협업필터링 오토인코더 기반 센서 데이터 결측치 처리

정재익, 구태연, 박완기
한국전자통신연구원
jaeik1210@etri.re.kr, kutai@etri.re.kr, wkpark@etri.re.kr

Sensor Data Missing Imputation based on Collaborative Filtering with Autoencoder

Jaeik Jeong, Tai-Yeon Ku, Wan-Ki Park
Electronics and Telecommunications Research Institute

요 약

최근 에너지 최적 관리를 위한 인공지능 기반 기술들이 활발히 개발되고 있으며 이를 위해 센서 데이터 신뢰성의 중요도가 높아지고 있다. 기존 실물 센서의 경우 불안정한 센싱 환경으로 데이터에 결측치가 발생하는 경우가 많기 때문에 실물 센서 데이터를 보정해주는 가상센서 기술 개발이 필요하며, 최근 오토인코더를 기반으로 한 시스템 단위 가상센싱 방법이 데이터 결측치 처리에 높은 성능을 보이고 있다. 그러나 기존 오토인코더 기반 기술은 훈련 데이터에 결측치가 많을수록 성능이 떨어지는 경향을 보인다. 본 논문에서는 훈련 데이터에 결측치가 많은 환경에서 오토인코더의 결측치 처리 능력을 더욱 향상시키기 위한 방법으로 협업필터링 오토인코더 방식을 제안한다. 기존 오토인코더와 달리 감지된 결측치는 학습에 참여시키지 않는 방식을 적용하여 훈련 데이터의 결측치에 더욱 강인한 모델을 학습시킬 수 있다. 시뮬레이션 결과 제안한 협업필터링 오토인코더 기반 알고리즘이 기존 오토인코더와 대비하여 결측치 처리 성능이 매우 우수하게 나타났음을 확인하였다.

Ⅰ. 서론

최근 인공지능 기술이 에너지 최적 관리를 위해서도 널리 사용되고 있으며, 기존 방법보다 더 높은 성능을 보여주고 있다. 인공지능 기술은 센서로부터 측정한 데이터의 품질에 크게 의존하기 때문에 기술 개발에 앞서 센서 데이터의 신뢰성을 확보하는 것이 매우 중요하다. 그러나 작게는 1분 단위로까지 측정해야 하는 센서의 특성상 하드웨어의 결함이나 네트워크 문제와 같은 불안정한 센싱 환경에 노출되기 쉽고 이는 센서 데이터에 이상치나 결측치를 만들어 데이터의 신뢰성을 크게 떨어뜨릴 수 있다. 가상센서는 이러한 실물 센서의 한계를 보완하기 위해 나타난 기술이다. 가상센서는 실물 센서를 설치할 수 없는 환경에서 실물 센서의 역할을 대체할 때 쓰일 수도 있으나, 실물 센서가 있는 환경에서도 보조적 역할을 수행하기 위해 쓰일 수도 있다 [1]. 특히 가상센서를 통한 결측치 처리 기술은 센서의 결함이 있는 환경에서도 센서 데이터의 신뢰성을 높일 수 있다.

이를 위해 최근 오토인코더(Autoencoder) 기반 시스템 단위 가상센싱 방법이 널리 활용되고 있다 [2,3]. 오토인코더는 비지도학습 방법의 일종으로 입력 데이터로부터 가장 중요한 특징(Feature)을 학습할 수 있다. 이때 입력 데이터에 결측치가 들어간 환경에서도 학습한 특징은 이러한 결측치에 매우 강인한 모습을 보인다 [4]. 따라서 특징으로부터 데이터를 복원할 때 데이터에 결측된 부분에 적절한 값이 주입될 수 있다. 디노이징 오토인코더(Denoising Autoencoder)는 입력 데이터에 추가로 결측치를 생성하고 생성한 결측치는 복원될 때 정상값으로 복원되도록 특징을 학습하여 결측치에 더욱 강인해질 수 있다 [5]. 그러나 오토인코더로부터 학습한 특징이 아무리 결측치에 강인하다고 하더라도 훈련 데이터에 결측치가 많은 환경에서는 결측치 위주로 학습이 될 수 있어서 결측치 처리 성능이 급격히 떨어진다.

