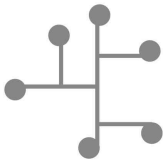


결측값을 포함한 센서 스트림에 대한 어텐션 메커니즘 및 합성곱 신경망 기반의 패턴 분류 기법

Pattern Classification based on Attention Mechanism and CNN for Sensor Stream Data including Missing Values

이은진(Eunjin Lee)[1] 오소연(Soyeon Oh)[2] 이민수(Minsoo Lee)[3]

요 약

다양한 센서로부터 수집된 IoT 스트림 데이터 분석은 대표적인 비선형 분석 문제로, 최근 이러한 문제들의 해결에 합성곱 신경망(Convolutional Neural Network, CNN)을 비롯한 딥러닝 기법들을 다방면으로 적용하고 있다. 또한, IoT 센서 스트림 데이터는 그 수집 과정에서, 센서와 서버 간의 통신 장애 또는 센서의 하드웨어적 결함 등으로 인한 결측값 즉, 손실 데이터를 포함하는 경우가 많으며, 이러한 손실 데이터는 분석의 정확도를 감소시킨다. 한편, 다양한 센서 스트림 데이터 중, 루프 센서를 통해 수집된 교통량 데이터 분석은 도시 계획, 교통 공학, 다양한 교통 및 위치 기반 서비스의 구현 등에 활용된다. 그러나 루프 센서를 통한 교통량 데이터 수집 과정에서 결측값이 발생하는 경우가 많다. 본 논문에서는 이렇게 결측값이 포함된 센서 스트림 데이터의 패턴 분류 정확도를 높이기 위한 기법을 제안한다. 제안하는 기법은 합성곱 신경망 기반의 패턴 분류 모델에 어텐션 메커니즘(Attention Mechanism)을 도입하여 비손실 데이터에 대한 가중치를 부여함으로써 결측값으로 인한 정확도의 손실을 보완한다. 본 논문에서는 결측값의 발생이 잦은 루프 센서 기반의 교통량 데이터를 대상으로 제안하는 패턴 분류 기법을 적용하였고, 제안하는 기법이 결측값을 포함한 센서 스트림 데이터에 대한 패턴 분류 정확도를 향상시킬 수 있음을 실험을 통해 확인하였다.

주제어: IoT, 스트림 데이터, 딥러닝, 합성곱 신경망, 어텐션 메커니즘

---

1 이화여자대학교 컴퓨터공학과, 학사과정.
2 이화여자대학교 컴퓨터공학과, 박사과정.
3 이화여자대학교 컴퓨터공학과, 교수, 교신저자.
+ 이 연구는 2019학년도 이화여자대학교 교내연구비 지원에 의한 연구임.
+ 논문접수: 2020년 07월 26일, 최종 심사완료: 2020년 08월 21일, 게재승인: 2020년 08월 25일.

Abstract

Analysis for IoT stream data collected from various sensors is a typical non-linear analysis problem, and recently, deep learning techniques including convolutional neural networks have been applied to these problems in various ways. In addition, the IoT sensor stream data often includes missing data, that is, loss data due to a communication failure between the sensor and the server or a hardware defect of the sensor during the collection process, and such loss data reduces the accuracy of analysis. Meanwhile, among the various sensor stream data, the analysis of traffic volume data collected through the loop coil sensor is used for urban planning, traffic engineering, and implementation of various traffic and location-based services. However, during the process of collecting traffic data through the loop coil sensor, missing values are often generated. In this paper, we propose a method to increase the accuracy of pattern classification of sensor stream data containing missing values. The proposed method compensates for the loss of accuracy due to missing values assigning weights to non-loss data by applying attention mechanism to the pattern classification model based on the convolutional neural network. In this paper, the proposed pattern classification method is applied to traffic volume data measured by loop coil sensors that frequently generate missing values, and it was confirmed through experiments that the proposed method can improve the accuracy of pattern classification for sensor stream data including missing values.

Keywords: IoT, Stream Data, Deep Learning, CNN, Attention Mechanism

1. 서 론

다양한 센서로부터 수집된 IoT 스트림 데이터 분석 문제는 대표적인 비선형 문제로, 최근 이러한 비선형 문제들을 해결하기 위하여 합성곱 신경망(Convolutional Neural Network, CNN)을 비롯한 여러 딥러닝 모델 및 기법들을 다방면으로 적용하고 있다[1, 2]. 한편, IoT 센서 스트림 데이터를 수집하는 과정에서는 센서와 서버 간의 통신 장애 또는 센서의 하드웨어적 결함 등으로 인한 결측값 즉, 손실 데이터가 발생하는 경우가 많다. 이러한 손실 데이터는 합성곱 신경망 기반으로 수행되는 일반적인 센서 데이터 패턴 분류 작업의 정확도를 감소시킨다. 따라서, IoT 센서 스트림 데이터에 대한 패턴 분류 기법을 실세계 환경에 적용하기 위해서는 결측값이 포함된 센서 스트림 데이터에 대한 패턴 분류 정확도를 높이기 위한 기법의 설계 및 도입이 필수적이다.

다양한 센서 스트림 데이터 중, 루프 센서[3]를 통해 수집되는 교통량 데이터 역시, 위에서 언급한 다른 IoT 센서 스트림 데이터 수집 과정에서와 유사한 여러 가지 요인에 의하여 결측값이 빈번하게 발생한다. 교통량 데이터를 활용한 각종 분석 결과는 도시 계획, 교통 공학, 다양한 교통 및 위치 기반 서비스의 구현 등에 다양하게 활용되는데[4, 5], 수집 과정에서 발생한 결측값들로 인하여 교통량 데이터에 대한 각종 분석 작업의 정확도가 떨어지게 된다.

