차량 주행 데이터 전처리에 관한 연구

방은진, 김창우, 최효섭, 강정훈
한국전자기술연구원

A Study on the Pre-Processing of non continuous Vehicle Driving Data

Bang Eun Jin, Kim Chang Woo, Choi Hyo Sub, Kang Jeong Hoon
Korean Electronics Technology Institute

요 약

본 논문에서는 실제 운행 차량으로부터 수집한 주행 데이터의 전처리 과정을 다루고 있다. 내연기관차가 전기차로 대체되고 있는 大전환 시대에 차량 데이터 수집 및 분석은 전기차 사업의 핵심이다. 전기차에서 주로 발생되고 있는 배터리 이슈뿐 아니라 신산업 창출을 통한 미래 주력 산업까지 해결할 수 있는 밑바탕이 되기 때문이다. 본 논문에서는 자동차 주행 데이터에서 서비스용으로 의미있는 부분을 분류하기 위해 데이터를 정리하고, 특징을 분석하기 위한 주행 데이터 전처리를 위한 효과적 소프트웨어 처리 단계를 제안한다.

Ⅰ. 서 론

본 논문에서는 차량 주행 데이터의 전처리 과정을 다루고 있으며 이때 사용된 차량 데이터는 차량에 부착된 IoT 기기를 이용하여 주행속도와 누적주행거리 위주로 수집하였다. 이러한 센서 데이터는 센서 및 장치의 오류 혹은 주행 중의 변수들로 인하여 이상치 및 결측치가 발생할 가능성이 많으며, 데이터 분석에 있어 왜곡된 해석을 야기할 수 있다. 따라서 이상치 및 결측치를 처리하는 '데이터 전처리' 과정이 데이터 분석 이전에 우선적으로 진행되어야 한다. 데이터 분포, 데이터 형식, 데이터 분석 목적에 따라 전처리 기법이 달라질 수 있으며 단일 데이터 전처리 기법이 아닌 여러 데이터 전처리 과정이 필요할 수도 있다. 본 논문에서는 이러한 데이터 전처리의 각 모듈을 도커 컨테이너로 이미지화하여 데이터 전처리 과정을 구성하였다. 컨테이너들을 통합 운영하기 위해 도커 컴포즈 기반 아키텍처를 적용하여 운영과 실행 효율을 향상 하였다

Ⅱ. 본론

2.1 장기간 데이터의 전처리

데이터 전처리 과정에는 크게 네 가지 기법을 들 수 있다. 1) 결측치를 채우거나 노이즈 데이터를 조정하는 Data Cleaning, 2) 중복된 데이터를 식별하는 Data Integration, 3) 관련 없는 데이터를 삭제하는 Data Reduction 4) 기존의 결과보다 나은 분석을 위해 데이터를 변환하는 Data Transformation 기법이다.

(1) Data Cleaning

데이터의 결측치 제거 및 노이즈 데이터를 조정하기 위하여 표준편차 기법, IQR, 가우시안 추정, Robust Covariance, SVM, Isolation Forest, Local Outlier Factor 기법을 사용하였다. 표준편차 기법은 데이터의 분포를 정규분포로 가정하고, ±(표준편차*시그마 계수)를 벗어나면 이상치(Outlier)라고 판단하는 기법이다. IQR은 정규분포에 벗어난 데이터의 사분위수 상위 75%지점 Q3와 하위 25%지점 Q1값의 차이를 말한다. 즉, 중앙에 위치하는 50% 데이터의 퍼진 정도를 나타내는 값이고 Q3에서 1.5배의 IQR 더한 값을 최대 상한선, Q1에서 1.5배의 IQR을 뺀 값을 최소 하한선으로 두고 이외의 범위를 이상치라고 판단한다. 가우시안 추정 기법은 데이터 샘플이 파라미터가 알려지지 않은 가우시안 분포에서 생성되었다고 가정하는 확률 모델로, 임계값을 정하여 밀도가 낮은 지역에 있는 모든 샘플을 이상치로 가정한다. Robust Covariance는 Determinant of sample covariance matrix를 최소로 만드는 h개의 데이터를 뽑아 그 데이터만 이용하여 분산이나 평균을 구사하는 기법이다. one-class SVM은 SVM이지만 클래스가 하나이고 원본 공간으로부터 다차원 공간에 있는 샘플을 분리하여 새로운 샘플이 이 영역 안에 놓이지 않는다면 이상치로 판단하는 기법이다. Isolation Forest는 무작위로 성장한 결정 트리로 구성된 랜덤 포레스트를 생성하여 모든 샘플이 다른 샘플과 격리될 때까지 진행하고 다른 샘플과 멀리 떨어져 있는 샘플을 이상치로 판단하는 기법이다. Local outlier Factor는 주어진 샘플 주위의 밀도와 이웃 주위의 밀도를 비교하여 이상치를 판단하는 기법이다. 데이터에서 이상치를 검색하여 제거한 후에, 서비스에 필요한 만큼 충분한 데이터량이 확보되었는지를 판단하는 것도 연속적으로 실행해야 한다.

