기계학습을 이용한 교통 정체 구간 예측 시스템 설계

전우혁, 최지인, 박경빈, 김경섭
충남대학교 컴퓨터공학과
e-mail: tororum4992@gmail.com, acji4878@naver.com, hjk1210p@nate.com, sclkim@cnu.ac.kr

Design of traffic congestion predictive system with Machine Learning

Woohyeok Jeon, Jiin Choi, Kyungbin Park, Kyungsup Kim
Dept of Computer Engineering, Chung-Nam National University

요 약

정보통신기술이 발전함에 따라 수많은 데이터가 발생하고 있다. 이러한 '빅데이터'의 활용은 국민의 니즈 파악, 공공서비스 제공 등 미래 경쟁력의 핵심 가치라 할 수 있다. 이에 본 논문에서는 기상데이터와 교통데이터를 수집한 후, 분산 시스템 환경 하에서 실행되는 기계학습 알고리즘을 이용하여 기상기후와 관련된 교통 정체 구간 예측 시스템에 대해 제안하고자 한다.

1. 서론

최근 정보통신기술의 발달로 등장한 '빅데이터'는 미래 경쟁력의 결정하는 핵심 이슈라고 할 수 있다. '빅데이터'란 기존의 소프트웨어로는 저장, 관리, 분석이 어려운 데이터라고 정의할 수 있다. '빅데이터'의 활용을 통해 사회 현안 및 국민의 니즈 파악, 미래전략 수립, 선제적 공공서비스 제공 등 국가 경쟁력 강화 및 정부 혁신을 지원할 것으로 예상된다. 1)

기상기후 빅데이터는 기상자료의 분석, 미래 예측과 에너지 관광, 날씨보험과 같은 타산업과의 융합을 통해 신사업 창출이 가능하다. 그리고 기상기후 빅데이터는 개인의 신상을 바탕으로 한 정보가 아니기 때문에 개인 정보 보호의 문제에서 자유로울 수 있다. 이러한 기상 빅 데이터와 사회경제적 자료를 접목하면 시너지 효과를 발휘할 수 있다. 예를 들면 기상 변화에 따른 실시간 교통정보 연계, 기상 현상과 물류의 연계를 통해서 새로운 가치창출이 가능하고 기상재해로 인한 도로 사고를 예방할 수 있다. 2)

이에 따라 본 논문에서는 기상 데이터와 교통 데이터를 수집한 후, 분산 시스템 환경 하에서 실행되는 기계학습 알고리즘을 이용하여 기상기후와 관련된 교통 정체 구간 예측 시스템에 대해 제안하고자 한다.

2장에서는 시스템 구조에 대한 설명과 데이터의 수집과 전처리과정을 대한 내용이 있다. 3장에서는 불안정성을 줄인 데이터의 예측 모델 생성과 정확도를 보여준다. 4장에는 실험결과와 결론이 있다.1)

2. 시스템 설계

2.1 시스템 개요

교통 정체는 사고나 공사, 날씨 등 많은 요인들의 영향을 받기 마련이다. 더 나아가 사고는 운전자의 부주의로 인한 것일지라도 날씨의 영향을 종종 받는 경우도 있다. 날씨가 교통에 많은 영향을 미치는 데 얼마나 미치는지 효과적으로 알 수 없었다. 이를 해결하기 위해 본 시스템에서는 대량의 날씨 데이터, 교통 데이터를 분석하여, 날씨에 따른 주요 도로의 교통 흐름이 어떻게 될 것인지 예측하는 시스템을 제안하고자 한다.

2.2 시스템 구성

본 논문에서 제안하는 시스템은 대량의 날씨 데이터와 교통 소통 데이터를 분석하고 예측하기 위해서 통계 분석 툴과 분산 처리 시스템을 구성하였다. 실제로 구축한 시스템은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있는 R과 분산 처리 시스템의 대표적인 솔루션 Hadoop 을 사용하였다. R의 기계학습 라이브러리와 Hadoop 위에 Spark를 사용하여 대량의 데이터를 효과적으로 분석할 수 있게 하였다. 예측 시스템에서 분석한 데이터를 바탕으로 기계학습 알고리즘을 이용하여 수행하였다. 분산 처리 시스템 상에서 대용량의 데이터를 안정적으로 저장하기 위해 HDFS를 사용하였으며, YARN 어플리케이션을 통해 분산 시스템 상태를 관리하였다.