본 논문에서는 이렇게 결측값이 포함된 센서 스트림 데이터의 패턴 분류 정확도를 높이기 위한 기법을 제안한다. 제안하는 기법은 합성곱 신경망 기반의 센서 스트림 데이터 패턴 분류 모델을 기반으로 하며, 여기에 어텐션 메커니즘(Attention Mechanism)을 도입하여 비손실 데이터 부분에 대한 가중치를 부여함으로써 결측값으로 인한 정확도의 손실을 보완한다. 본 논문에서는 결측값의 발생이 잦은 루프 센서 기반의 교통량 데이터를 대상으로, 제안하는 패턴 분류 기법을 적용하였고, 제안하는 기법이 결측값을 포함한 센서 스트림 데이터에 대한 패턴 분류 정확도를 향상시킬 수 있음을 실험을 통해 확인하였다. 특히, 결측값에 대한 특별한 전처리 과정을 거치지 않더라도 본 논문에서 제안하는 어텐션 메커니즘 기반의 패턴 분류 기법을 통하여 분류 정확도의 향상이 가능하다는 것을 실세계 교통량 데이터에 대한 실험을 통해 확인하였다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 본 논문에서 제안하는 기법에서 활용한 어텐션 메커니즘을 비롯한 딥러닝 기반의 시계열 데이터 분석 관련 연구 동향을 소개하고, 수집 과정에서 결측값의 생성이 잦은 센서 스트림 데이터인 교통량 데이터의 특성을 소개하면서, 본 논문에서 제안하는 기법에 대한 실험에 교통량 데이터를 활용하게 된 배경을 소개한다. 3장에서는 본 논문에서 제안하는 어텐션 메커니즘 및 합성곱 신경망 기반의 패턴 분류 모델을 소개한다. 4장에서는 이와 관련한 본 논문의 실험에 사용된 교통량 데이터와 실험 조건을 명시한 뒤, 본 논문에서 제안하는 기법의 효과와 성능을 실험을 통하여 분석하면서, 5장의 결론으로 마무리 짓는다.

2. 관련 연구

2.1 딥러닝 기반의 시계열 데이터 분석

최근 딥러닝 기반의 예측 및 분류 기법들이 센서 스트림과 같은 시계열 데이터의 분석에 활발히 활용되고 있다[6-8]. 앞먹임 신경망(Feed-forward Network)과 합성곱 신경망은 주어진 시계열 데이터 패턴을 특정 클래스로 분류하는 문제에 주로 활용되고[9-11], 순환 신경망 (Recurrent Neural Networks, RNN)은 순차적인 데이터에 대한 회귀적인 특징을 학습하여 시계열 데이터의 미래 값을 예측하는 문제에 주로 활용된다[12-14]. 기존 통계적

관점에서의 가우시안 기반 선형 모델을 활용하는 분석 기법들과 비교하여, 이러한 딥러닝 기반 분석 기법들을 통하여 각종 비선형 분석 문제들을 해결할 수 있게 되었기 때문에, 센서 스트림 데이터와 같은 실세계의 시계열 데이터에 대한 보다 향상된 정확도의 분석이 가능해졌다.

본 논문에서는 어텐션 메커니즘과 합성곱 신경망 기반의 교통량 데이터 패턴 분류 모델을 설계하였다. 합성곱 신경망은 커널을 통해 주변 데이터들과의 연관성을 학습할 수 있기 때문에, 이미지 데이터뿐만 아니라 여러 채널로부터 수집되는 다변량 시계열 데이터 처리에 효과적임이 여러 연구들에 의해 증명되었다[9-11]. 이 때, 모델의 학습 과정에서 조건부 드롭아웃(conditional dropout)을 활용하면 모델의 성능 저하를 최소화하면서 학습 속도를 높일 수 있다[15].

한편, 어텐션 메커니즘은 주로 컴퓨터 비전에서의 시각적 집중의 구현이나 순차적인 입력 데이터에 대한 회귀 분석 문제에서 입력 데이터에 가중치를 부여하기 위하여 활용된다. 어텐션 메커니즘은 특징 추출과 관련하여 특별히 높은 중요도로 고려해야 하는 부분에는 상대적으로 높은 가중치를, 그 외의 부분에는 상대적으로 낮은 가중치를 갖도록 표현한 벡터인 어텐션 스코어 (attention score)를 산출할 수 있도록 학습된다. 어텐션 메커니즘은 높은 가중치를 둘 위치를 반영하는 메타 데이터의 도입, 순환 신경망의 은닉 상태 (hidden state) 벡터간의 유사도 계산 등 입력 데이터와 목표 출력에 적합한 다양한 방식을 통해서 어텐션 스코어를 학습하도록 설계된다. 이렇게 학습된 어텐션 스코어와 입력 데이터를 곱하면 기존 입력 데이터의 특정 부분에 높은 가중치를 부여한 새로운 입력 데이터를 얻을 수 있다. 이러한 과정을 통해서 어텐션 메커니즘은 특징 학습을 용이하게 할 수 있게 되며, 이미지 처리 성능 향상을 위한 입력 이미지에 대한 시각적 집중의 구현이나 자연어 처리 또는 회귀 분석을 위한 순환 신경망 기반 모델의 성능을 향상 시키는 데에 큰 도움을 줄 수 있음이 알려져 있다[16-18].

2.2 교통량 데이터

교통량 데이터는 도시 계획, 교통 공학, 다양한 교통 및 위치 기반 서비스의 구현 등에 다양하게 활용되는 실세계 데이터로[4, 5], 센서를 통해 수집되는 시계열 데이터의 한 종류이다. 교통량 데이터의 수집은 주로, 루프 센서[3]를 통해 이루어지는데, 이 과정에서 센서와 서버 간의 통신 장애 또는 센서의 하드웨어적 결함 등으로 인한 결측값이 빈번하게 발생하게 된다. 본 논문에서 실험에 활용한 데이터는 캘리포니아 주의 교통량 측정 시스템인 Caltrans PeMS[19]에 의해 수집된 실제 교통량 데이터인데, Caltrans PeMS의 경우에도 루프 센서의 하드웨어적 결함, 여러 루프 센서를 통합해 관리하는 스테이션(station)에서의 결함, 센서 스트림 데이터 전송을 담당하는 네트워크의 결함 등의 여러 가지 요인에 의하여 데이터 수집 과정 중 결측값이 발생하게 된다. 그리고 실제로 Caltrans PeMS에 의해 수집된 교통량 데이터들은 이러한 결측값들이 포함된 손실 데이터의 형태를 보이고 있으며, 결측값 발생 원인이 다양하기 때문에 데이터의 손실 패턴도 무작위하다.

