LGBM을 통한 대기 오염 정보 데이터 결측치 예측에 대한 연구

A Study on The Prediction of Missing Value in Data of Air Pollution Using LightGBM

저자
(Authors)

김선욱, 방준일, 홍성은, 김화종
Kim Seon Uk, Bang Jun Il, Hong Seong Eun, Kim Hwa Jong

출처
(Source)

한국통신학회 학술대회논문집 , 2019.6, 1029-1030(2 pages)
Proceedings of Symposium of the Korean Institute of communications and Information Sciences , 2019.6, 1029-1030(2 pages)

발행처
(Publisher)

한국통신학회
Korea Institute Of Communication Sciences

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09234687

APA Style

김선욱, 방준일, 홍성은, 김화종 (2019). LGBM을 통한 대기 오염 정보 데이터 결측치 예측에 대한 연구. 한국통신학회 학술대회논문집, 1029-1030

LGBM을 통한 대기 오염 정보 데이터 결측치 예측에 대한 연구

김선욱, 방준일, 홍성은, 김화종*
강원대학교, *강원대학교
king950411@gmail.com, tkfka965@gmail.com,
sungkenh@gmail.com, *hjkim3@gmail.com

A Study on The Prediction of Missing Value in Data of Air Pollution Using LightGBM

Kim Seon Uk, Bang Jun Il, Hong Seong Eun, Kim Hwa Jong*
Kangwon Univ., *Kangwon Univ.

요 약

최근 미세먼지, 초미세먼지 등의 대기 오염 문제가 지속적으로 대두되고 있다. 이러한 대기 오염 문제를 해결할 수는 없지만 이를 미리 알릴 수 있다면 좋을 것이다. 2015년부터 2017년까지의 일자 별로 기록된 대기 오염 정보 데이터와 그래디언트 부스팅 모델 중 하나인 Light GBM을 이용하여 대기 오염 정보 데이터 간의 관계를 분석하여 패턴을 학습시키고, 이를 이용하여 2018년의 미세먼지에 대해 예측해본 결과를 실제 관측 값과 그래프에 동시에 표시하여 차이를 알아보기 쉽게 시각화하고, 규칙 기반 분류를 통해 정확도를 확인한 결과, 79%라는 정확도가 나왔다. 그러나 이상치가 발견되는 점과, 다차원 임베딩을 사용하기에 부족한 학습에 쓰일 데이터에 대해서는 개선해야할 필요가 있다. 이 논문은 최근에 구현된 그래디언트 부스팅 모델인 Light GBM에 대한 지식 습득과 데이터 패턴 분석, 다차원 임베딩을 이용한 유사 데이터 추출 및 성능 향상 방안, 대기 오염 정보 데이터의 결측 치 문제의 해결방안을 제시하는 것을 목표로 하는 논문이다.

Ⅰ. 서 론

최근 미세먼지, 초미세먼지 오존 등의 대기 오염에 대한 관심이 커지면서, 대기 기상 상태에 따라 야외 활동을 결정하게 되고 있다. 이로 인하여 사람들이 날씨를 확인하듯이 대기 기상 상태도 확인하고 있으며, 이러한 대기 오염 정보를 예측하기 위한 연구도 활발히 진행되고 있다.

예측에 활용되는 특성(변수)도 굉장히 다양한데, 풍향, 풍속, 온도 등에 대한 일반 기상 정보뿐 아니라, 그동안 관측되어온 미세먼지(PM_{10}), 초미세먼지($PM_{2.5}$), 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2) 등에 대해 여러 위치의 측정소에서 관측하고 기록된 대기 오염 정보 데이터도 쓰인다.[1]

해당 대기 오염 정보를 관측하는 측정기기는 정확성 확보를 위해 일정한 주기의 교정, 수리, 점검이 필수적이다. 이러한 활동이 진행 중일 때, 측정기기가 측정을 중지하므로 해당 기간 동안에는 자료 생산이 중단된다. 해당 기록된 데이터는 곳곳에 기록이 되지못한 결측 치가 존재한다. 이러한 결측 치는 대기 오염 정보 상태를 예측에 치명적이다. 정확한 대기 오염 정보 상태 예측을 위해 이러한 결측 치 문제가 해결되어야 한다.

본 논문의 연구의 목표는 기존의 대기 오염 정보 데이터의 특정 항목 결측 치를 예측하기 위한 다차원 임베딩 기술 활용과 유사 데이터 추출의 속도, 정확도 등의 예측 모델의 성능 향상을 목표로 연구를 진행하였다.