본 논문에서는 결측치에 더욱 강인한 오토인코더 모델을 개발하기 위해 협업필터링 오토인코더(Collaborative Fitlering Autoencoder)를 제안한다. 협업필터링은 추천시스템에서 많은 사용자들로부터 수집한 정보로부터 사용자들의 기호를 예측하는데 널리 사용되는 기술이다. 이때 사용자가 평점을 준 아이템에 대해서는 오토인코더가 해당 평점을 복원하도록 학습시키고, 평점을 주지 않았다면 그 비어있는 평점을 복원하는 것이 아닌 학습에 참여시키지 않는 방법으로 학습시켰다 [6]. 이 경우 특징은 평점이 매겨진 값으로부터만 학습이 되기 때문에 입력된 데이터에 더욱 적합한 특징이 학습되어질 수 있다. 이 방법을 센서 데이터 결측치 처리에도 적용할 수 있다. 결측치를 학습에 참여시키지 않으면 결측치가 아닌 값으로부터만 특징을 학습할 수 있다. 따라서 특징으로부터 데이터를 복원할 때 결측치에 더욱 적절한 값이 주입되어 효과적인 결측치 처리 결과를 기대할 수 있다.

Ⅱ. 본론

오토인코더는 출력값이 입력값과 동일하도록 학습된다. K차원의 훈련 데이터가 총 N개가 있고, x_{ij}를 i번째 데이터의 j번째 값이라고 하자. 파라미터가 θ인 오토인코더는 아래의 목적함수에 따라 학습이 진행된다.

$$\min_{\theta} \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{K} \left(x_{ij} - \hat{x}_{ij} \right)^2. \quad (1)$$

여기서 소수의 x_{ij}가 결측되어 들어오더라도 (주로 0으로 처리된다) 오토인코더는 이에 강인하게 특징을 학습할 수 있기 때문에 복원된 값 $\hat{x}_{ij}$은 적절한 값이 주입된 상태로 출력될 수 있다.

디노이징 오토인코더는 이러한 결측치 처리 성능을 더 강력하게 할 수 있다. 기존 오토인코더와의 차이는 입력 데이터에 추가적으로 결측치를 넣은 데이터를 입력값으로 넣는다는 점이다. 위의 목적함수에서 일부 $\hat{x}_{ij}$

는 결측치로 된 입력값으로부터 (0으로 처리된 값으로부터) 출력되도록 바뀌게 되지만 복원된 값은 x_{ij}가 되도록 학습이 진행되어 결측치에 더욱 강인해질 수 있다.

협업필터링 오토인코더는 결측치를 학습에 참여시키지 않는 방법으로 목적함수가 변경되어, 아래의 목적함수에 따라 학습이 진행된다.

$$\min_{\theta} \frac{1}{N \times |O_i|} \sum_{i=1}^{N} \sum_{j=1}^{K} (x_{ij} - \hat{x}_{ij})^2 \times I(x_{ij} \in O_i). \quad (2)$$

여기서 O_i는 i번째 데이터의 K개의 변수 $x_{i1}, x_{i2}, \cdots, x_{iK}$ 중에서 결측치가 아닌 값들의 집합이며, $I(x_{ij} \in O_i)$는 $x_{ij} \in O_i$을 만족할 경우 1을, 그렇지 않을 경우 0을 출력한다. 위와 같은 방식으로 결측치를 학습에 참여시키지 않지만 해당 부분이 결측되지 않은 다른 데이터들로부터 그 부분으로 인해 나타나는 특징을 충분히 학습할 수 있는 협업필터링 기반 방법이기 때문에 결측치에 적절한 값이 주입된 채로 복원될 수 있다. 또한 결측치가 많을 경우 기존 오토인코더나 디노이징 오토인코더의 경우 결측치 위주로 학습되어 학습 성능이 저하될 수 있지만 협업필터링 오토인코더는 이러한 결측치가 학습에 참여되지 않으므로 정상값 위주로 학습되어 학습 성능을 저하시키지 않을 수 있다.

Ⅲ. 실험

제안한 협업필터링 오토인코더(Collaborative Filtering Autoencoder; CFAE)의 성능 평가를 위해 기존 오토인코더(Autoencoder; AE) 및 디노이징 오토인코더(Denoising Autoencoder; DAE)와 결측치 처리 성능을 비교하였다. AI 허브에서 제공하는 전력 설비 에너지 품질 데이터 중 열처리 설비 8개의 전력 소비량 데이터를 사용하였다 [7]. 해당 데이터는 2021년 1월 16일부터 2월 4일까지 20일간 1분 단위로 측정되었고, 본 실험에서는 10분 단위로 샘플링하여 8차원의 데이터 2880개를 형성하였다. 각 데이터에서 20%의 비율로 결측치를 생성하였으며, 모든 데이터를 각 설비 전력 소비량의 최대값으로 나누어 0과 1 사이의 값으로 만들었다.

