형 순환 유닛

회로형 순환 유닛(Gated Recurrent Unit: GRU)은 재설정(reset)과 갱신(update)이라는 두 개의 게이트를 가진 LSTM 변형 모델이다[7].

$$z=\sigma(W_z x_t+U_z h_{t-1}+b_z) \quad (5)$$
$$r=\sigma(W_r x_t+U_r h_{t-1}+b_z)$$
$$m=\phi(W_m x_t+U_m(h_{t-1}\bigcirc r)+b_m)$$
$$h_t=(1-z)\bigcirc h_{t-1}+z\bigcirc m$$

Equation (5)에서 σ는 시그모이드 함수이며, x_t는 t시점에서 입력 값을, h_{t-1}는 $t-1$ 시점에서 출력 값을 의미한다. 그리고 $W_z, U_z, W_r, U_r, W_m, U_m$은 각 게이트와 셀 기억을 위한 가중치 행렬(weight matrix)이다. r은 재설정 게이트(reset gate)를 의미하는 것으로 직전 상태를 유닛의 입력에 반영하는 비율을 결정한다. z는 갱신 게이트(update gate)로 유닛의 출력에 직전 상태를 보존하여 반영하는 정보를 조정한다. ○ 기호는 원소별 곱셈(element-wise product)을 나타내는 기호다. ϕ 기호는 활성 함수(activation function)를 의미한다.

GRU는 LSTM 네트워크에서 망각 게이트와 입력 게이트를 하나의 갱신 게이트로 통합하고, 셀 상태와 은닉 상태(hidden state)를 하나로 통합하여 재설정 게이트를 만든 것으로, LSTM 모델보다 출력 값에 영향을 주는 게이트의 수를 하나 더 줄인 시계열 데이터 기반 기억 기계 학습 모델이다.

5. 실험 및 결과

5.1 선형 회귀 분석에 따른 예측 결과

Table 3은 딥 러닝을 사용한 시계열 예측 분석과 비교를 하기 위해 각각의 센서 노드의 추세 예측 분류율을 평균 낸 값을 결과로 최종 성능을 비교하였다.

Table 3에서 확인할 수 있듯 센서 노드 각각에 대한 분류율은 실험당 편차가 크다. 가장 높은 분류율을 보이는 선형 회귀 분석 모델은 이산화탄소이고(89.31%), 가장 낮은 분류율을 보이는 선형 회귀 분석 모델은 온도 모델이다(9.77%). 이 결과를 복합 모달리티를 분류하는 모델과 비교하기 위해서는 여섯 가지 단일 노드 선형 회귀 분석 모델을 통합할 필요가 있다. 그러나 복합 모달리티 분석 모델의 경우 모든 벡터가 테스트 데이터가 예측 결과와 일치해야 하고, 이 기준을 본 모델의 기준에 적용한다면 전체 모델 분류율은 가장 낮은 분류율을 보이는 모델보다(9.77%) 작거나 같게 된다. 따라서 본 실험의 테스트 데이터가 모수가 같다는 것을 근거로 전체 분류율의 평균을 내는 방식으로 실험 번호 분류율을 결정하였다. 따라서 단일 노드 선형 모델을 결합한 모델 분류율은 60.96%이다.

Table 3. Classification Using Linear Regression

Learning model	Component	Classification(%)
Linear Regression	Dust	85.49
Linear Regression	Light	72.80
Linear Regression	VOC	21.90
Linear Regression	CO2	89.31
Linear Regression	Temperature	9.77
Linear Regression	Humidity	86.50
Linear Regression	Total average	60.96

5.2 GRU를 이용한 시계열 데이터 예측 결과

GRU로 학습 모델을 구축하기 위해서는 다양한 초모수들의 결정이 필요하다. 예를 들어 은닉층의 계측 수, 은닉층의 노드 수, timestep, 활성함수, 배치의 크기 등이 있다. 이때 최적의 모델을 위한 초모수 조합은 timestep t=109를 기준으로 20회의 실험으로 결정했다.

GRU 1과 2의 은닉 노드는 1,270개이고, 활성 함수는 시그모이드(sigmoid)를 사용하였다. Dense Function 부분은 은닉 노드에서 시계열 데이터가 분할되어 표현된 데이터를 2차원 텐서로 압축하는 함수로 활성함수는 소프트맥스(softmax) 함수를 사용하였다. 마지막으로 최적화 함수는 ADAM (Adaptive Moment Estimation) 방식을 사용하였다[10]. 공기 품질 예측 시스템의 전체 구조는 Fig. 3과 같다. ADAM 최적화 함수는 최근의 딥 러닝에서 표준으로 가장 많이 사용되는 방법으로, 파라미터의 업데이트를 기울기 평균과 분산으로부터 직접 추정하고 편향(bias) 조정 항이 추가된 최적화 방식이다. 이 방식은 AdaGrad 알고리즘[11]과 RMSProp 알고리즘이 결합한 방법으로, 두 알고리즘은 각각 시간에 따라 잘 업데이트 되지 않는 파라메터의 학습률을 높이는 방식으로 동작한다. 결과적으로 ADAM 최적화 함수는 초모수의 변동이 학습률에 지대한 영향을 미치지 않으면서 학습률을 조정할 수 있도록 하는 최적화 알고리즘이다. Table 4는 최적의 초모수 조합을 바탕으로 한 GRU 학습 알고리즘이다.