예측에 필요한 학습 모델은 Light GBM이라는 그래디언트 부스팅 모델을 사용하였으며, Light GBM은 그래디언트 부스팅(Gradient Boosting)의 모델 중 하나로, 기존의 XG Boost에서 과도한 노드 생성을 보완한 모델이다. 정확도 평가는 세계보건기구에서 제시한 기준으로 실제 미세먼지 수치를 분류하여 정확도를 계산했다.

Ⅱ. 본론

기계학습에서 학습기(Learner)는 그 정확도와 복잡성에서 약한 학습기와(Weak Learner)와 강한 학습기(Strong Learner)로 나뉜다. 비교적 단순하고 약한 학습기를 결합해서 정확한 결과를 낼 수 있는 강한 학습기를 만들어 낼 수 있는데, 이를 부스팅(Boosting)이라 한다. 다수의 모델 중 이전 모델의 약점을 그 다음 모델이 보완하는 과정을 반복하는 원리다.[2]

현재 모델이 예측한 값과 실제 값의 차이를 추정하여 이를 타겟으로 다음 모델을 피팅(Fitting)한다. 여기서 피팅된 새로운 모델을 기존의 모델이 흡수하여 편향(Bias)를 줄인다. 그리고 다시 예측한 값과 실제 값의 차이를 추정하고 모델 피팅 후, 다시 기존 모델에 흡수시키기를 반복한다. 이를 Gradient Boosting이라 한다.[3]

본 연구에서 학습에 쓰인 데이터는 서울특별시의 중구에 소재한 측정소에서 측정한 2015년부터 2017년까지의 대기 오염 정보 관측 값으로, 한국환경 공단의 에어코리아에서 제공받았다.

분석을 위한 학습에 쓰인 데이터는 일자에 따라 기록된 비선형 데이터이다. 데이터 중 일부 항목인 미세먼지의 그래프지만, 미세먼지 이외의 항목인 초미세먼지($PM_{2.5}$), 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2)에 대한 값들도 마찬가지로 비선형 데이터이다. 예측에 쓰일 파라미터 값은 해당 일자의 대기 오염 정보 데이터 중 미세먼지의 값을 제외한 다른 항목들의 관측 값으로, 마찬가지로 비선형 데이터이다.

그러므로 예측에 쓰일 모델은 제공되는 파라미터에 대한 유연성이 필요하다. 그에 따라 파라미터에 대해 새로운 모델을 흡수하며 손실함수를 줄여나가는 Light GBM이라는 Gradient Boosting 모델을 사용하였다. Light GBM의 구조는 아래 그림 1과 같다.

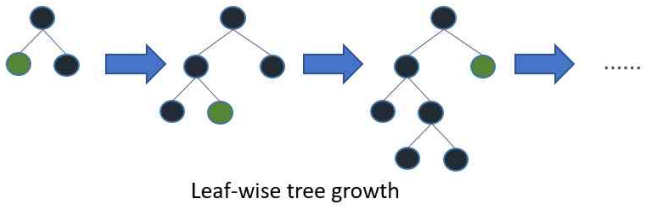

그림 1 . Light GBM 구조

Light GBM은 각 노드에서 분기점이 나뉠 때, 매번 2개씩 분리되었던 이전의 GBM과 달리, 모든 노드를 분리하는 게 아닌 잘 맞는 노드를 기준으로 분리하며, 잘 맞지 않는 노드는 분기점으로 선택되지 않는다. 다만, 해당 부적합 노드는 학습이 진행됨에 따라 다시 분기점이 될 수 있다. 이 GBM은 Gradient를 기준으로 인스턴스를 샘플링하는 GOSS (Gradient-Based One-Side Sampling) 방법을 사용한다. 이 방법은 큰 경사도의 인스턴스를 보유하는 동시에 작은 경사도의 인스턴스를 무작위로 추출한다. 그리고 이러한 방법을 이용하여 인스턴스를 필터링한다. 이 필터링을 통해 분할 값을 찾아내는 것이 Light GBM 모델이다.[4]

Light GBM의 객체 선언을 위한 objective 파라미터에는 regression(선형 회귀), binary(이진), multiclass(다중 클래스)가 모두 가능하다. 이를 사용하기 위한 파라미터로는 objective에 multiclass(다중클래스)를, random_state는 5를 주었으며, x, y는 각각 학습에 필요한 데이터이고, x_test는 x에서 뽑아낸 테스트를 위한 데이터이다.

이 모델을 통하여 분석한 결과에 대한 그래프는 그림 2와 같다.