(2) Data Reduction and summarization

관련성 없는 데이터를 삭제하기 위해서는 Dimensionality Reduction을 사용하였으며, 그 중에서도 PCA, LDA, LLE, Isomap 기법을 사용했다. PCA는 데이터에 가장 가까운 초평면을 정의한 다음 데이터를 그 초평면에 투영시키는 기법이다. LDA는 분류 알고리즘이나 훈련 과정에서 클래스 사이를 잘 구분하는 축을 학습하여 이 축을 데이터가 투영되는 초평면을 정의하는 데 사용하는 기법이다. LLE는 샘플을 고정하고 최적의 가중치를 찾아 투영시키는 기법이다. Isomap은 각 샘플을 가장 가까운 이웃과 연결하는 식으로 그래프를 생성하고 샘플간의 거리를 유지하면서 차원을 축소 시키는 기법이다. 데이터를 간략화 하고 저주파 대역의 특징을 확인하는 것은 대체적인 데이터 특성을 이해하기 위해 중요한 단계인데, 해당 단계에서 너무 많은 특징이 제거되지 않도록, 여러 가지 경우에 대해 반복적인 연산과 확인을 적용해야 한다.

(3) Data Transformation

데이터 분석 변환을 위해서 정규화, 표준화, 십진 스케일링 기법을 사용하였다. 정규화 기법은 데이터의 측정 단위가 데이터 분석에 영향을 줄 수 있으므로 데이터의 모든 속성에 동일한 가중치를 적용하는 기법이다. $\frac{\text{해당값} - \text{최소값}}{\text{최대값} - \text{최소값}}$ 으

로 계산한다. 표준화 기법은 각 해당 값이 평균을 기준으로 어느 정도 떨어져 있는지 나타내는 기법이다. 각 해당 값에서 평균을 뺀 다음 표준편차로 나누어 계산한다. 십진 스케일링 기법은 10배수의 값으로 이동시켜 정규화하는 기법이다. 각 해당 값을 10의 배수 값으로 나누어 계산한다.

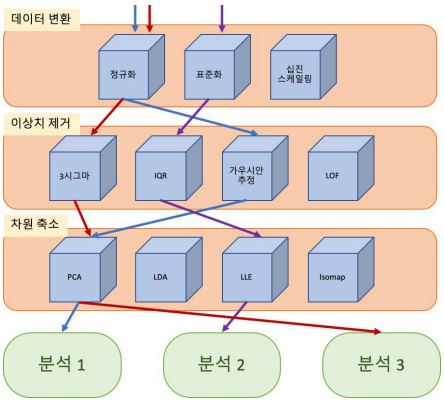

그림 1 데이터 전처리 기능 조합 구조

2.2 데이터 전처리 모듈 이미지화

데이터의 형식, 분포, 분석 목적에 따라 적용해야 할 전처리 과정이 달라진다. 본 논문은 본론에서 소개한 전처리 기법들을 모듈화하여 각 데이터 분석에 맞게 전처리 모듈을 조합하여 적용하고자 한다. 우선 각 데이터 전처리 모듈을 빠르게 배포하고 관리하기 위해 도커 컨테이너로 이미지화하고 각 컨테이너를 조합 운영하기 위한 도커 컴포즈로 실행한다. 도커는 애플리케이션을 신속하게 구축, 테스트 및 배포할 수 있는 소프트웨어 플랫폼이다. 이 컨테이너에는 라이브러리, 시스템 도구, 코드, 런타임 등 소프트웨어를 실행하는 데 필요한 모든 것이 포함 되어있다. 그러므로 도커를 사용하면 환경에 구애받지 않고 애플리케이션을 신속하게 배포 및 확장 할 수 있으며 재사용이 많은 코드를 이미지화 하여 효율성을 높일 수 있다. 이러한 도커를 기반으로 전체적인 아키텍처는 그림 1과 같다. 우선 각 모듈에 필요한 인자는 환경 변수로 지정을 하여 필요한 인자 값을 변경하고 독립적으로 실행 될 수 있도록 지정 하였고 구성 된 컨테이너들은 하나의 네트워크로 묶여 서로 통신이 가능하며 실행 순서를 제어할 수 있다. 이러한 구조를 기반으로 데이터 특성에 따라 필요한 데이터 전처리 모듈을 선택적으로 조합하여 하나의 Data Flow 구성이 가능하다.