3. 결측값을 포함한 센서 스트림에 대한 대한 어텐션 메커니즘 및 합성곱 신경망 기반의 패턴 분류 기법

이 장에서는 본 논문에서 제안하는 어텐션 메커니즘 및 합성곱 신경망 기반의 패턴 분류 기법을 소개한다. 먼저, 제안하는 기법의 개요를 기술하고, 손실

데이터에 의한 정확도 손실을 보완하기 위한 어텐션 메커니즘의 설계와 패턴 분류를 위한 합성곱 신경망 기반의 모델의 설계를 기술한다. 마지막으로, 제안하는 기법에 앞서 적용할 수 있는 결측값에 대한 전처리 기법에 대하여 기술한다.

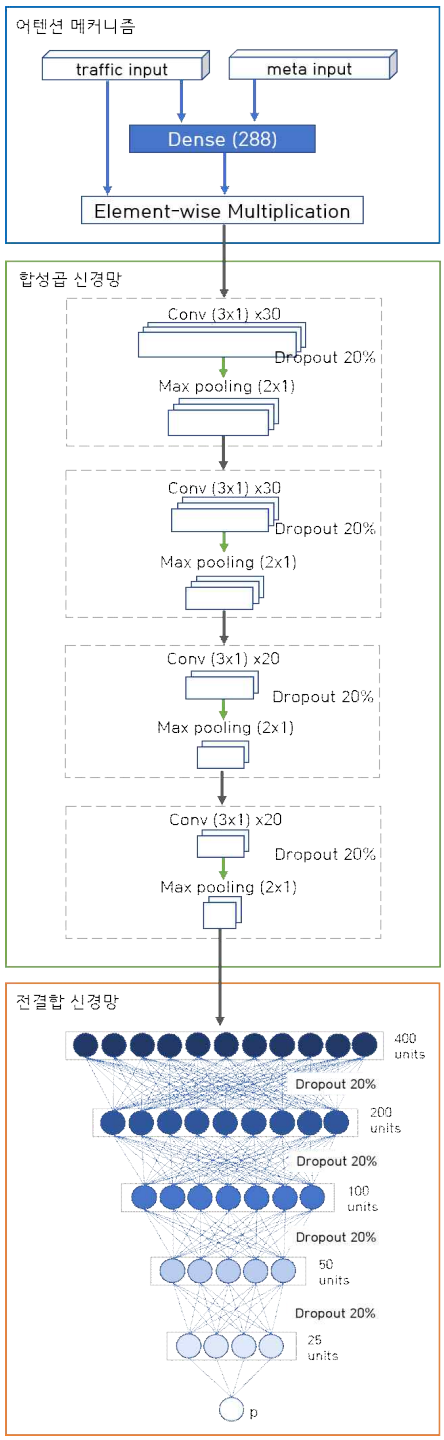
3.1 제안하는 기법

제안하는 기법은 [그림 1]과 같이, 결측값에 의한 정확도 손실을 보완하기 위한 어텐션 메커니즘 부분과 패턴 분류를 수행하는 합성곱 신경망 기반의 모델을 결합한 딥러닝 기반의 분석 기법이다. 어텐션 메커니즘은 입력 데이터로부터 생성된 메타 데이터에 기반하여, 입력 데이터의 비손실 부분에 대한 가중치를 모델링 한다. 이러한 어텐션 메커니즘에 의하여 생성된, 입력 데이터의 비손실 부분에 가중치가 부여된 벡터는 합성곱 신경망과 전결합 신경망으로 구성된 딥러닝 모델의 입력으로 주어지고, 최종적인 패턴 분류 결과를 얻게 된다.

3.1.1 어텐션 메커니즘 설계

제안하는 패턴 분류 기법의 어텐션 메커니즘은 입력 데이터에서 중요한 부분 즉, 결측값이 포함되지 않은 비손실 부분에 높은 가중치를 부여하여, 합성곱 신경망과 전결합 신경망으로 구성된 딥러닝 모델이 입력 데이터의 비손실 부분에 더욱 집중하여 패턴 분류를 수행할 수 있게 한다. 따라서 결측값에 의한 패턴 분류 정확도의 손실을 보완하여 전체적인 분류 성능을 높일 수 있게 된다.

이러한 어텐션 메커니즘의 구조는 [그림 1]의 최상단과 같이 도식화 할 수 있다. 어텐션 메커니즘의 구현을 위해서는 입력 데이터와 메타 데이터를 필요로 한다. 메타 데이터는 입력 데이터로부터 생성되는 데이터로, 입력 데이터의 결측값에 대한 정보를 반영하는 두 개 부분 벡터의 연결(concatenation)로 구성



[그림 1] 제안하는 패턴 분류 기법

된다. 첫 번째 부분 벡터는 입력 데이터와 같은 차원을 가지며, 입력 데이터의 결측값 위치를 표현한다. 두 번째 부분 벡터는 입력 데이터의 결측값 포함 여부를 표현하는 (2, 1) 차원의 벡터이다. 이 두 벡터를 연결하여 메타 데이터를 구성한다.

위와 같은 과정을 거쳐 얻어진 메타 데이터와 기존 입력 데이터를 입력으로 받아 시그모이드(sigmoid) 활성화 함수를 갖는 유닛으로 구성된 한 층의 신경망을 거치면 입력 데이터와 같은 차원의 출력 벡터를 얻을 수 있다. 이러한 출력 벡터는 입력 데이터에 대한 결측값과 비손실 부분에 대한 가중치를 표현하는 가중치 벡터이다. 이러한 가중치 벡터와 기존 입력 벡터를 요소별 곱(element-wise multiplication)하여 어텐션 메커니즘의 출력 벡터이자, 이후 합성곱 신경망과 전결합 신경망으로 구성된 패턴 분류 모델의 입력이 되는 벡터를 얻는다.