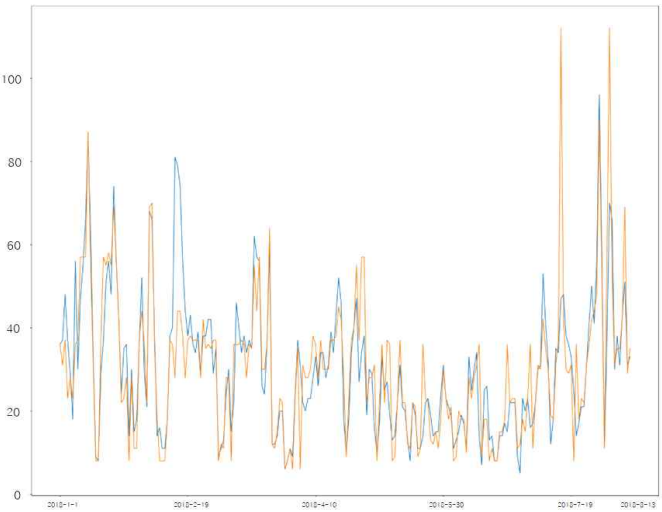

그림 2. 결과 그래프

그림 2의 그래프는 2018년 중 일부의 미세먼지에 대해 예측한 값과 해당 일자에 대한 실제 미세먼지 관측 값을 표현한 것이다. x축은 미세먼지의 수치, y축은 시간의 경과를 일자별로 나열한 것이다. 파란색은 실제 관측된 값이고, 주황색은 학습을 통해 예측한 값이다.

수치를 정확하게 맞추는 것은 매우 어려운 일이다. 그래프를 보면 수치를 정확하게 맞추지는 못하지만 비슷한 값까지의 예측은 보여 지고 있다.

그러므로 정확도에 대한 검증은 규칙 기반 분류를 통해, 미세먼지의 농도의 정도를 범위별로 등급을 나누어 등급을 맞추었을 때를 기준으로 정확도를 테스트 해보았다. 규칙 기반 분류는 특정 규칙에 근거하여 데이터를 분류하여 특성을 기준으로 구분한다.[5] 이번 연구는 미세먼지를 기준으로 학습한 후, 예측하므로 미세먼지 수치에 대한 등급에 따라 분류하여 예측 값과 실제 값을 비교하였다.

Ⅲ. 결론

본 논문에서는 에어코리아의 과거의 대기 오염 정보 데이터를 이용한 학습으로 대기 오염 정보 데이터 항목 중 일부를 예측해 보았다. 대기 오염 정보 데이터 중 미세먼지의 수치만 예측해 보았지만, 대기 오염 정보의 초미세먼지($PM_{2.5}$), 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2) 등의 다른 항목에 대해서도 예측이 가능할 것이다. 이러한 예측을 통해 측정기기의 점검이나 교정 등의 자료 생산 중단으로 인한 결측 치 문제에 대한 해결책이 될 수 있다.

예측의 결과 값은 정확성 측면에서는 수치가 아닌 세계보건기구에서 제시한 미세먼지 농도 기준에 따라 분류를 통한 미세먼지 등급에 대한 정확도는 79%정도로, 정확한 수치 예측은 어렵지만 미세먼지 등급에 대한 예측은 정확한 편이고, 전체적인 증감의 시기와 비슷한 값까지의 예측이 가능하다. 하지만 세부적으로 보았을 때 이상치가 다수 발견되고 있다.

이후의 연구에서는 이러한 이상치를 줄여나가는 방향으로 추가적인 Ensemble과 내부, 외부 등의 구조형 클러스터링을 통한 다차원 임베딩을 사용하고, 학습에 쓰일 데이터를 확장하여 진행할 것이다. 예를 들어, 해당 데이터에 영향을 미치는 풍속이나 풍향, 습도와 온도 등의 기상 데이터를 추가하여 학습시키고, 이번 연구에 쓰인 2015년부터 2017년까지의 데이터보다 더 장기간의 데이터를 축적하고, 하루에 한 번에 해당하는 측정 데이터가 아닌 하루에 여러 번 측정한 대기오염 데이터를 사용하여, 깊은 학습을 통해 이전 모델보다 정확한 수치를 예측할 수 있도록 추가적인 연구를 진행할 것이다.

ACKNOWLEDGMENT

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2019007059, 시퀀스 데이터 분석 성능 향상을 위한 다차원 임베딩 기술 연구)

참 고 문 헌

[1] 이현우, 유지훈, 신동일, 신동규, “Apache Spark 기반의 미세먼지 농도 예측 및 대기오염 정보 제공 시스템” 한국통신학회 학술대회논문집, 2018

[2] Aurelien Geron, “Hands-On Machine Learning with Scikit-Learn & TensorFlow”, 한빛미디어 , 2018.

[3] Alexey Natekin, Alois Knoll, " Gradient boosting machines, a tutorial" Front Neurorobot., 2013

[4] Guolin ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017

[5] Charu C. Aggarwal, “Data Classification: Algorithms and Applications”, CRC Press, pp. 121-156, 2014