합성곱 신경망을 이용한 데이터스트림 환경에서의 개념 변화 검출 기법

김대원, 임효상
연세대학교 원주캠퍼스 컴퓨터정보통신공학부
e-mail: {dwkim1214, hyosang}@yonsei.ac.kr

A Method for Detecting Concept Drift in Data Stream by Using Convolutional Neural Network

Daewon Kim and Hyo-Sang Lim
Computer and Telecommunications Engineering Division,
Yonsei University, Wonju

요 약

본 논문에서는 데이터스트림 환경에서 개념 변화를 탐지하기 위해 합성곱 신경망(CNN)을 사용하는 방법을 제시한다. 데이터스트림 환경에서 입력될 수 있는 데이터를 패턴화하여 신경망 모델에 학습시키고, 패턴화한 데이터를 학습시킨 신경망 모델을 이용하여 스트림 환경에서 개념 변화를 검출 가능함을 보인다.

1. 서론

스마트 디바이스의 보급과 더불어 사물인터넷과 같은 센서 컴퓨팅 환경의 발전에 따라 많은 양의 스트림 데이터가 발생하고 있다[1]. 스트림 데이터는 빠르고 지속적으로 발생하는 데이터의 시퀀스를 말한다. 스트림 데이터는 지속적으로 대용량의 데이터로써 입력되는 특성을 가지고 있어, 빠르고 정확하게 데이터를 처리 하는 것이 중요하다. 본 논문은 데이터 스트림 환경에서 발생하는 특성 중 하나인 개념 변화를 검출하기 위한 기법을 수행한다.

개념 변화(concept drift)란 입력되는 데이터의 특성이 시간에 따라 변화하는 것을 의미한다[2]. 스트림 데이터에서의 개념 변화란 시퀀스로써 입력되는 데이터의 특성이 시간에 따라 변화하는 것을 의미하는데, 데이터의 특성의 예로는 데이터의 분포 혹은 통계량 등이 있다. 스트림 데이터에서 개념 변화가 발생하게 되면, 변화한 데이터의 특성에 맞는 처리를 수행할 필요가 있다. 예로써, 시퀀스로써 입력되는 심박수 데이터의 통계적 변화를 감지하여 긴급한 상황을 탐지하는 상황 등이 있다.

본 논문에서는 데이터스트림 환경에서 개념 변화를 탐지하기 위해 신경망을 응용하는 기법을 제시한다. 신경망은 인간의 두뇌 조직의 기본 단위인 뉴런의 동작을 모사하기 위해 구성한 네트워크를 말한다. 신경망은 분류, 회귀 등의 문제를 해결하기 위해 사용되며 본 논문은 신경망을 통한 데이터의 분류를 통해 데이터스트림 환경에서 개념 변화를 탐지한다.

2. 관련 연구

2.1. 데이터스트림 환경에서의 개념 변화 관련 연구

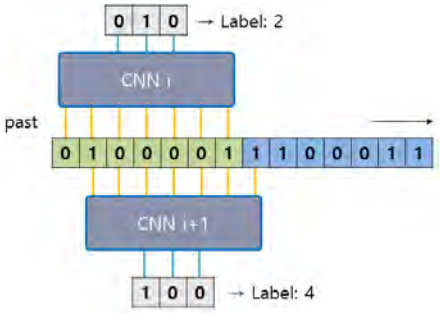
지속적으로 대용량의 데이터가 입력되는 특성을 갖는 데이터스트림 환경에서는, 입력되는 데이터를 모두 저장하는 것이 어렵기 때문에 윈도우를 사용한다. 윈도우는, 일정량의 데이터를 저장할 수 있는 공간을 말한다. 데이터가 입력될 때마다 입력된 데이터를 윈도우에 쌓아가며 윈도우 내에 있는 데이터에 대해 처리를 수행하다가 윈도우가 가득차면 윈도우에서 가장 오래된 데이터를 제거해가며 새로운 데이터를 쌓아 처리하는 방식으로 데이터를 처리한다. 인접한 두 시점의 데이터를 담는 두 윈도우 $i, i+1$의 데이터 확률 분포를 p_i, p_{i+1}이라 할 때, $p_i \neq p_{i+1}$인 경우 $i+1$ 시점에서 개념 변화가 발생한 것으로 판단한다[3].

2.2. 합성곱 신경망(convolutional neural network) 관련 연구

합성곱 신경망이란 이미지 인식을 위해 사용되는 합성곱 필터와 신경망 기술을 결합시킨 신경망이다[4]. 합성곱 필터를 이용하여 데이터의 특성은 유지하면서 데이터의 양을 줄이면, 신경망이 이를 학습하여 분류를 수행하는 방식으로 동작한다.

3. 신경망을 이용한 개념 변화 검출 방법

본 논문에서는 신경망을 이용하여 데이터스트림 환경에서의 개념 변화를 탐지하기 위해 [그림 1]과 같이 신경망을 배치한다.



[그림 1] 신경망을 이용한 개념 변화 검출 방법

학습된 신경망에 1개 윈도우 크기의 데이터를 입력하면, 학습 내용에 따라 스트림데이터의 분포와 연관된 데이터인 레이블을 얻을 수 있다. 이러한 신경망을 이용하여 개념 변화를 탐지하기 위하여 두 개의 신경망을 배치한다. 1개의 신경망에 i시점의 데이터를, 나머지 1개의 신경망에는 $i+1$시점의 데이터를 입력한다. 두 신경망의 레이블을 상호 비교하여 두 레이블이 서로 다르다면 개념 변화가 발생한 것으로 판단하고, 그렇지 않다면 개념 변화가 발생하지 않은 것으로 판단한다.