3.1.2 합성곱 신경망 기반의 모델 설계

어텐션 메커니즘의 출력 벡터는 합성곱 신경망과 전결합 신경망으로 구성된 딥러닝 모델의 입력 벡터로 전달되며, 이 모델의 구조는 [그림 1]의 중하단과 같다. 이 모델은 24시간 교통량 데이터 패턴에 따라, 이 날이 다저스 구단의 경기일인지 아닌지를 구분하는 이진 분류를 수행한다. 따라서 전결합 신경망의 출력층에서는 시그모이드 함수를 활성화 함수로 사용하였으며, 그 외 나머지 은닉층 및 합성곱 층에서는 ReLu를 활성화 함수로 사용하였다.

본 논문에서는 실험을 통하여 패턴 분류를 수행하기 위한 합성곱 신경망과 전결합 신경망의 구조와 관련된 파라미터 값들의 최적화를 수행하였다. 파라미터 최적화 실험 결과, 합성곱 신경망의 레이어(layer) 수는 7개, 유닛 수는 10개일 때 가장 좋은 성능을 보였으며, 레이어 수가 10개 이상이거나, 유닛 수가 20개 이상일 경우에는 성능이 감소했다. 필터 크기는 2일 때 가장 좋은 성능을 보였으며, 필터 크기가 3보다 클 경우에도 성능이 감소했다. 풀링 크기는 3일 때 가장 좋은 성능을 보였으며, 풀링 크기가 1이거나 3보다 클 경우에는 정확도가 많이 낮아졌다. 스트라이드 크기가 1보다 클 경우 역시 정확도가 많이 낮아졌으며, 평균 풀링 보다는 최대 풀링을 수행할 경우의 성능이 더 높았다. 이러한 최적화 과정을 거쳐 결정된 합성곱 신경망의 파라미터 최적값은 [표 1]과 같다.

합성곱 신경망 구조 관련 파라미터	최적값
레이어 수	7
유닛 수	20
필터 크기	2
스트라이드 크기	1
풀링 크기	3
풀링 방식	최대 풀링

[표 1] 합성곱 신경망 구조 관련 최적 파라미터

3.2 결측값에 대한 전처리

본 논문에서는 결측값에 의한 정확도 손실을 보완하기 위하여 어텐션 메커니즘을 도입한 패턴 분류 기법을 설계 및 제안하였다. 한편, 일반적으로 결측값이 포함된 손실 데이터를 분석할 때에는 주로 결측값에 대한 전처리를 수행한 뒤에 목표로 하는 분석을 수행한다. 결측값에 대한 전처리는 대표적으로 두 가지 방식이 있는데, 첫 번째는 결측값을 특정한 동일값 또는 데이터 특성을 반영하는 대푯값으로 대체(imputation)하는 방식이며, 두 번째는 결측값 주변의 실측값들을 이용하여 결측값을 예측하여 보간(interpolation)하는 방식이다.

본 논문에서 제안하는 어텐션 메커니즘 및 합성곱 신경망 기반의 패턴 분류 기법을 적용하기에 앞서, 위와 같은 두 가지 방식으로 결측값에 대한 전처리를

수행할 수 있다. 본 논문에서는 결측값에 대해 다음과 같은 두 가지 방식의 전처리 기법을 적용하여 보았다.

① 결측값을 모두 -1로 치환

② 직선 보간법을 수행하여 결측값 치환

결측값을 모두 -1의 동일한 값으로 대체하는 전처리 방식은 대표적인 결측값 전처리 방식 중 첫 번째 방식의 일종으로 볼 수 있다. 결측값을 모두 -1로 치환하는 이유는 루프 센서의 실측값 범위가 0 이상의 정수이기 때문이다. 한편, 직선 보간법을 수행하는 전처리 방식은, 대표적인 전처리 방식 중 두 번째와 같이 실측값들을 이용하여 결측값을 보간하는 방식이다. 본 논문에서 실험에 사용한 교통량 패턴 데이터는 비선형적 특징을 갖는 IoT 데이터의 한 종류로, 데이터 특징을 고려한 적합한 보간법을 특정하는 것이 어렵다. 따라서 결측값이 발생한 위치의 데이터가 선형적 증감패턴을 갖는다는 가정에 기반하여 직선 보간법을 통한 전처리를 수행하였다. 이러한 두 가지 전처리 기법들은 결측이 발생한 위치의 참값에 대한 높은 신뢰도로 결측값을 복원할 수는 없지만, 결측이 발생한 위치에 분석 가능한 수치 데이터를 할당하여 이후의 분석 작업 수행이 가능하도록 한다.

한편, 본 논문에서 제안하는 기법에서는 합성곱 신경망에 의해 패턴 분류를 수행하게 된다. 합성곱 신경망은 패턴을 구성하는 모든 데이터에 대하여 동일한 중요도를 가지고 특징 추출을 위한 합성곱 및 풀링 연산을 수행한다. 그런데 IoT 센서 스트림 데이터 패턴과 같은 비선형적인 특징을 갖는 데이터에 대해서는, 결측값을 대체하기 위한 적합한 대푯값이나 결측값을 효과적으로 유추할 수 있는 보간법을 결정하는 것이 쉽지 않다. 즉, 비선형적인 데이터의 특징 추출에 합성곱 신경망을 이용하는 경우에는, 결측값에 대한 전처리를 수행하는 것만으로는 결측값에 의한 분석 정확도 손실을 완전히 보완하기 어렵다.

