JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 22, No. 4: 595~601, Apr. 2018

미세먼지 수치 예측 모델 구현을 위한 데이터마이닝 알고리즘 개발

차진욱[1] · 김장영[2*]

Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model

Jinwook Cha[1] · Jangyoung Kim[2*]

[1]Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea
[2*]Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

요 약

최근 미세먼지 수치가 급격히 상승함에 따라 이에 대한 관심도가 굉장히 높아지고 있다. 미세먼지의 노출은 호흡기 및 심혈관계 질환의 발생과 관련이 있으며, 사망률도 증가시키는 것으로 보고되고 있다. 뿐만 아니라, 산업현장에서도 미세먼지에 대한 피해가 속출한다. 그러나 현대인의 삶에서 미세먼지 노출은 불가피하다. 그러므로 미세먼지를 예측하여, 이에 대한 노출을 최소화하는 것이 건강 및 산업 피해축소에 가장 효율적인 방법일 것이다. 기존의 미세먼지 예측 모델은 농도 수치가 아닌 미세먼지의 농도 범위에 따라 좋음, 보통, 나쁨, 매우 나쁨으로만 나누어 예보하고 있다. 본 논문은 기존의 실제 기상 및 대기 질 데이터를 이용, 기계학습 알고리즘인 Artificial Neural Network (ANN)알고리즘과 K-Nearest Neighbor (K-NN)알고리즘을 상호 응용하여 미세먼지 수치 (PM 10)를 예측하고자 하였다.

ABSTRACT

Recently, as the fine dust level has risen rapidly, there is a great interest. Exposure to fine dust is associated with the development of respiratory and cardiovascular diseases and has been reported to increase death rate. In addition, there exist damage to fine dusts continues at industrial sites. However, exposure to fine dust is inevitable in modern life. Therefore, predicting and minimizing exposure to fine dust is the most efficient way to reduce health and industrial damages. Existing fine dust prediction model is estimated as good, normal, poor, and very bad, depending on the concentration range of the fine dust rather than the concentration value. In this paper, we study and implement to predict the PM10 level by applying the Artificial neural network algorithm and the K-Nearest Neighbor algorithm, which are machine learning algorithms, using the actual weather and air quality data.

키워드 : 미세먼지, 데이터 마이닝, ANN 알고리즘, K-NN 알고리즘

Key word : Fine Dust, Data Mining, ANN Algorithm, K-NN Algorithm

Received 21 February 2018, **Revised** 8 March 2018, **Accepted** 30 March 2018

*** Corresponding Author** Jangyoung Kim (E-mail: jykim77@suwon.ac.kr, Tel: +82-31-229-8345)
Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2018.22.4.595 pISSN:2234-4772

Ⅰ. 서 론

1.1. 미세먼지의 정의

미세먼지는 아황산가스, 질소 산화물, 납, 오존, 일산화탄소 등과 함께 수많은 대기오염물질을 포함하는 대기오염 물질로 자동차, 공장 등에서 발생하여 대기 중 장기간 떠다니는 입경 10㎛ 이하의 미세한 먼지이며, PM10이라 한다. 미세먼지는 부유분진, 입자상물질 등으로도 불리며 명칭에 따라 약간씩 다른 의미를 가지고 있다.

1.2. 미세먼지의 원인

미세먼지가 생성되는 원인은 다양하게 존재한다.

일상생활에서 생성되는 미세먼지의 요인은 예측이 불가능할 정도로 굉장히 많다. 일상생활에서 발생 하는 먼지뿐만 아니라, 산업 활동에 있어서도 미세먼지는 발생한다. 위에 언급한 것 외에도 우리나라 자체에서 발생하는 미세먼지와 더불어, 외국에서 들어오는 미세먼지 또한 무시할 수 있는 수준이 아니다. 중국 공업지대에서 편서풍을 타고 날아오는 산업먼지와, 각종 사막과 고원에서 발생하여 많은 양의 큰 입자와 작은 입자의 먼지를 동반 수송하는 황사에 의해 미세먼지 수치가 급격히 증가하기도 한다.

1.3. 미세먼지가 인체에 미치는 영향

미세먼지의 농도가 증가할수록 전체 사망률과 심장혈관 및 호흡기계 질환으로 인한 사망률이 증가한다 [1].미세먼지 장기 노출과 폐암 및 심혈관질환 사망률의 연관성이 높고 단순히 호흡기계 뿐만 아니라 암과 심혈관계 등과 같은 전신 질환과도 깊은 관련이 있다 [2].