4. 실험

4.1. 실험 신경망 및 데이터스트림 환경

실험으로 사용할 합성곱 신경망으로 google TensorFlow [5]에서 제공하는 CIFAR-10 모델을 사용하였다. CIFAR-10은 32x32 크기의 RGB 이미지와 10종류의 레이블 중 하나가 세트로 구성된 데이터 60,000개로 구성된 데이터 셋[6]이며, CIFAR-10 모델은 이미지와 레이블 쌍을 학습하여 학습 이후에 입력되는 이미지에 대한 레이블을 출력하는 모델이다.

본 논문에서는 0 혹은 1만 입력되는 데이터스트림 환경을 가정하여 실험을 수행한다. 스트림 데이터로써 입력될 수 있는 이진수의 패턴을 윈도우 길이에 대한 1의 비율에 따라 아래 표와 같이 분류하여 데이터를 랜덤하게 50,000개 생성하여 CIFAR-10 모델의 학습을 시도한다.

1의 비율	레이블
0%	0
20%	1
40%	2
60%	3
80%	4
100%	5

[표 1] 윈도우 내 데이터 비율에 따라 분류한 스트림 데이터 및 레이블

4.2. 실험 데이터

학습된 신경망의 결과를 이용하여 개념 변화를 탐지하기 위하여, [그림 2]와 같은 실험 데이터를 생성한다.

시간 방향 →

1의 비율	60	80	20	40	0	60	20	0	40	100	80
윈도우 개수	5	10	10	10	10	10	10	10	10	10	5

[그림 2] 실험 데이터

실험 데이터는 순서가 있는 이진수의 배열이며, 처음에 1의 비율이 60%로 일정한 데이터가 윈도우 5개 분량만큼 입력된 후 1의 비율이 80%로 올라간 데이터 윈도우 10개 분량의 데이터가 입력되는 방식이다. 실험에 사용할 두 신경망은 입력된 데이터를 처리한 뒤 일정한 간격으로 데이터 접근 범위를 우측으로 이동하는 방식으로 데이터에 접근한다. 접근한 두 시점 데이터에 대한 레이블을 각각의 신경망이 계산하면 이 두 값을 상호 비교하여 개념 변화 여부를 판단한다.

실험 결과의 비교를 위하여, 학습 회수를 구분하여 따로 실험을 하였다. 학습은 1회에 1,000개의 데이터를 학습하며 이를 20만 회(200K), 10만 회(100K) 학습한 모델을 준비하였으며, 정보는 [표 2]와 같다. 두 모델은 임의의 레이블을 갖는 데이터에 대한 적중률로 각각 97%, 78%를 가진다.

모델 구분	200K	100K
학습 회수	200000	100000
임의의 1개 윈도우에 대한 레이블 적중률 평균	97%	78%

[표 2] 실험에 사용한 두 CNN 모델

4.3. 실험 결과

두 모델이 실험 데이터에 대해 개념 변화 탐지를 수행한 결과는 [표 3]과 같다. 각 값의 단위는 윈도우 1개 크기에 대한 비율이다. 데이터 분포가 변하는 상황에서 시간에 따라 두 개의 레이블이 번갈아가며 나타나는 상황을 진동이라 표현하였다. 변화를 임의 시점에 부여하였을 때, 변화에 대응하여 레이블의 변화를 일으키는 시점은 200K 모델이 변화가 일어난 후 평균적으로 윈도우 1.00개 크기만큼 더 이동한 후였으며, 100K는 윈도우 239.1개만큼 더 이동한 후에 개념 변화를 감지하였다. 개념 변화 횟수 이상으로 개념 변화를 탐지한 횟수인 중복 알람의 수와 개념 변화 탐지 성공률은 각 모델의 레이블 적중률에 따라 크게 달라지는 것을 확인할 수 있었다.

모델 구분	200K	100K
윈도우 크기	24576(bit)	
진동 폭	0.93	7573.8
변화 부여 후 개념 변화 감지가 끝날 때까지의 윈도우 개수	1.00	239.1
변화 부여 후 개념 변화 감지까지의 윈도우 개수	0.08	7334.7
중복 알람의 수	75.38	1179.4
탐지 성공율	90%	50%

5. 결론

본 논문은 데이터스트림 환경에서 입력될 수 있는 데이터를 패턴화하여 CNN 모델에 학습시키고, 패턴화한 데이터를 학습시킨 신경망 모델을 이용한 실험을 통해 스트림 환경에서 개념 변화를 검출 가능함을 보였다.

참고문헌

[1] A. Haque, L. Khan, M. Baron, B. Thuraisingham and C. Aggarwal, "Efficient handling of concept drift and concept evolution over Stream Data," 2016 IEEE 32nd International Conference on Data Engineering

[2] A. Haque, L. Khan, and M. Baron, "Semi supervised adaptive framework for classifying evolving data stream", In Advances in Knowledge Discovery and Data Mining, volume 9078 of Lecture Notes in Computer Science, Springer International Publishing.

[3] Indr˙e Zliobait˙e, "Learning under Concept Drift: an Overview", Technical report, Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania, 2009.

[4] Yoon Kim, "Convolutional Neural Networks for Sentence Classification", New York University

[5] Convolutional Neural Networks, Google, https://www.tensorflow.org/tutorials/deep_cnn, 2017

[6] The CIFAR-10 dataset, Alex Krizhevsky, http://www.cs.toronto.edu/~kriz/cifar.html