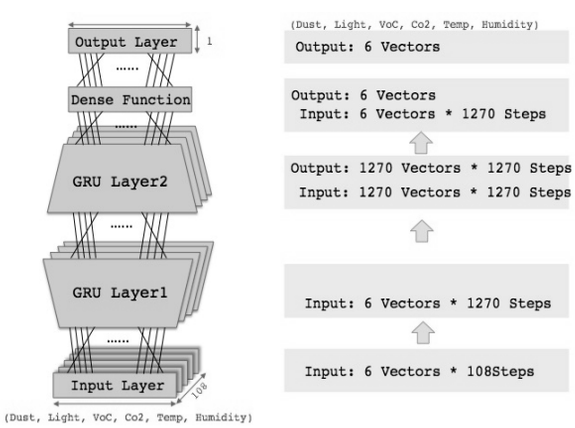

Fig. 3. Air Quality Regression Model Construction

Table 4. GRU Learning Algorithm

Algorithm 2. GRU Learning

Input : Training Dataset $\{x_{v\times t}, y_{v\times 1}\}_{train}$
Input/Output vectors 6
Hidden layer features 1270
Time step t

1. GRU_model() ← Train $\{x_{v\times t}\}_{train}$ with the corresponding event $\{y_{v\times 1}\}_{train}$ with parameters(6, 1270, t)
2. $\{y_{v\times 1}\}_{test}$ ← Predict event class based on GRU_model() ($\{x_{v\times t}, y_{v\times 1}\}_{test}$)

Output : Predicted classes $\{y_{v\times 1}\}_{test}$

실내 공기 품질 데이터를 시계열 복합 모달리티 분류 모델에 적용한 결과는 Table 5과 같다. Table 5의 실험은 GRU와 LSTM 네트워크에 대하여 수행하였고, 은닉 노드의 숫자와 은닉층의 숫자를 늘리거나 줄이는 방식으로 최적의 분류 모델을 탐색하였다.

실험 결과, GRU의 모델의 분류율이 가장 높은 것을 확인할 수 있었고(84.69%, 실험 번호13), 가장 낮은 시계열 복합 모달리티 분류 모델은 장기-단기 주기 메모리 네트워크에서 나타났다(60.23%, 실험 번호 18).

본 실험의 경우 일반적인 음성인식에서 사용하는 깊은 층위(deep layer) 구조를 사용하는 대신, 상대적으로 얕고 넓은 표현형의 층위를 사용하였을 때 분류 성능이 더욱 좋아지는 것을 확인할 수 있었다. 이것은 음성 데이터의 특징 벡터(feature vectors)가 다양하고 그 진폭이 한정적인 것과 달리, 본 데이터의 특징 벡터가 여섯 종인데 비해 변동 폭이 크고, 단위 시간당 데이터의 의존도가 크기 때문인 것으로 고려된다.

Table 6은 실험 모델로 사용하였던 Linear regression, LSTM, GRU에 대한 가장 높은 분류율을 나타낸 것이다.

Linear regression의 분류율은 60.96%, LSTM의 분류율은 70.13%, GRU의 분유율은 82.43%로 GRU의 분류율이 가장 높았다. 비교를 통해 GRU가 테스트 데이터 세트에 대해서 기존의 기계 학습 모델과 시계열 분석, 그리고 딥 러닝 기계학습 모델보다 얼마나 우수한지를 알 수 있다.

Table 6. Maximum Classification for the Learning Models

Learning model	Classification
GRU	82.43%
LSTM	70.13%
Linear Regression	60.96%

Table 5. Experiment to Construct Optimal Time Series Complexity Classification Model

Experiment number	Learning model	Layer	Number of layers	Hidden node	Number of hidden layers	Total number of layer	Classification
1	GRU	in/out	2	128	1	3	79.26%
2	GRU	in/out	2	32	3	5	77.40%
3	GRU	in/out	2	32	2	4	67.55%
4	GRU	in/out	2	32	4	6	73.32%
5	GRU	in/out	2	32	4	6	72.13%
6	GRU	in/out	2	256	2	4	81.96%
7	GRU	in/out	2	256	1	3	81.34%
8	GRU	in/out	2	384	1	3	80.03%
9	GRU	in/out	2	16	4	6	70.39%
10	GRU	in/out	2	6	4	6	60.31%
11	GRU	in/out	2	384	3	5	81.58%
12	GRU	in/out	2	1536	3	5	83.16%
13	GRU	in/out	2	1270	2	4	84.69%
14	GRU	in/out	2	512	2	4	83.80%
15	GRU	in/out	2	1024	2	4	82.43%
16	GRU	in/out	2	1024	3	5	82.43%
17	GRU	in/out	2	2048	-	2	Out of Memory
18	LSTM	in/out	2	32	3	5	60.23%
19	LSTM	in/out	2	32	4	6	61.22%
20	LSTM	in/out	2	1024	3	5	70.13%