Ⅱ. 관련 연구 및 배경 지식

2.1. 관련 연구

미세먼지에 영향을 끼치는 기상인자로 일기유형, 기온, 상대습도, 풍속, 풍향으로 설정한 후, Data Mining Tool인 WEKA를 활용하여 Machine-Learning알고리즘을 사용 미세먼지 수치를 예측한 연구가 있다. 기상인자를 변수로 두어, 각기 다른 알고리즘에 적용하여 미세먼지 예측 정확도를 비교하였다 [3].

2.2. 최근 동향

미세 먼지는 정확한 발생원에 대한 조사에 한계가 있기 때문에, 때문에 물리적 수치모델 사용과, 화학수송모델을 적용한 현실적인 예측이 용이하지 않다. 그러한 이유로 대기질 측정망자료와 기상 관측자료와의 상관관계를 분석한 통계 모델을 이용한 예측기법이 주로 사용된다. 때문에 미세먼지 예보를 담당하는 환경부 소속의 국립환경과학원에서는 대기오염도 측정자료와 지표 및 고층 기상대 관측자료, 기상예보자료를 이용하여, 신경망 모형의 오류 역전파 학습기능을 갖는 다층 인식 자신경망 모형을 기반으로 회귀 분석과 의사결정 모형을 결합한 알고리즘으로 미세 먼지를 수치를 예측하고 있다 [4].

2.3. ANN Algorithm

인공신경망(Artificial Neural Network, ANN)은 기계학습과 인지과학에서 생물학의 신경망(동물의 중추신경계중 특히 뇌)에서 영감을 얻은 통계학적 학습 알고리즘이다 [5]. 다른 층의 뉴런(노드)들 사이의 연결 패턴, 연결의 가중치를 갱신하는 학습과정, 마지막으로 뉴런의 가중 입력을 활성화도 출력으로 바꿔주는 활성화 함수, 이 세 가지의 인자를 이용해 정의된다. 여기서 활성화 함수는 입력 신호의 총합이 활성화를 일으키는지를 정하는 역할을 하게 된다. 신경망의 신호 전달 방법을 설명하기 전에 퍼셉트론에 대해 먼저 선행이 되어야한다. x_1과x_2라는 두 신호를 입력 받아 y를 출력하는 퍼셉트론이다.

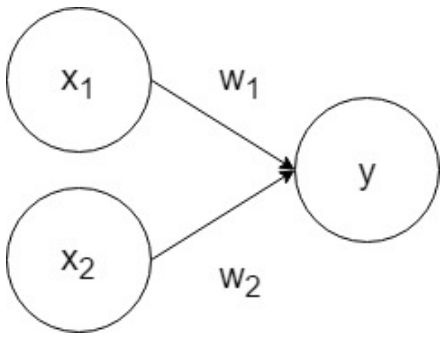

Fig. 1 Example of Perceptron

이를 수식으로 나타내면 다음과 같다.

y = 0일 경우

$$y = a + w_1x_1 + w_2x_2 \leq 0 \quad (1)$$

y = 1일 경우

$$y = a + w_1x_1 + w_2x_2 > 0 \quad (2)$$

그림 1에는 명시하지 않았지만 여기서 a는 편향을 나타내는 매개변수이며 얼마나 쉽게 활성화되느냐를 제어한다. w_1과w_2는 각 신호의 가중치를 나타내는 매개변수로, 각 신호의 영향력을 제어한다. 그림 1에서 x_1, x_2라는 2개의 신호가 각각 뉴런에 입력되어, 각 신호에 가중치를 곱한 후, 다음 뉴런에 전달된다. 다음 뉴런에서는 이들 신호의 값을 더하여 그 합이 0을 넘으면 1을 출력하고 그렇지 않으면 0을 출력하게 된다. 이 함수를 $h(x)$라고 하였을 때, 그림 1을 다음과 같은 수식으로 표현할 수 있다.