※ 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(R7115-16-1007)

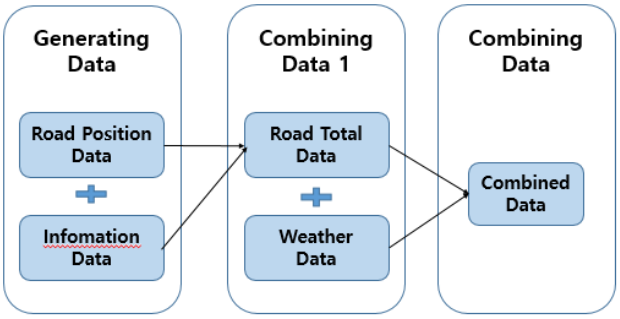
2.3 데이터 수집

본 논문에서 제안하는 '교통 정체 구간 예측 시스템'을 설계하기 위해서는 데이터 구축을 위한 데이터들과 그 데이터들을 효율적으로 정리, 축소하기 위해서 전처리가 필요하다. 각 지역별로 데이터를 전처리를 하기 위해서 날씨 데이터를 여러 대도시(고속도로) 별로 나누고, 정체구간의 시간과 날씨 데이터에서의 시간을 하루치로 통합한다.

날씨 데이터의 경우에는 하루당 강수량, 일조량, 온도 등 데이터가 포맷 되어 있지만 교통 데이터의 경우에는 5분당 교통 정보를 담고 있으므로 이를 전처리 시켜주어야 한다.

이때 데이터를 전처리하기 위해서 대용량의 데이터를 다루기에 적합한 R을 이용한다.

전처리 방식은 아래 그림과 같으며 교통 정체 데이터와 각 지역별 Road ID 값 리스트들을 먼저 통합 처리해주고 통합해준 데이터와 날씨 데이터를 다시 날짜별로 통합시켜준 데이터를 예측 분석할 시에 사용한다.



<그림 1> 교통 정체 구간 예측 시스템 데이터 전처리 과정

1일 치 전체 소통 데이터를 각 지역별로 나누어서 쪼개어 주는 역할을 해야 한다. 이때 도로 장소별 RoadId 리스트가 들어 있는 Road Position Data를 사용하여 각 지역별 도로로 우선 나누어 주는 작업을 한다. 그리고 각 도로 당 최대 지연시간, 통행시간, 최소 소통 속도를 각 도로마다 통합시켜준다.

해당 지역의 1일 치 날씨 데이터와 2530여 개의 도로별 소통 데이터를 결합 위해서는 우선 기준 값을 맞춰야 하는데 이때 전처리한 소통 데이터는 날짜 값이 없으므로 해당 날짜 값을 추가시켜주어야 한다.

그리고 해당 날짜 값을 기준으로 날씨 데이터와 전처리한 소통 데이터를 통합시켜준다. 이때 날씨 데이터에서 전처리한 소통 데이터에 추가시켜주는 것은 온도, 습도, 강수량, 강우량, 풍량 등이 있다.

2.4 시스템 설계

우리는 날씨 데이터와 교통소통 데이터를 이용하여 소통 속도와 시간 등을 기준으로 Decision Tree(결정 트리) 모델을 통한 날씨에 따른 소통 예상시간이나 속도 등을 예측하는 시스템을 설계한다.

결정 트리 학습은 널리 사용되며 귀납적 추론에 매우 실용적인 방법 중 하나이다. 3)

이 Decision Tree(결정 트리)는 다른 알고리즘에 비해 직관적인 이해, 설명이 용이한 장점을 지니고 있기 때문에 이 예측 기법을 사용했으며, 이 예측 기법을 사용하기 위해서 R에서 제공하는 C5.0 라이브러리 속에 Decision Tree를 사용하였다. 또한 Bootstrapping으로 Decision Tree 모델의 불안정성을 줄이고 정확도를 늘리도록 하였다.

3. 구현 및 평가 (제안한 알고리즘)

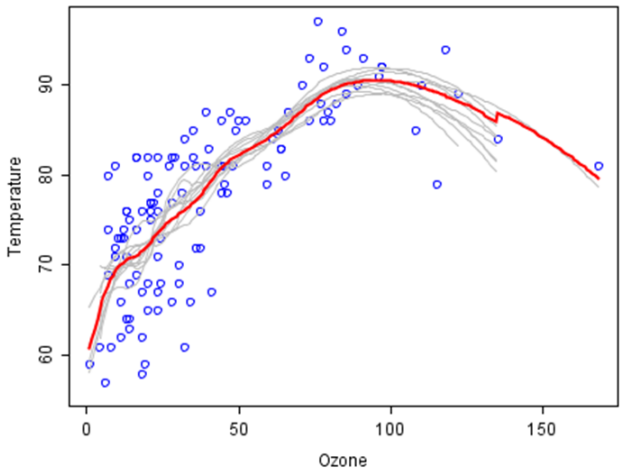
3.1 예측 시스템 위한 기본 데이터 학습

기계학습을 위해서는 분석 시스템에서 분석되어서 나온 데이터를 사용하는데 이때 기계학습을 위한 데이터인 Training Data와 트레이닝 된 Decision Tree에 결과를 확인하기 위한 데이터인 Testing Data가 필요하다. 우선적으로 예측 시스템 설계를 위해서 60%의 데이터를 트레이닝 시키고 40%를 Testing 하여서 예측 모델의 정확도 및 그려지는 Decision Tree를 확인해 보았다.

3.2 Training Data Bootstrapping

위에서 설명한 대로 이 방법은 가설 검증을 하거나 메트릭을 계산하기 전의 Random Sampling을 적용하는 방법을 일컫는다. 이때 중복을 허용해 주어야 한다.