따라서 본 논문에서 제안하는 기법은 분석 작업에서의 수치 계산을 위하여 결측값에 대한 전처리를 마친 데이터라 하더라도, 전처리된 결측값에 비해서 실측값에 높은 가중치를 두고 패턴 분류를 위한 특징 추출을 수행하기 위하여 어텐션 메커니즘을 도입하였다. 즉, 본 논문에서 제안하는 패턴 분류 기법을 적용하는 경우, 어텐션 메커니즘의 입력 데이터는 전처리를 통하여 결측값의 대체 혹은 보간을 수행한 벡터 데이터가 된다. 그러나 전처리 되지 않은 기존 입력 데이터에서의 결측값 발생 여부와 발생 위치를 표현하는 메타 데이터가 함께 입력으로 주어지기 때문에, 전처리 되지 않은 비손실 데이터 부분이 전처리 된 부분에 비해 높은 가중치를 갖게 된다. 이러한 어텐션 메커니즘을 통하여 비손실 데이터에 높은 가중치를 두어 결측값으로 인한 분류 정확도 손실을 보완할 수 있게 된다. 본 논문에서는 위의 두 가지 전처리 기법들과 제안하는 패턴 분류 기법을 조합하였을 때의 효과를 4장의 실험을 통하여 분석하였다.

4. 실험 및 결과 분석

이 장에서는 결측값을 포함한 손실 데이터에 대한 제안하는 기법의 패턴 분류 성능 검증을 위하여, 실제 결측값이 발생한 루프 센서 데이터 스트림인 캘리포니아 주의 교통량 데이터 셋에 대하여 실험을 수행하였다. 먼저, 실험에 사용한 교통량 데이터 셋을 구성하는 과정에 대해 기술한다. 그런 뒤 실험에 사용한 모델의 학습 조건과 성능 검증을 위한 실험 개요를 기술한 뒤, 실험 결과에 대한 분석을 수행한다.

4.1 실험 데이터 셋 구성

본 논문에서는 미국 캘리포니아 주의 교통량 측정 시스템인 Caltrans PeMS[19]의 루프 센서[3]를 통해 수집된 실제 교통량 데이터로부터 데이터 셋을 구

성하여 제안하는 기법의 검증에 활용하였다. 본 논문의 실험은 캘리포니아 주의 각 도로에 분포된 Caltrans PeMS 루프 센서들 중, 다저스 구단의 경기가 열리는 경기장 근처 도로의 루프 센서에서 수집된 교통량 데이터를 대상으로 하였다. 즉, 본 논문에서는 다저스 구단의 경기가 열리는 경기장 주변의 교통량 데이터 패턴에 따라 다저스 구단의 경기일 여부를 분류하는 이진 분류 실험을 통하여 제안하는 기법의 검증을 수행하였다. 따라서, 다저스 구단의 경기 시즌이 아닌 날짜의 교통량 데이터는 이러한 이진 분류 문제의 범위 및 조건에 적합하지 않으므로, 다저스 구단의 경기 일정을 조사하여 경기 시즌인 4월부터 9월까지의 날짜에 해당하는 교통량 데이터를 대상으로 하였다.

한편, 제안하는 기법의 검증을 위해서는 데이터 셋에 결측값이 포함 되어야만 한다. 하지만 정상 측정된 데이터에 비하여 결측값이 차지하는 비율이 너무 커지면, 이 또한 실험 결과에 영향을 주는 변인으로 작용하게 된다. 따라서 제안하는 기법 외에 실험 결과에 영향을 줄 수 있는 다른 변인들을 통제하기 위해서는 적절한 수준의 결측값이 포함된 데이터 셋을 구성할 필요가 있다. 본 논문에서는 이러한 변인 통제를 위하여 Caltrans PeMS의 연도별 교통량 데이터 중, 결측값 비율이 너무 크지 않은 [표 2]의 5개 연도 교통량 데이터를 선정하였고, 결측값 비율은 연도별로 약간의 차이는 존재 하지만 평균 2.68% 정도이다. 또한, [표 2]의 3열과 같이, 실험에 사용한 5개 데이터 셋 각각에 대하여 다저스 구단의 경기 개최일과 경기 개최일이 아닌 날의 비율을 산출하였다. 이는, 경기 개최일 여부를 이진 분류하는 본 논문의 실험에 대한 데이터 레이블의 분포를 살펴보기 위함이다. 경기 개최일 레이블 비율 산출 결과, 실험에 사용한 5개 데이터 셋 모두에 대하여 최소 42.62%, 최대 44.26%의 비율을 확인할 수 있었으며, 실험에 사용한 데이터 셋은 크게 편향되지 않은 레이블 분포를 가진다고 볼 수 있다.

연도	데이터셋 크기 (날짜 수)	경기개최일 비율 (%)	결측값 비율 (%)
2007	158	44.26	1.66
2008	181	43.72	2.91
2011	148	43.72	2.80
2012	183	42.62	0.44
2017	162	44.26	5.60

[표 2] 실험 데이터 셋 명세

모델 학습 관련 파라미터	설정값
드롭아웃 비율	0.2
학습률	0.001
에폭 수	100
배치 크기	20

[표 3] 모델 학습 관련 파라미터와 설정값

선정된 5개 연도로부터 [표 2] 같은 5개의 실험 데이터 셋을 도출하기 위해서, 5분 단위로 측정된 교통량 데이터를 하루 단위의 시계열 데이터 벡터로 구성하였다. 즉, 한 개의 시계열 데이터 벡터는 5분 단위로 측정된 하루 동안의 교통량 벡터로, (288, 1) 차원을 갖는다. 이렇게 구성된 데이터 셋 중, 하루 동안의 교통량 벡터가 모두 손실된 경우는 데이터 셋에서 제외 하였다. 그 결과, 최종적으로 실험에 사용된 5개 연도 데이터 셋의 크기는 [표 2]와 같다. 학습 및 테스트를 위한 레이블 즉, 다저스 구단의 경기일인지의 여부는 해당 연도의 다저스 구단 경기 일정을 참조하여 레이블링하였다.