$$y = h(b + w1x1 + w2x2) \quad (3)$$

$$h(x) = \begin{cases} 0\ (x \leq 0) \\ 1\ (x > 1) \end{cases} \quad (4)$$

위의 식 (3)은 입력 신호의 총합이 $h(x)$라는 함수를 거쳐 변환되어, 그 변환된 값이 y의 출력이 되는 것을 보여준다. 그리고 입력이 0이 넘으면 1을 출력하고, 넘지 않을 경우 0을 출력하게 된다. (4)의 $h(x)$는 활성화 함수로, 입력 신호의 총합을 출력 신호로 변환하는 함수이다. 이러한 단순 퍼셉트론(단층 네트워크에서 계단 함수 - 임계값을 경계로 출력이 바뀌는 함수)이 여러 층으로 구성되는 것을 다층 퍼셉트론(신경망)이라고 한다. 신경망은 시그모이드 함수 등의 활성화 함수를 사용하는 네트워크이다. 가령 3층 신경망을 도식화하면 그림 2처럼 된다. 여기에서 가장 왼쪽 줄을 input layer, 맨 오른쪽 줄을 output layer, 가운데 줄을 hidden layer이라고 통칭한다. 은닉층의 뉴런은 입력층이나 출력층과 달리 사람 눈에 보이지 않는다 [6].

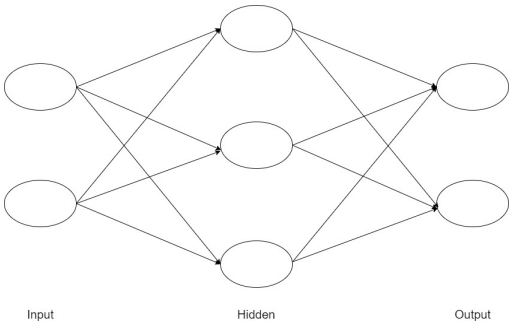

Fig. 2 Example of Artificial Neural Networks

2.4. K-Nearest Neighbor Algorithm

패턴 인식에서, K-NN알고리즘은 분류나 회귀에 사용되는 방식이다. 두 경우 모두 입력이 특정 공간 내에 k개의 가장 가까운 훈련데이터로 구성이 되는데, 출력은 K-NN이 분류로 사용되었는지 혹은 회귀로 사용되었는지에 따라 다르다. K-NN 분류에서 출력은 객체 k개의 최근접 이웃 사이에서 가장 공통적인 항목에 할당되는 객체로 과반수 의결에 의해 분류된다. 이 때 k는 양의 정수이며, 만약 k = 1이라면 객체는 단순히 하나의 최근접 이웃이다. K-NN 회귀에서 출력은 객체의 특성 값이며, 이 값은 k개의 최근접 이웃이 가진 값의 평균이다 [7]. 분류와 회귀 모두 더 가까운 이웃일수록 평균에 더 많은 기여를 하기 때문에, 이에 대한 가중치 여부를 적절히 주는 것이 유용하다. 즉, 더 가까운 이웃에 더 많은 가중치를 주는 것이 더 정확한 결과를 가져올 수 있다. 가령 어떠한 점에서 이웃까지의 거리가 d라고 할 때, 각각의 이웃에게 1/d의 가중치를 주는 것이다. 예를 들어 그림 3을 보면 a라는 점과 b, c라는 이웃이 있다. 각각 거리를 1과 10이라고 하면, b에는 1/1의 가중치를, c에는 1/10의 가중치를 주는 것이다.

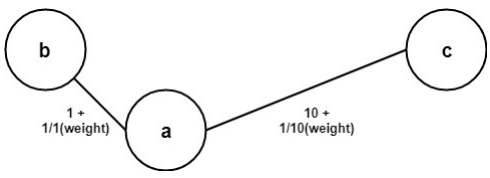

Fig. 3 Example of Weight

기존 데이터 중 가장 유사한 k개의 데이터를 이용하여 새로운 데이터를 예측하는 K-NN은 k의 수치에 따라 결과가 확연히 달라지기 때문에, k의 개수를 적절하게

부여하는 것이 관건이다. 너무 작은 수의 k는 과적합의 우려가 벌어지며, 너무 큰 수의 k는 데이터 구조 파악의 어려움이 있다 [8].

Ⅲ. 제안 알고리즘

본 논문에서는 ANN algorithm과 K-NN algorithm을 상호 응용하여 미세먼지 농도 수치를 예측하려 한다 (그림 4).

