다차원 스트림 데이터의 온라인 점진적 학습을 위한 순차적 예측 모델

허민오[O] 이상우 장병탁
서울대학교 컴퓨터공학부[1]
{moheo, slee, btzhang}@bi.snu.ac.kr

A Sequential Prediction Model for Online Incremental Learning of Multidimensional Stream Data

Min-Oh Heo[O] Sang-Woo Lee Byoung-Tak Zhang
School of Computer Science & Engineering, Seoul National University

요 약

최근 실생활 속에 다양한 센서 데이터가 나타남에 따라, 실시간으로 끊임없이 유입되면서도 장기적으로 분포가 변할 수 있는 데이터가 늘어나고 있다. 이러한 데이터의 학습에는 점진적인 학습을 수행하면서도 동시에 순차적 예측이 가능한 모델을 필요로 한다. 이에 따라, 본 고에서는 다차원 스트림 데이터를 위한 시계열 예측을 다루는 새로운 모델을 제안한다. 이 모델은 순차적인 정보를 지닌 패턴들의 집합과 해당 패턴들이 나타난 빈도를 가지고 있으며, 이를 기반으로 시계열 예측을 시도한다. 또한, 파라미터의 폭발적인 증가를 막고 효율적인 표현을 위하여, 패턴 집합의 개수가 특정 수 이상 늘어나지 않도록 제약을 둘 수 있다. 본 모델의 효용성을 평가하기 위해, 분포를 아는 문자서열을 생성하여 조건부 확률분포가 학습됨을 보였다. 또한, 스마트폰으로 수집한 GPS 데이터를 이산화하여 이제까지 이동해온 도로를 통해 다음에 이동할 도로를 예측하는 문제를 해결하고, 그 분석 결과를 통해 본 모델의 특성을 확인하였다.

1. 서 론

최근 스마트폰과 같이 실생활 속에 쉽게 다양한 센서 데이터를 제공하는 기기가 나타남에 따라, 실시간으로 끊임없이 유입되면서도 장기적으로 분포가 변할 수 있는 데이터가 늘어나고 있다[1]. 전통적인 시계열 형태의 순차적 데이터 모델링 기법은 오프라인 학습이 주로 사용되었으며, 장기적인 변화를 다루기 위해서는 긴 기간의 데이터 수집이 불가피 하였다[2-4]. 즉, 한번 확보하였던 데이터를 이용하여 모델을 구축한 후, 문제해결에 활용하는 방식을 주로 취해왔다. 흔히 사용하는 시계열 데이터 모델링 방법으로 재귀적 뉴럴 네트워크(recurrent neural network (RNN))[2], 은닉 마코프 모델(hidden Markov model (HMM))[3]이 사용되는데, 주로 단기적인 패턴학습에 사용된다. 특히, HMM과 같이 모델에 마코프 가정을 적용하였을 경우, 이전 시간의 값에 따라 현재값이 결정된다.

끊임없이 유입되는 데이터, 즉, 스트림 데이터를 학습하기 위해서는 오프라인 형태가 아니라, 시간의 흐름에 따라 함께 학습을 수행하는 온라인 학습(online learning)이 가능해야 한다[5]. 장기적 관점에서 사용자의 환경변화에 영향을 받아 유입되는 데이터의 분포가 변할 경우에도 지속적인 학습을 통해 concept drift를 다룰 수 있어야 함을 의미한다[6, 7]. 또한, 이러한 과정을 지속적으로 수행하면서도 사용자의 요청에 따라 현재 모델을 이용하여 추론하는 것이 가능해야 한다. 즉, 스트림 데이터를 이용하여 점진적인 온라인 학습을 수행하면서도 동시에 순차적 예측이 가능한 모델이 요구되고 있다.