4.2 모델 학습 조건 및 실험 개요

제안하는 기법은 어텐션 메커니즘 및 합성곱 신경망 기반의 딥러닝 모델을 활용하며, 이 딥러닝 모델을 구성하는 합성곱 신경망의 구조에 대한 파라미터들은 3절에서 기술한 최적화 과정에 따라 [표 1]과

같은 최적값을 갖도록 하였다. 본 논문의 실험을 위해서는 이렇게 설계 완료된 모델의 학습이 우선 수행되어야 한다. 학습을 위하여 [표 2]의 각 데이터 셋에 대하여, 각 데이터 셋의 75%를 학습 데이터 셋으로, 10%를 검증 데이터 셋으로, 15%를 실험 결과 도출을 위한 테스트 데이터 셋으로 사용하였다. 또한, 드롭아웃 비율, 학습률, 에폭 수, 배치 크기와 같이 모델의 학습에 관련된 파라미터의 값은 [표 3]과 같이 설정하였고, 이러한 과정을 거쳐 학습이 완료된 모델을 실험에 사용하였다. 제안하는 기법의 성능 검증을 위한 실험은 두 가지 방법으로 진행되었으며, 각각 실험 1과 실험 2로 표기한다.

실험 1은 [표 4]와 같이 설계 및 수행되었으며, 제안하는 기법에서 도입한 어텐션 메커니즘의 효과 검증과 3절에서 기술한 두 가지 전처리 기법들을 제안하는 기법과 조합하였을 때의 성능 비교를 목표로 한다. 따라서 먼저, [표 2]의 5개 실험 데이터 셋에 대하여 3절에서 기술한 손실 데이터에 대한 두 가지 전처리를 각각 수행한 10개 케이스를 도출하였다. 그런 뒤, 10개 케이스 각각에 대해서 제안하는 기법 즉, 어텐션 메커니즘을 도입한 패턴 분류 기법과 어텐션 메커니즘을 도입하지 않고 합성곱 신경망 및 전결합 신경망으로만 구성된 패턴 분류 기법을 수행하였을 때의 분류 정확도를 측정하여 [표 4]와 같이 총 20개 케이스에 대한 데이터를 얻었다.

실험 2는 [표 5]와 같이 설계 및 수행되었으며, 어텐션 메커니즘이 실제 결측값을 포함한 데이터의 분류 성능에 미치는 영향의 정도를 확인하는 것을 첫 번째 목표, 그리고 반대로 결측값을 전혀 포함하지 않은 비손실 데이터의 분류 성능에는 어떠한 영향을 미치는지 확인하는 것을 두 번째 목표로 한다. 따라서, 각 연도별 데이터 셋에 대해서, 결측값이 포함된 교통량 벡터들로 구성된 부분 데이터 셋과 결측값이

연도	전처리	어텐션 메커니즘	정확도 (%)
2007	직선 보간	적용	91.30
	-1 치환		92.75
2008	직선 보간		85.19
	-1 치환		74.07
2011	직선 보간		92.42
	-1 치환		81.82
2012	직선 보간		90.12
	-1 치환		91.36
2017	직선 보간		70.83
	-1 치환		54.17
2007	직선 보간	적용 안함	86.96
	-1 치환		85.51
2008	직선 보간		75.31
	-1 치환		79.01
2011	직선 보간		62.12
	-1 치환		65.15
2012	직선 보간		75.31
	-1 치환		88.89
2017	직선 보간		66.67
	-1 치환		61.11

[표 4] 실험 1 수행 결과

연도	결측값	어텐션 메커니즘	정확도 (%)
2007	있음	적용	100
		적용 안함	100
2008		적용	95.00
		적용 안함	90.00
2011		적용	100
		적용 안함	88.24
2012		적용	100
		적용 안함	100
2017		적용	87.72
		적용 안함	85.96
2007	없음	적용	98.44
		적용 안함	97.66
2008		적용	95.74
		적용 안함	92.20
2011		적용	92.11
		적용 안함	95.61
2012		적용	99.40
		적용 안함	96.41
2017		적용	81.90
		적용 안함	85.71

[표 5] 실험 2 수행 결과

포함되지 않은 교통량 벡터들로 구성된 부분 데이터 셋, 이렇게 두 개의 부분 데이터 셋을 도출한다. 이러한 방식으로 도출된 총 10개의 부분 데이터 셋에 대하여, 각각 제안하는 기법 즉, 어텐션 메커니즘을 적용한 패턴 분류 기법과 어텐션 메커니즘의 적용 없이 합성곱 신경망 및 전처리 신경망만을 활용하는 패턴 분류 기법의 정확도를 측정한다. 이 때 전처리는 [표 4]의 실험 1 결과를 참고하여, 각 케이스에 대해 더 높은 성능을 보였던 전처리 방식을 적용한다.

4.3 실험 결과 및 분석

[표 4]와 같이 진행된 실험 1의 결과를 보면, 5개 연도 데이터 셋에 대하여 각각 두 가지 전처리를 수행한 총 10개 데이터 모두에 대하여, 제안하는 패턴 분류 기법을 적용하였을 때 즉, 어텐션 메커니즘을 도입한 패턴 분류 기법을 적용하였을 때 더 높은 분류 정확도를 보이는 것을 확인할 수 있다. 특히, 같은 전처리 조건에서 가장 높은 성능 향상을 보인 경우는 2011년 데이터 셋에 직선 보간 전처리를 수행한 경우, 어텐션 메커니즘을 도입함으로써 분류 정확도가 30.3% 향상 되었다. 가장 적은 성능 향상을 보인 경우는 2012년 데이터 셋에 -1 치환 전처리를 수행한 경우로, 어텐션 메커니즘을 도입함으로써 분류 정확도가 2.47% 향상 되었다. 나머지 3개 데이터 셋의 경우에도 어텐션 메커니즘을 도입하였을 때 분류 정확도가 평균 7.1% 향상하는 것을 확인할 수 있었다. 이를 통해서, 어텐션 메커니즘을 도입한, 제안하는 패턴 분류 기법은 결측값이 포함된 센서 스트림 데이터의 분류 정확도를 효과적으로 향상시킴을 확인할 수 있다.