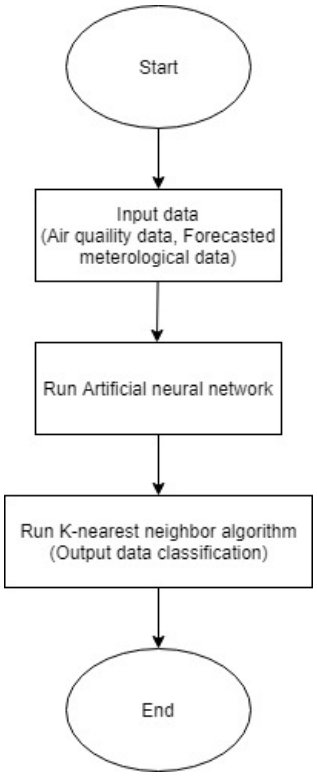

Fig. 4 Proposed model of flow chart

input data인 air quality data는 대기질 측정 자료로 미세먼지에 영향을 주는 SO2, NO2, CO, O3.에 대한 데이터이다. forecasted meterological data는 강수량, 풍속, 습도, 일조량, 기온의 기상 인자 데이터를 말한다. 이러한 데이터들을 바탕으로 대기질 인자들과 기후인자들을 앞서 설명한 Artificial Neural Network의 input data layer로 입력하였다. 각 인자들은 일평균으로 일괄 처리하였고, 인자들의 상관도를 고려하여 활성화 함수를 이용, 가중치를 부여하였다. 이 과정은 분야가 다르기 때문에 많은 시간과 배경지식이 요구되었다. 2014년 1월 1일부터 2017년 6월 31일까지, 3년 반 동안의 데이터를 학습시켜 출력된 데이터를, 앞서 설명한 K-Nearest Neighbor algorithm을 이용하여, 분류 하였다. 그리하여 Artificial Neural Network알고리즘만 이용 했을 때와, K-Nearest Neighbor algorithm알고리즘만 사용 하였을 때, 그리고 제안 알고리즘에 의한 실험 결과 데이터 등을 이용하여 알고리즘 정확도를 계산하였다.

Ⅳ. 실험과정 및 실험결과

4.1. 실험과정 (구현내용)

본 논문에서의 실험과정은 다음과 같다.

4.1.1. 데이터 수집 및 가공

수집된 데이터는 기상청에서 제공한 대기질 측정자료 통계와 기상 실제측정 데이터, 그리고 기상 예보 데이터이다. 제안 알고리즘엔 앞서 말한 10개의 인자가 사용된다. 10개의 인자는 각각 일평균 수치 값으로 통일하여 대입하였다. 2014년부터 2016년 6월까지의 데이터는 각각 모두 입력하여 학습하였다. 그 후 2017년 6월 30일까지 1년 데이터를 사용 및 예측하여 비교하였다. 기상청에서 공개한 데이터들을 .xslx파일로 수집하였고, 이를 알고리즘에 입력하기 위하여 일차적으로 가공하였다. 또한 알고리즘의 결과를 추출하기 위하여 각 인자들의 관계에 대한 상관관계와 상당히 많은 배경지식이 요구되었다. 또한 기상인자 데이터, 대기질 데이터, 그리고 기상 예보데이터와 대기질 예보 데이터는 각각 배포방식과 정렬상태가 달라, 이를 통합하여 가공하는데 많은 시간이 소모되었다. 또한 기상청에서 발표한 예보 데이터와, 실제 측정 데이터, 그리고 제안 알고리즘에 의한 실험 결과 데이터를 비교하기 위하여 같은 실험 환경을 구축하는 데에도 많은 어려움이 있었다. 실험에 사용된 데이터 수와 출처는 다음과 같다 (표 1).