6. 결론 및 향후 과제

본 연구에서는 공기 품질 데이터, 그중에서도 여섯 종의 주요 공기 품질을 측정하고, 예측하는 모델을 제안하였다. 실내 공간 공기 품질은 대기라는 범주 안에서 다양한 성분의 데이터가 혼합된 자료형이다. 이 데이터가 변인에 따라 다양하게 상호 작용하고 패턴이 검출된다는 가설 아래 복합 모달리티 분석 모델을 실험에 적용하였고, 그 결과 논문의 모델이 단일 선형 회귀 분석법보다 우수한 성능의 분류 능력을 보인다는 것을 확인할 수 있었다.

향후에는 여섯 가지 대기 품질 측정 기준뿐 아니라 방사성 동위 원소, 제논 검출기 등 약 열 가지 종류의 실내 공간 공기 품질 지표를 가지고 실험을 하는 모델을 구축해 볼 수 있을 것이다. 그리고 각 센서 노드 하나에 대하여 나머지 센서 노드들이 얼마나 많은 의존도를 가지고 영향을 주고받는지 측정하고, 이 영향력을 수치로 검증하는 분석을 진행할 수 있을 것이다.

References

[1] R. Allen, T. Larson, L. Sheppard, L. Wallace, and L. J. S. Liu, “Use of Real-Time Light Scattering Data to Estimate the Contribution of Infiltrated and Indoor-Generated Particles to Indoor Air,” *Environmental Science & Technology*, Vol.37, No.16, pp.3484-3492, 2003.

[2] H. K. Lai, L. Bayer-Oglesby, R. Colvile, T. Götschi, M. J. Jantunen, N. Künzli, E. Kulinskaya, C. Schweizer, and M. J. Nieuwenhuijsen “Determinants of indoor air concentrations of $PM_{2.5}$, black smoke and NO_2 in six European cities (EXPOLIS study),” *Atmos. Environ.*, Vol.40, No.7, pp.1299-1313, 2006.

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, Vol.9, Issue 8, pp.1735-1780, 1997.

[4] A. Temko and N. Climent, “Classification of acousticevents using SVM-based clustering schems,” *Pattern Recognition*, Vol.39, No.4, pp.682-694, 2006.

[5] T. Zhao and H. Xue. “Regression Analysis and Indoor Air Temperature Model of Greenhouse in Northern Dry and Cold Regions,” *International Conference on Computer and Computing Technologies in Agriculture*, Springer, Berlin, Heidelberg, 2010.

[6] A. Graves, “Supervised sequence labelling with recurrent neural networks,” Springer, Vol.385, 2012.

[7] Cho Kyunghyun, et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv preprint arXiv:1406.1078, 2014.

[8] R. Jozefowicz, W. Zaremba, and I. Sutskever, “An empirical exploration of recurrentnetwork architectures,” *Proceedings of the 32ndInternational Conference on Machine Learning (ICML-15)*, 2015.

[9] R. E. Walpole and R. H. Myers, “Probability and Statistics for Engineers and Scientists,” New York: Macmillan, ISBN 10: 0024241709, ISBN 13: 9780024241702, 1985.

[10] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[11] J. Duchi, E. Hazan, and Y. Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization,” *Journal of Machine Learning Research*, Vol.12, pp.2121-2159, 2011.

안 재 현

e-mail : jaehyunahn@sogang.ac.kr

2015년 서강대학교 컴퓨터공학과(학사)

2017년 서강대학교 컴퓨터공학과(석사)

2017년～현 재 버즈니(주) 연구원

관심분야 : 기계학습, 인공지능, 패턴인식

신 동 일

e-mail : shindi91@sogang.ac.kr

2017년 서경대학교 전자공학과(학사)

2017년～현 재 서강대학교 컴퓨터공학과 석사과정

관심분야 : 기계학습, 인공지능, 패턴인식

김 규 호

e-mail : ekyuho@sogang.ac.kr

1984년 서울대학교 전자계산기공학과 (학사)

1986년 서울대학교 전자계산기공학과 (석사)

1996년 한국과학기술원(박사)

2014년～현 재 서강대학교 산학협력중점 교수

관심분야 : IoT, 빅 데이터, 센서 데이터

양 지 훈

e-mail : yangjh@sogang.ac.kr

1983년 서강대학교 전산학과(학사)

1989년 Iowa State University Computer Science(석사)

1999년 Iowa State University Computer Science(박사)

2002년~현 재 서강대학교 컴퓨터공학과 교수

관심분야 : 기계학습, 인공지능, 패턴인식, 데이터마이닝, 생물정보학