오토인코더의 구조를 결정하기 위해 데이터에 10% 비율의 결측치를 추가로 생성하여 Validation Set으로써 활용하였다. 그 결과 3개의 은닉층에 각각 4개, 2개, 4개의 뉴런이 있는 구조로 구성하였다. 이에 따라 오토인코더가 학습하는 특징은 2차원이 된다. 학습 알고리즘으로 학습률이 0.0001인 Adam을 사용하였고, 과적합 방지를 위해 정규화(Regularization) 파라미터를 1로 설정하였으며, 파이토치를 이용하여 구현하였다. 성능 평가는 생성한 결측치를 얼마나 잘 복원하는지로 확인하였으며, 성능 평가 지표로는 Root Mean Square Error (RMSE)와 Mean Absolute Error (MAE)를 0과 1 사이로 정규화(Normalization)한 NRMSE와 NMAE를 사용하였다.

표 1은 제안한 모델과 비교군 모델들의 NRMSE와 NMAE를 비교한 것이다. 오토인코더 모델이 가장 안 좋은 성능을 보였고, 디노이징 오토인코더 모델은 성능을 조금 향상시켰으며, 제안한 협업필터링 오토인코더 모델이 가장 좋은 성능을 보였다. 이는 기존 모델들이 결측치 비율이 20%로 높은 환경에서 결측치 위주로 학습이 진행되어 학습 성능이 나빠지게 되었기 때문으로 제안한 모델의 결측치를 학습에 참여시키지 않은 것에 대한 효과를 확인할 수 있었다. 그림 1은 결측치 처리 오차의 Probability Density Function (PDF)을 보여주며, 제안한 모델의 그래프가 다른 모델에 비해 낮은 오차에 분포되어 있는 것을 확인할 수 있다. 그림 2는 결측치 처리 오차의 Cumulative Distribution Function (CDF)를 나타내며 제안한 모델의 그래프가 좌측 상단에 위치하여 있어 마찬가지로 낮은 오차에 더 많이 분포되어 있음을 확인할 수 있다.

	AE	DAE	CFAE
NRMSE	0.0649	0.0593	0.0477
NMAE	0.0506	0.0482	0.0351

표 1 결측치 처리 성능 비교

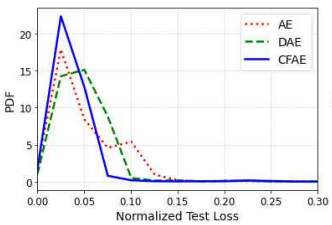

그림 1 결측치 처리 오차의 PDF

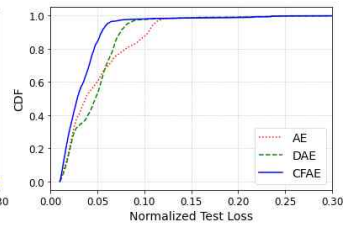

그림 2 결측치 처리 오차의 CDF

Ⅳ. 결론

본 논문에서는 센서 데이터의 결측치를 처리하기 위해 협업필터링 오토인코더를 제안하였다. 협업필터링 오토인코더는 감지된 결측치를 학습에 참여시키지 않음으로써 훈련 데이터에 결측치가 많은 환경에서도 보다 강인한 모델을 학습할 수 있다. 시뮬레이션 결과 기존의 오토인코더 또는 디노이징 오토인코더 방법보다 결측치 처리에서 더 뛰어난 성능을 보이는 것을 확인하였다.

향후 연구에서는 결측치 처리뿐만 아니라 이상치까지 탐지하고 처리하는 연구를 진행할 필요가 있다. 또한 Transformer와 같이 시계열 데이터 처리에 용이한 모델 또는 Variational Autoencoder와 같이 확률적으로 이상치를 탐지할 수 있는 알고리즘을 적용하면 이상치 탐지 및 처리 성능을 더욱 향상시킬 수 있다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. 20212020090028C)

참 고 문 헌

[1] 윤성민. (2020). 건물에너지시스템의 가상센싱기술과 커미셔닝 자동화 방안. 건축환경설비, 14(3), 31-40.

[2] 홍예진, 최영웅, 윤성민, & 김용식. (2020). 오토인코더 기반 건축설비시스템의 가상센싱 방법. 대한설비공학회 학술발표대회논문집, 336-339.

[3] 고홍철, 석규한, 이정한, 박종현, & 김선우. (2022). 가상센서를 이용한 제철소 가열로의 연소가스 분석기 고장 감시 및 백업 기술개발. 제어로봇시스템학회 논문지, 28(8), 708-713.

[4] Park, K., Jeong, J., Kim, D., & Kim, H. (2020). Missing-insensitive short-term load forecasting leveraging autoencoder and LSTM. IEEE Access, 8, 206039-206048.

[5] Ryu, S., Kim, M., & Kim, H. (2020). Denoising autoencoder-based missing value imputation for smart meters. IEEE Access, 8, 40656-40666.

[6] Sedhain, S., Menon, A. K., Sanner, S., & Xie, L. (2015, May). Autorec: Autoencoders meet collaborative filtering. In Proceedings of the 24th international conference on World Wide Web (pp. 111-112).

[7] AI 허브 https://aihub.or.kr/