Table. 1 Data Collection Source

Source		Entry
http://www.kma.go.kr	2014 Data : 365 * 14	csv, xlsx
	2015 Data : 365 * 14	
	2016 Data : 366 * 14	
	2017 Data : 161 * 14	
Total	17598	

4.1.2. ANN

Input Layer	Factors	1	VAR00001
		2	VAR00002
		3	VAR00003
		4	VAR00004
		5	VAR00005
		6	VAR00006
		7	VAR00007
		8	VAR00008
		9	VAR00009
		10	VAR00010
		11	VAR00011
		12	VAR00012
		13	VAR00013
	Number of Units[a]		2478
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		15
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	PM10
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares

Fig. 5 Result of ANN

ANN알고리즘을 가동하기 위해선 먼저 Input layer에서 hidden layer에 대한 weight와 hidden layer에서 output에 대한 weight를 알아야 한다. weight를 구하기 위해, PM10을 포함한 총 14개의 변수들을 통하여 다중 회귀 분석을 실시하여 값을 추출하였다. 그림5를 보면 추출한 weight를 이용하여, input layer에 13개의 노드, hidden layer에서의 15개의 노드, 그리고 output layer에 1개의 노드로 ,총 3-layer로 이루어진 network를 구성하였다는 것을 알 수 있다. (그림 5)

4.1.3. K-NN

그 이후, K-NN알고리즘을 이용하여 ANN알고리즘에 의한 output을 분류하였다. 여러 실험에 걸친 후에 K = 9 일 때 가장 높은 정확도를 가진다는 것을 알 수 있었다.(그림 6)

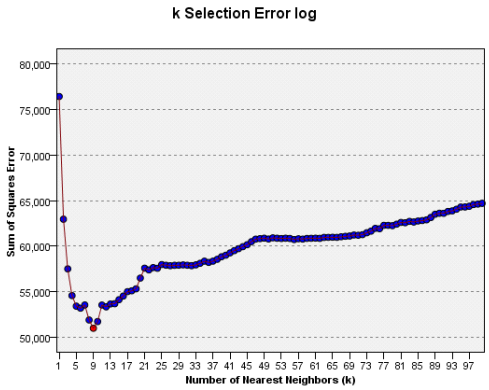

Fig. 6 K selection Error log

그림 7은 k = 9일 때의 예측 결과이다.

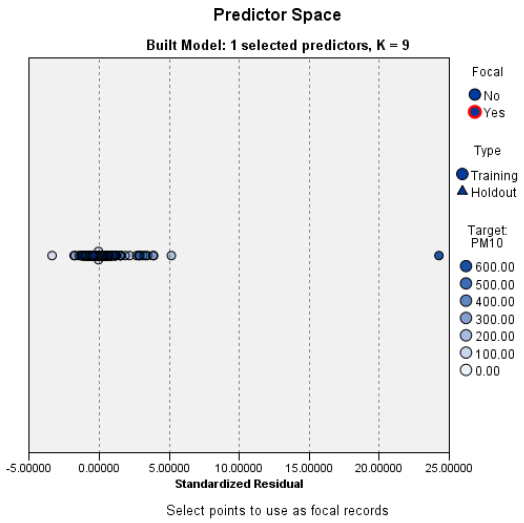

Fig. 7 The prediction result when k = 9

4.2. 실험 결과

표 2는 그림 에 나온 최종 결과를 각 50일씩 나누어 PM10수치를 예상한 차트이다. Actual: 실제 PM10수치 Proposed: 제안모델 예상 PM10수치, ANN: ANN 알고리즘 예상 PM10수치, K-NN: K-NN알고리즘 PM10.

Table. 2 PM10 Comparison

PM10 date	Actual (PM10)	Proposed (PM10)	ANN (PM10)	K-NN (PM10)
50	44	62	21	45
100	63	49	66	60
150	77	47	120	61
200	14	28	15	75
250	31	13	32	76
300	16	31	35	81
350	53	54	42	65

표 3에는 미세먼지 수치 예측에 대한 Artificial neural network와 K-nearest neighbor, 그리고 Proposed Model의 정확도와 오차율을 비교하여 표기하였다. Proposed Model이 다른 두 가지의 알고리즘에 비해 더 높은 정확성과 낮은 오차율을 보임을 알 수 있다.

Table. 3 Accuracy and error rate comparison

	ANN	K-NN	Proposed Model
Accuracy(%)	62.2794	58.4151	83.4221
Error rate(%)	13.5364	24.2794	2.9943

그림8은 최종결과를 비교하여 그래프로 나타낸 결과이다. ANN과 K-NN보다 제안 모델이 실제 데이터와 가장 비슷한 추이와 지표를 나타내는 것을 확인할 수 있다.