또한, [표 4]의 실험 1 결과에서 제안하는 기법과 함께 적용한 직선 보간과 -1 치환의 두 전처리 기법의 효과를 살펴보면, 2008, 2011, 2017년의 경우에는 직선 보간 전처리를 수행하였을 때 더 높은 성능 향상이 이루어졌으나, 2007, 2012년의 경우에는 -1 치환 전처리를 수행하였을 때 더 높은 성능 향상이 이루어졌다. 이는 각 데이터 셋이 가지고 있는 고유의 데이터 특징 분포에 따라 적합한 전처리 기법이 다르기 때문인 것으로 생각할 수 있다. 그러나, 제안하는 기법에 -1 치환 전처리를 수행하여 최고 성능을 달성한 경우는 직선 보간 전처리를 수행한 경우에 대하여 각각 1.45%, 1.24%로 정확도 향상 폭이 매우 작았다. 반면, 제안하는 기법에 직선 보간 전처리를 수행하여 최고 성능을 달성한 경우는 -1 치환 전처리를 수행한 경우에 대하여 각각 11.12%, 10.6%, 16.66%의 높은 정확도 향상 폭을 보인다. 따라서, -1 치환 전처리보다는 직선 보간 전처리가 제안하는 기법에 앞서 적용하기에 더 적합한 전처리 기법이라고 생각할 수 있다.

[표 5]와 같이 진행된 실험 2의 결과를 보면, 5개 연도 데이터 셋 각각에 대해서, 결측값이 포함된 교통량 벡터로만 구성된 부분 데이터 셋 5개를 추출했을 때, 이 5개의 부분 데이터 셋 모두에 대해서 어텐션 메커니즘을 도입한, 제안하는 패턴 분류 기법의 분류 정확도가 더 높았다. 2007, 2012년 데이터 셋은 어텐션 메커니즘 적용 여부에 관계 없이, 결측값에 대해서는 모두 100%의 분류 정확도를 보였다. 즉, 이 두 데이터 셋은 고유의 특성의 분포가 분류 기준에 매우 명확하게 분포되어 있는 특별한 경우였다고 생각할 수 있다. [표 4]의 실험 1 수행 결과에서도 이 두 데이터 셋은 실험 조건의 변화에 거의 무관하게 높은 분류 정확도를 달성한 것을 확인할 수 있다. 반면, 2008, 2011, 2017년 데이터 셋의 경우에는 결측값에 대해서는 어텐션 메커니즘을 도입하였을 때 각각, 5%, 11.76%, 1.76%의 분류 정확도 향상

을 보였다. 즉, 제안하는 기법은 결측값이 포함된 센서 스트림 데이터에 대해서는 매우 효과적으로 분류 성능을 향상시킨다고 할 수 있다.

반면, [표 5]의 실험 2 수행 결과에서, 5개 연도 데이터 셋 각각에 대해서 결측값이 전혀 포함되지 않은 비손실 교통량 벡터로만 구성된 부분 데이터의 경우에는 어텐션 메커니즘을 적용함으로써 분류 정확도가 향상된 케이스 3개, 어텐션 메커니즘을 적용하지 않을 때 오히려 분류 정확도가 향상된 케이스가 2개이다. 또한, 각 케이스의 정확도 향상 폭도 최대 약 3% 정도로 비슷한 수준이다. 따라서, 비손실 데이터에 대해서는 어텐션 메커니즘의 적용이 정확도 향상에 큰 영향을 미치는 요소라고 보기는 어렵다. 다만, 일반적으로 전체 데이터 셋은 결측값이 포함된 손실 교통량 벡터와 비손실 교통량 벡터가 모두 포함되어 있고, 손실 교통량 벡터에 대한 정확도 향상 폭이 비손실 교통량 벡터에 대한 정확도 득실 폭에 비해서 훨씬 크기 때문에, 전체적인 관점에서 봤을 때 제안하는 패턴 분류 기법은 결측값을 포함한 센서 스트림 데이터에 대해 높은 성능 향상을 이끌어 낼 수 있다고 할 수 있다.

5. 결론

본 논문에서는 실세계 또는 IoT 환경의 센서 스트림 데이터에서 흔히 발생하는 결측값에 의한 패턴 분류 정확도의 손실을 해결하기 위한, 어텐션 메커니즘 및 합성곱 신경망 기반의 패턴 분류 기법을 제안하였다. 제안하는 기법은 결측값에 의한 정확도 손실을 보완하기 위하여 어텐션 메커니즘을 도입하였고, 이를 통하여 결측값이 아닌 비손실 데이터에 더 높은 가중치를 부여할 수 있도록 하였다. 따라서, 합성곱 신경망 및 전결합 신경망으로 구성된 패턴 분류 모델은 원래 입력 데이터에 결측값이 포함되어 있더라도 비손실 데이터에 집중하여 패턴 분류를 수행할 수 있기 때문에, 향상된 분류 정확도를 달성할 수 있다.

본 논문에서는 측정 과정에서의 결측값 발생이 잦은 실세계 교통량 센서 데이터 스트림을 대상으로, 제안하는 기법의 효과를 여러 측면에서 검증하였다. 실험 결과, 제안하는 기법의 어텐션 메커니즘 도입이 결측값을 포함한 손실 데이터의 패턴 분류 성능을 최대 30.3%, 평균 10.8% 정도 향상시킴을 확인할 수 있었다. 또한, 제안하는 기법을 적용하기에 앞서 결측값에 대한 직선 보간 전처리를 수행하는 것이 성능 향상에 도움이 될 수 있음을 실험을 통해 확인하였다.