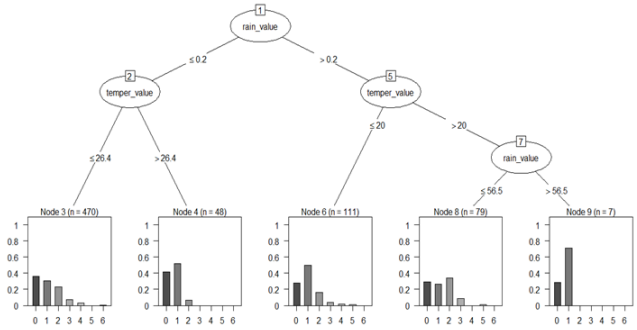
이것을 해주는 이유는 Training 시에 어떠한 상황 쪽에 너무 많은 데이터가 Training 되어있으면 올바른 Decision Tree 모델이 형성되지 않아서 올바른 예측 값이 나오지 않기 때문이다. Random Sampling을 통하여 데이터를 뽑아내고 그것을 100번 반복하여 평균을 취한 데이터를 사용한다.



<그림 2> 기온과 오존층의 관계를 Bootstrapping 하여 나온 결과

예를 들어 회색선이 Random Sampling 된 데이터이고 빨간 선이 Bootstrapping이 적용된 데이터이다.

3.3 데이터를 통한 예측 및 정확도



<그림 3> Decision Tree Model

날씨 데이터 중에 기온, 눈, 비, 바람 등을 부가 요소로 하고 시간을 7분위로 나누어서
Decision Tree를 모델링 한 예측 모델이다. 각 모델별로 기준 값을 가져오기는 힘들지만 predict() 함수를 사용하여 Testing Data를 넣어 나오는 결과 테이블은 가능하다. 그리고 예측 모델의 정확도와 테이블 형성을 나타낸 그림은 아래와 같다.

```
-----  Trial 0:  -----

Decision tree:

rain_value <= 0.2:
:...temper_value <= 26.4: 0 (470/299)
:   temper_value > 26.4: 1 (48/23)
rain_value > 0.2:
:...temper_value <= 20: 1 (111/56)
    temper_value > 20:
    :...rain_value <= 56.5: 2 (79/52)
        rain_value > 56.5: 1 (7/2)

*** boosting reduced to 1 trial since last classifier is very inaccurate

*** boosting abandoned (too few classifiers)

Evaluation on training data (715 cases):

          Decision Tree
        ----------------
        Size      Errors

           5  432(60.4%)   <<


         (a)   (b)   (c)   (d)   (e)   (f)   (g)    <-classified as
        ----  ----  ----  ----  ----  ----  ----
         171    53    23                           (a): class 0
         144    85    21                           (b): class 1
         107    21    27                           (c): class 2
          34     4     7                           (d): class 3
          13     2                                 (e): class 4
                 1     1                           (f): class 5
           1                                       (g): class 6


       Attribute usage:

       100.00% rain_value
       100.00% temper_value


Time: 0.0 secs
```

<그림 4> 예측 모델의 정확성과 classified 테이블

실험 결과 총 5개의 리프 노드가 생성되었으며 각 classified 테이블도 확인할 수 있었다. 정확도는 39.6%로 낮은 정확도를 보여준 것으로 파악되었다.

4. 결론 (실험결과 및 요약)

본 논문에서는 Decision Tree 모델링 예측 결과를 통해서 수집한 날씨 데이터와 교통 데이터가 어떤 상관관계가 있는지 확인할 수 있다.

예측 시스템에서 낮은 정확도는 더 많은 데이터를 활용하여 학습을 시킨다면 개선될 것 같다. 본 논문에서는 Decision Tree를 모델링했지만 개선점으로 Regression 및 다른 기계학습 알고리즘을 활용한다면 더 높은 정확도와 흥미로운 결과를 볼 수 있을 것이다.

과거의 날씨 데이터와 교통 데이터를 분석하여 예측하는 시스템에 대해 제안하였다. 제안한 시스템을 통해 사용자는 해당 도로의 날씨에 따른 예측 소통 정보를 활용할 수 있다.

이러한 시스템은 도로 교통 정보를 예상하고 제공하는 새로운 방법의 일환으로 사용자들이 예상되는 도로 교통 정보를 통해 최적의 길을 안내할 수 있을 것으로 판단된다.

참고문헌

1) 정우수, 『경쟁전략 모형을 이용한 빅데이터 산업구조 분석』, 『2013 정보통신설비 학술대회 논문집』, 2013년, pp.104-107

2) 한국기상산업진흥원『빅데이터 관점에서의 기상정보와 타산업간의 융합방안』, 2013년 12월, pp.3, 14

3) 손지은, 김성범 『의사결정나무 모델에서의 중요 룰 선택기법』, 『2013 대한산업공학회 추계학술대회 논문집』, 2013년, pp.13-23.