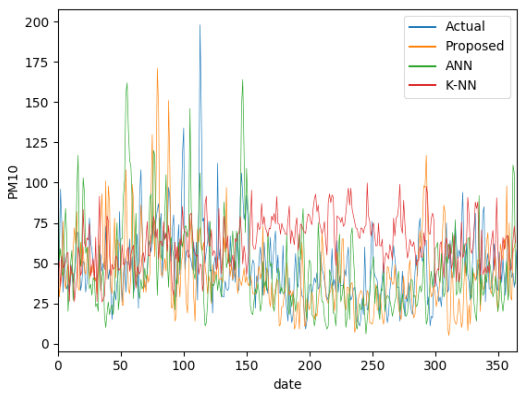

Fig. 8 Comparison graphs

Ⅴ. 결론 및 향후과제

본 논문에서는 기계학습 알고리즘들을 상호 응용하여 데이터 마이닝을 하였고, 서울지역의 미세먼지를 대상으로 농도 수치를 예측해보는 새로운 모델을 제안해 보았다. 그러나, 대기질 예측이 100%정확한 데이터가 아니기 때문에 이를 바탕으로 한 예측 또한 원하는 정확도를 얻지 못한다는 한계가 있었다. 만약 기상 예보나 대기질 예측 정확도가 상승한다면, 제안한 모델을 이용하여 미세먼지 예측도 또한 자연히 상승할 것으로 기대해본다. 미세 먼지는 대기질과 기상뿐만 아니라, 배출, 황사, 장마 등으로 인한 영향도 상당히 크다. 향후 연구로는 이러한 요소들도 고려하여 더 정확한 예측이 가능할 것으로 예상된다 [9, 10].

ACKNOWLEDGEMENT

The paper was supported by the research grant of the University of Suwon in 2017.

REFERENCES

[1] J. S. Oh, S. H. Park, M. K. Kwak, C. H. Pyo, K. H. Park, , H. B. Kim, S. Y. Shin, and H. J. Choi, “Ambient Particulate Matter and Emergency Department Visit for Chronic Obstructive Pulmonary Disease,” Journal of *The Korean Society of Emergency Medicine*, vol. 28, no. 1, pp. 32-39, Jan. 2017.

[2] H. J. Bae, “Effect of Short-term Exposure to PM10 and PM2.5 on Mortality in Seoul,” Journal of *Korea Society of Environmental Health* vol.40, no.5, pp. 346-354, May 2014.

[3] B. D. Oh, J. H. Park, and Y. S. Kim, "Prediction of the concentration of PM10 using Machine-Learning," *Journal of Korea Information Science Society*, vol. 20, no. 12, pp. 1674 - 1676. Dec. 2016.

[4] Y. S. Koo, H. Y. Yun, H. Y. Kwon, and S. H. Yu, "A Develope of PM10 Forecasting System," *Journal of Korean Society for Atmospheric Environment*, vol. 26, no. 6, pp. 666-682, Nov. 2010.

[5] Artificial neural network [Internet] Available : http://en.wikipedia.org/wiki/Artificial_neural_network

[6] W. Goki, *Study of deep learning from scratch* translated

Korean, Hanbit Media, Seoul. 2016.
[7] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, Mar. 1992.
[8] M. L. Zhang, and Z. H. Zhou, "A k-Nearest Neighbor Based Algorithm for Multi-label Classification," *Granular Computing, 2005 IEEE International Conference on Beijing: China*, pp. 718 - 721, May 2005.
[9] Gitae Kim, "Data Mining for Spam Email Classification," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, ISSN:2383-5285, Vol.6, No.7, pp. 37-47, July 2016.
[10] Jin-Ho Noh, and Han-Ho Tack, "The Implementation of the Fine Dust Measuring System based on Internet of Things(IoT)," *The Korea Institute of Information and Communication Engineering, Journal of the Korea Institute of Information and Communication Engineering*, vol. 21, no. 4, pp. 829-835, April 2017.

차진욱 (Jinwook Cha)

2017년 3월: 수원대학교 컴퓨터학과 석사 재학

※관심분야 : Big data, Networks

김장영(Jangyoung Kim)

2005년 2월: 연세대학교 컴퓨터과학 공학사
2010년 5월: Pennsylvania State Univ. 공학석사
2013년 7월: State University of New York 공학박사
2013년 8월: University of South Carolina 조교수
2014년 3월: 수원대학교 컴퓨터학부 조교수
※관심분야 : Big data, Cloud computing, Networks