한편, 본 논문에서 실험에 사용한 루프 센서를 통해 수집된 교통량 데이터의 경우, 결측값이 랜덤하게 발생한다는 특성이 있다. 따라서 향후, 이러한 랜덤한 결측값의 발생 패턴 외에, 특정한 결측값 발생 패턴을 갖는 데이터 셋에 제안하는 기법을 적용하여 추가 실험을 진행한 뒤, 연구를 확장할 예정이다. 또한, 본 논문에서는 주로 정확도 측면의 분석을 수행하였는데, 향후, 연구를 확장하는 과정에서는 정확도 측면의 분석 외에도 다른 분석 기준들을 도입하여 다양한 측면에서의 분석이 가능하도록 할 예정이다. 특히, 실험 2를 확장하여 결측값으로만 구성된 데이터 셋과 비손실 데이터로만 구성된 데이터 셋에 대하여 각각 어텐션 메커니즘을 통한 가중치 측면에서 분석을 수행하여[20], 본 논문에서 제안하는 어텐션 메커니즘이 결측값이 포함된 데이터에 대한 패턴 분류 성능 향상에 도움을 주는 작동 원리를 더욱 심도 있게 규명하기 위한 향후 연구를 수행할 예정이다.

참고 문헌

[1] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, “A brief survey of machine learning methods and their sensor and IoT

applications," in 2017 8th International Conference on Information, Intelligence, Systems Applications (IISA), Aug. 2017, pp. 1–8, doi: 10.1109/IISA.2017.8316459.

[2] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," IEEE Communications Surveys Tutorials, vol. 20, no. 4, pp. 2923–2960, Fourthquarter 2018, doi: 10.1109/COMST.2018.2844341.

[3] R. H. Lees and R. A. Lees, "Loop sensing apparatus for traffic detection," US6483443B1, Nov. 19, 2002.

[4] S. Latif, H. Afzaal, and N. A. Zafar, "Intelligent traffic monitoring and guidance system for smart city," in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Mar. 2018, pp. 1–6, doi: 10.1109/ICOMET.2018.8346327.

[5] J. Qiu, L. Du, D. Zhang, S. Su, and Z. Tian, "Nei-TTE: Intelligent Traffic Time Estimation Based on Fine-Grained Time Derivation of Road Segments for Smart City," IEEE Transactions on Industrial Informatics, vol. 16, no. 4, pp. 2659–2666, Apr. 2020, doi: 10.1109/TII.2019.2943906.

[6] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 4, pp. 758–769, Apr. 2018, doi: 10.1109/TNSRE.2018.2813138.

[7] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," Data Min Knowl Disc, vol. 33, no. 4, pp. 917–963, Jul. 2019, doi: 10.1007/s10618-019-00619-1.

[8] J. C. B. Gamboa, "Deep Learning for Time-Series Analysis," arXiv:1701.01887 [cs], Jan. 2017, Accessed: Jul. 23, 2020. [Online]. Available: http://arxiv.org/abs/1701.01887.

[9] S. Hershey et al., "CNN architectures for large-scale audio classification," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017, pp. 131–135, doi: 10.1109/ICASSP.2017.7952132.

[10] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," Journal of Systems Engineering and Electronics, vol. 28, no. 1, pp. 162–169, Feb. 2017, doi: 10.21629/JSEE.2017.01.18.

[11] C.-L. Liu, W.-H. Hsaio, and Y.-C. Tu, "Time Series Classification With Multivariate Convolutional Neural Network," IEEE Transactions on Industrial Electronics, vol. 66, no. 6, pp. 4788–4797, Jun. 2019, doi: 10.1109/TIE.2018.2864702.

[12] A. V, G. P, V. R, and S. K p, "DeepAirNet: Applying Recurrent Networks for Air Quality Prediction," Procedia Computer Science, vol. 132, pp. 1394–1403, Jan. 2018, doi: 10.1016/j.procs.2018.05.068.

[13] S. Kumar, L. Hussain, S. Banarjee, and M. Reza, "Energy Load Forecasting using Deep Learning Approach-LSTM and GRU in Spark Cluster," in 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), Jan. 2018, pp. 1–4, doi: 10.1109/EAIT.2018.8470406.

[14] N. Ramakrishnan and T. Soni, "Network Traffic Prediction Using Recurrent Neural Networks," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA),

Dec. 2018, pp. 187-193, doi: 10.1109/ICMLA.2018.00035.

[15] E. Bengio, P.-L. Bacon, J. Pineau, and D. Precup, “Conditional Computation in Neural Networks for faster models,” arXiv:1511.06297 [cs], Jan. 2016, Accessed: Jul. 23, 2020. [Online]. Available: http://arxiv.org/abs/1511.06297.

[16] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” arXiv:1508.04025 [cs], Sep. 2015, Accessed: Jul. 23, 2020. [Online]. Available: http://arxiv.org/abs/1508.04025.

[17] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2016, pp. 4945-4949, doi: 10.1109/ICASSP.2016.7472618.

[18] B. Wang, K. Liu, and J. Zhao, “Inner Attention based Recurrent Neural Networks for Answer Selection,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Aug. 2016, pp. 1288-1297, doi: 10.18653/v1/P16-1122.

[19] “Caltrans PeMS > State of California > Overview > Dashboard.” http://pems.dot.ca.gov/ (accessed Jul. 23, 2020).

[20] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, Jun. 2016, pp. 1480-1489, doi: 10.18653/v1/N16-1174.

이 은 진

2016년~현재 이화여자대학교 컴퓨터공학과 재학

관심분야 : 빅 데이터 분석, 데이터마이닝

오 소 연

2016년 이화여자대학교 컴퓨터공학과 (공학사)

2018년 이화여자대학교 대학원 컴퓨터공학과 (공학석사)

2018년~현재 이화여자대학교 대학원 컴퓨터공학과 박사과정

관심분야 : 빅 데이터 분석, 스트림 데이터 분석, 딥러닝, 대용량 데이터 마이닝, 대용량 그래프 마이닝

이 민 수

1992년 서울대학교 컴퓨터공학과 (공학사)

1995년 서울대학교 대학원 컴퓨터공학과 (공학석사)

2000년 University of Florida 컴퓨터공학과 (공학박사)

1995~1996 LG전자 연구소 연구원

2000~2002 미국 Oracle Corporation Senior Member of Technical Staff

2002~현재 이화여자대학교 컴퓨터공학과 교수

관심분야 : 데이터웨어하우스, 빅 데이터 분석, 워크플로우, 스트림 데이터, 데이터마이닝, 스마트 응용