본 고에서는 스트림 데이터를 위한 시계열 예측을 다루는 새로운 모델을 제안한다. 이 모델은 순차적인 정보를 나타내는 패턴들의 집합과 해당 패턴들이 나타난 빈도를 모델로서 가지고 있으며, 이를 기반으로 시계열 예측을 시도한다. 점진적인 온라인 학습은 데이터가 유입되는 매 시점마다 등장한 패턴의 빈도를 늘리는 것으로 단순하게 진행되며, 모델을 이용해 예측을 시도한다. 모델의 학습으로서, 유입된 데이터와 비교하여 정보가 부족하거나 수정이 필요하다고 판단될 경우, 패턴집합 내에 새로운 패턴을 추가하거나 수정/삭제하는 방법을 사용할 수 있다.

이후, 본 논문은 다음과 같이 구성된다. 2 장에서는 스트림 데이터의 점진적 온라인 학습을 위한 모델을 소개하고, 학습 및 추론 방법을 설명한다. 3 장에서는 실험을 통해 모델의 적용 예를 보인 후, 4 장에서 결론을 맺는다.

2. 제안 모델

2.1 문제 정의

본 모델에서 다루고자 하는 순차적 예측 문제는 최근 k 개의 이산변수에 대한 입력이 주어졌을 때, 바로 다음 단계 또는 n 단계 후의 입력이 무엇일지 예측하는

문제로서, 다음과 같이 표현된다.

$$\hat{x}_{t+n} = \arg\max_{x} P(X_{t+n} = x \mid X_{t-k:t})$$

여기서, n과 k는 임의의 양의 정수이다. 여기서, n이 1이고 k가 시계열 길이 전체를 나타내는 t-1인 경우에는 filtering 문제이며, n이 1보다 큰 경우에는 prediction 문제이다. 이러한 문제를 다루는 일반적인 접근 방법은 시계열 데이터에 대한 결합확률분포를 정의하여 확률이 최대가 되는 값을 구하는 방법이며, 대표적 방법으로 은닉 마코프 모델(HMM) 이 있다[3]. 하지만, 은닉 변수를 도입하게 되면 EM 기반의 학습방법을 사용해야 하므로, 스트림 데이터에 대한 온라인 학습방법으로 적절하지 않다.

2.2 모델 정의 및 추론

점진적으로 온라인 학습을 수행하기 위하여, 스트림 데이터로부터 등장하는 순차적 패턴의 빈도를 갱신하는 것으로 학습이 수행되도록 하는 모델을 설명한다. 그림 1에서와 같이, 스트림 데이터를 순차적 패턴의 등장빈도로 간결히 표현하고, 예측 수행 시, 이 모델로부터 사후확률을 구하는 것으로 요약할 수 있다.

모델 M=<H, w> 은 패턴의 집합 H와 패턴의 빈도 w의 쌍으로 정의된다. H는 임의의 순차적 패턴 E를 원소로 갖는 색인된 집합이며, i 번째 패턴을 E_i로 표기한다. 또한, w는 각 E가 스트림 데이터에서 등장한 빈도를 원소로 갖는 색인된 집합이며, E_i에 대한 원소를 w_i로 표기한다. 모든 E는 특정 시점을 기준으로 시간 차이와 값을 표현하는 순차패턴을 나타낸다. 이를 표현하기 위해, 상대적 시간 색인 u와 인자 색인 f가 부여된 이산 변수 X^f_u와 변수에 할당된 값 c로 이루어진 3-tuple (f,u,c)를 원소로 하는 집합 $\mathcal{X}$를 정의하자. $\mathcal{P}_E(\mathcal{X})$는 $\mathcal{X}$의 멱집합 $\mathcal{P}(\mathcal{X})$의 부분집합으로, 시간과 인자 쌍이 중복되지 않는 것들을 원소로 하는 집합이며, 패턴 E는 $\mathcal{P}_E(\mathcal{X})$의 임의의 원소이자, ($f$,$u$,$c$)들의 집합이 된다. 즉,

$$\mathcal{P}_E(\mathcal{X}) = \{E \mid E \in \mathcal{P}(\mathcal{X}),\ \forall i, j\ in\ E,\ (f_i, u_i) \neq (f_j, u_j)\}$$

단, i, j는 E 안에 있는 원소의 색인을 나타내는 양의 정수이다.