2.3 실제 주행 데이터 전처리 테스트

자동차 주행속도 및 누적주행거리 수집 데이터에서 발견된 문제점은 존재하는 원본 데이터가 전체 존재해야하는 시간 구간에서 실제로는 50% 이하로 존재하여 결측치가 상당히 높다는 점이었다. 따라서 전체 원본 데이터 구간을 작게 분류하여 데이터 존재비율이 50%이상인 경우와 이하인 경우로 나누어 특성에 맞게 전처리를 진행하였다.

(1) 원본 데이터 존재비율이 50% 이하인 시간 구간의 전처리
그림2와 같이 구간 데이터 존재 비율이 50%이하인 차량들에 대해서 주차라고 생각되는 구간의 time stamp 값을 1로 지정하여 저장하였다. 빈 구간의 바로 전 데이터와 바로 후의 속도가 0이 아니거나, 두 지점의 누적 주행거리가 차이가 1km이상이어서 주차인지 아닌지 확인이 불가한 구간은 채울 수 없는 구간(Unknown section)이라 정의하였다. 이때 '주차'의 조건범위는 2시간 이상 데이터가 없고 빈 구간 바로 전 time stamp의 속도가 0, 빈 구간 속도가 0 또는 비존재이면서 누적주행거리 차이가 1km미만인 상태이다. 이러한 정의에 따라 데이터를 분류하여 전처리한 결과로 결측치를 80% 복구할수 있었다. 이외 20% 부분에서는 중간값을 기준으로 네가지 경우에 대해, 조건별로 반복하여 해당 구간 복구를 수행하였다.

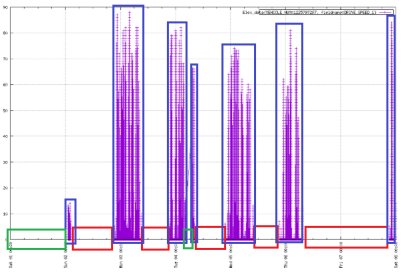

그림 2 원본데이터 비존재구간의 보간

(2) 원본 데이터 존재비율이 50% 이상인 경우 전처리
원본 데이터 존재비율이 50% 이상인 경우, 각 조건에 따라 추출된 데이터를 각기 다른 이름의 메트릭으로 저장하으며 2시간 이상 속도가 0으로 지속되는 구간과 빈 구간의 바로 전 데이터와 바로 후의 속도가 0이 아니거나 두 지점의 누적주행 거리가 1km 이상인 구간, 주행이라고 생각하는 구간으로 나누어 전처리 하여, 95%의 데이터를 처리할 수 있었다.

Ⅲ. 결론

본 논문에서는 이러한 전처리 기법을 통하여 속도기반 정차구간 추출 및 GPS 기반 정차구간 추출이 가능하였고 정차구간의 조건에 따른 통계적 기법으로 데이터 비존재 구간에서의 결측치를 파악하여, 주차상태 데이터를 복구할 수 있었다. 이러한 전처리 기법을 이용하여 주행 차량 데이터를 분석한다면 정확한 분석 조건을 검색할 수 있기 때문에, 전기차의 부품, 시스템 성능 향상 뿐 아니라 소프트웨어 서비스 기능 개발에도 활용될 수 있다.

ACKNOWLEDGMENT

This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE), Korea, under the "Infrastructure Support Program for Industry Innovation"(reference number P0014715, Construction and verification for electric vehicle parts data platform) supervised by the Korea Institute for Advancement of Technology (KIAT).

참 고 문 헌

[1] JunHyeog Cho, Sunghae Ju "Big Data Smoothing and Outlier Removal for Patent Big Data Analysis" Journal of The Korea Society of Computer and InformationVol. 21 No. 8, pp. 77-84, August 2016.

[2] 이미영, "클라우드 기반 대규모 데이터 처리 및 관리 기술," 전자통신동향분석, 2009.