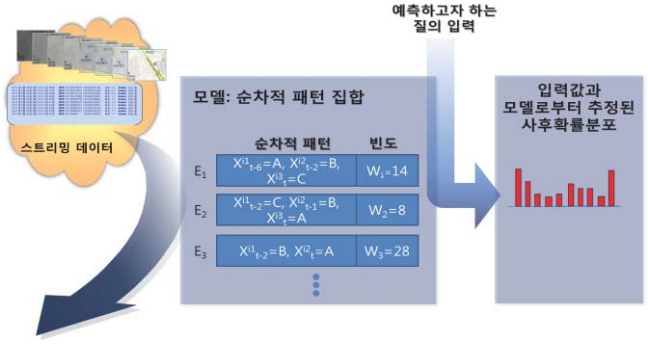

그림 1. 스트림 데이터를 점진적 온라인 학습할 수 있도록 모델의 갱신하는 과정을 단순하게 한 모델

t-k시점부터 t까지의 값이 질의로서 주어졌을 때, t+n 시점에 대한 예측을 다룰 경우, 조건부 확률 P가 최대가 되는 x가 결정되며, 주어진 M으로부터 다음과 같이 근사적인 비례관계를 얻는다.

$$P(X_{t+n} \mid X_{t-k:t} = x_{t-k:t}) \propto \frac{\sum_{E_i \in H} |c|^{|E_{i,u\geq n}|} \cdot w_i \cdot \delta(E_{i,u\geq n}, x_{t-k:t}) \cdot \delta(E_{i,u=0}, x_{t+n})}{\sum_{E_i \in H} |c|^{|E_{i,u\geq n}|} \cdot w_i \cdot \delta(E_{i,u\geq n}, x_{t-k:t})}$$

상기 식에서 $\delta(E_{i,u\geq n}, x_{t-k:t})$ 은 i 번째 패턴 E_i에서 시간 색인이 n 보다 큰 부분의 값이 질의와 동일하면 (또는 가까우면) 1, 아니면 0을 출력하는 함수이며, $\delta(E_{i,u=0}, x_{t+n})$ 는 패턴 E_i의 u=0인 원소와 예측하려는 값이 동일 (또는 가까움) 여부를 출력하는 함수이다.

2.3 학습 방법

모델 M은 이상적으로는 고려하는 시간 색인 길이 k에 따라 패턴 E의 개수가 지수적으로 증가한다. 이를 극복하기 위해 데이터에 나타나지 않은 패턴의 빈도는 0이므로, 실제 모델에는 아직 추가될 필요가 없으며, k가 클 경우, 모든 조합을 다 다룰 수 없으므로 가능한 패턴 중 일부를 무작위로 선택하여 모델에 반영한다.

이를 수행하는 학습 알고리즘을 그림 2에 제시하였다. 스트림 데이터가 입력됨에 따라, 스트림 데이터의 X_{t-w} ~ X_t 상에서 무작위로 시점과 인자를 선택하여 h개의 후보를 만든 후, M 안에 해당 패턴이 없을 경우 M에 포함시킨다. 그 후, 포함된 패턴에 대해서 입력되는 데이터와 비교, 맞는 패턴에 대해 해당 w를 증가시킨다.

여기서, 새로운 패턴이 무한히 발견될 경우, 모델도 이를 포함시켜야 하는 부담이 있다. 이를 회피하기 위해, 패턴 수의 상한선을 정하고, 패턴을 가질 수 있는데 의 수를 늘릴 수 없으므로, w, $|E|$가 모두 작은 패턴부터 모델에서 삭제할 수 있다. (그림 2에서 아래 revise 부분)

```
1.   M = empty          ( M=< H, w> )
2.   For  time_step t = 1 to T          %% T can go to infinity
3.        pred_t = predict(M)
4.        PE = {set of possible pattern E on sequence s at time t-k ~ t}
5.        new_E = {select_random_and_check_exist(PE, h)}
6.        H = H ∪ new_E
7.        w = w ∪ { w of new_E = '0' }
8.        M = incrementCount(M)
9.        M = revise(M)
10.  endFor
```

그림 2 간략히 표현한 학습 알고리즘

3. 실험 결과

실험을 통해 분포를 아는 인공 데이터를 통해 제안한 모델이 얼마나 실제 분포를 근사적으로 표현할 수 있는지 확인하고, [7]에서 쓰인 스마트폰으로 수집한 위치정보 데이터를 스트림 형태로 입력하게 하여 다음 시점의 위치할 도로 예측 문제의 성능을 보인다.

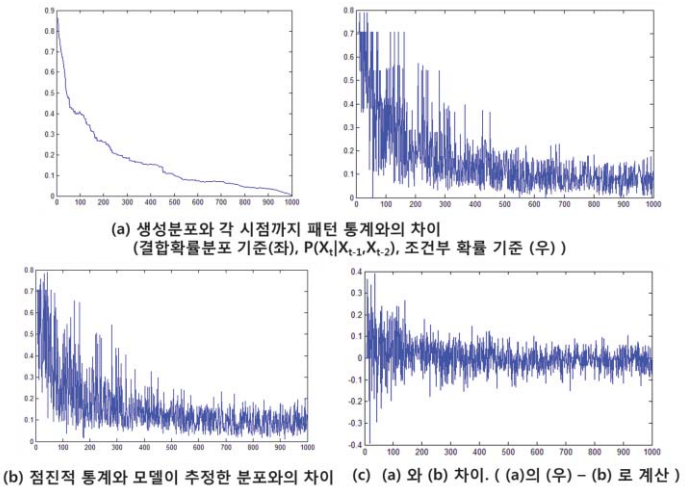

그림 3. 생성분포를 아는 스트림 데이터를 통해 모델이 분포를 학습한 결과. (x축은 시간, y축은 분포 사이의 차이(Hellinger Distance를 이용))

먼저 약간의 규칙을 부여한 분포를 통해 스트림 데이터의 대체 데이터를 생성하였다. 숫자 '1,2,3,4'를 가능한 c로 하고, 패턴 '1 2 3', '2 3 4', '3 4 1', ' 4 1 2'이 다른 패턴에 비해 4배 더 나타날 확률이 높도록 하여 길이 1000짜리 시계열 데이터를 생성하면서 이를 학습하였다. 사용한 모델 파라미터로서, n=1, k=4, 각 E가 가질 수 있는 최대 원소수는 4로 하였다. 시계열 그림 3 (a)에서 볼 수 있듯이 생성한 분포와 해당 시점까지 등장한 패턴의 수에 따른 통계는 다소 차이가 있으며 시계열이 길어질수록 그 차이가 줄어든다. 또한, 제안한 모델이 예측한 확률과 등장 패턴에 따른 통계 사이의 거리도 점차 줄어든다. (그림 4 (b), (c))

스마트폰 센서 데이터를 통해 현재 위치한 거리에서 다음에 이동하려는 거리를 예측하는 문제를 이 모델을 통해 적용해볼 수 있으며[8], 그 결과가 그림 4에 나타나 있다. 이를 위해 [8]에서 사용된 Action Logger를 이용하여, 실제 이동한 로그를 1초간격으로 4건(각 1247초, 1170초, 1719초, 1338초)을 수집하여 순서대로 누적되도록 학습하였다. (n=1, E의 최대 원소수는 3)

동일한 데이터를 반복하여 추가로 입력하는 것은 학습에 도움이 될 것이므로, 1 → 2 → 3 → 4 → 1 → 2 → 3 → 4 와 같은 순서로 4회 로그를 입력하여 학습하였고 그 결과가 그림 4에 도시 되어있다. 예측성능과 관련하여, 여러 로그가 추가되면서 방해가 생겨 예측 정확도는 다소 하락하지만 k를 늘리면 더 많은 정보를 이용하게 되어 성능이 증가한다.

4. 결론

본 고에서는 스트림 데이터를 다룰 수 있는 점진적 학습 모델을 소개하고 추론 및 학습 방법을 보였으며, 분포 비교 및 응용에의 예를 함께 제시하였다. 향후, 더욱 개선된 추론/학습 방법 연구와 다른 응용에의 적용을 기대한다.

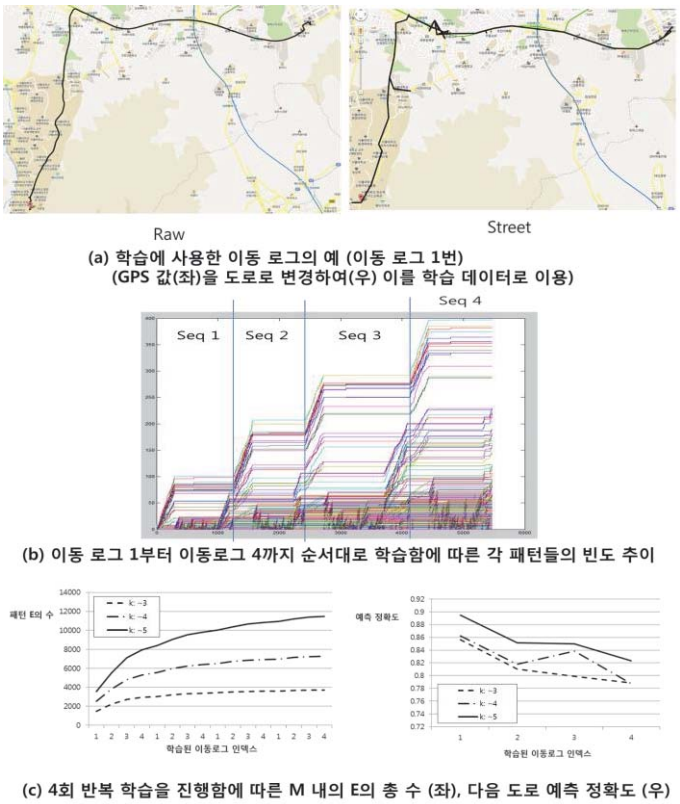

그림 4. 스마트폰 센서 데이터를 이용한 예측 결과

감사의 글

이 논문은 2012년도 정부(지식경제부)의 재원으로 수행된 연구이며(10035348, mLife), 한국연구재단의 지원(No. 2012-0005643, Videome) 및 BK21-IT 프로그램에서 일부 지원되었음.

참고문헌

[1] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, A Survey of Mobile Phone Sensing, *IEEE Communications Magazine*, vol.48, no.9, pp.140-150, 2010.

[2] Simon Haykin, Neural Networks and Learning Machine, Prentice Hall, 2009.

[3] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

[4] D. Barber, T. Cemgil, and S. Chiappa, Bayesian Time Series Models, Cambridge University Press, 2011.

[5] S. Shalev-Shwartz, Online Learning and Online Convex Optimization, *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp.107-194, 2012.

[6] G. Widmer and M. Kurat, Learning in the Presence of Concept Drift and Hidden Contexts, Machine Learning, vol. 23, pp.69-101, 1996.

[7] I. Zliobaite, Learning under Concept Drift: an Overview, Technical Report, Faculty of Mathematics and Informatics, Vilnius University, 2009.

[8] 허민오, 강명구, 임병권, 황규백, 박영택, 장병탁, 확률 그래프 모델을 이용한 스마트폰 사용자의 이동경로 학습 및 실시간 예측 기법, *정보과학회 논문지: 소프트웨어 및 응용*, 제39권 제6호, pp.425-435, 2012.6.