데이터 불균형 문제를 해결하기 위한 Generative Adversarial Networks (GAN) 기반의 가상 데이터 합성

Data Generation with Generative Adversarial Networks to Solve Data Imbalance Problem in Manufacturing

저자 (Authors)
이호진, 박승태, 박범수, 정해동, 이승철
Hojin Lee, Seungtae Park, Bumsoo Park, Haedong Jeong, Seungchul Lee

출처 (Source)
대한기계학회 춘추학술대회 , 2017.11, 1142-1143(2 pages)

발행처 (Publisher)
대한기계학회
The Korean Society of Mechanical Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07287680

APA Style
이호진, 박승태, 박범수, 정해동, 이승철 (2017). 데이터 불균형 문제를 해결하기 위한 Generative Adversarial Networks (GAN) 기반의 가상 데이터 합성. 대한기계학회 춘추학술대회, 1142-1143

데이터 불균형 문제를 해결하기 위한 Generative Adversarial Networks (GAN) 기반의 가상 데이터 합성

이호진† · 박승태 · 박범수 · 정해동 · 이승철

Data Generation with Generative Adversarial Networks to Solve Data Imbalance Problem in Manufacturing

Hojin Lee†, Seungtae Park, Bumsoo Park, Haedong Jeong, and Seungchul Lee

Key Words: Generative Adversarial Networks (GAN), Data Imbalance (데이터 불균형), Data Generation (데이터 생성), Deep Learning (딥러닝), Machine Learning (기계학습).

Abstract

With the development of manufacturing technology and the introduction of the smart factory, the productivity and the amount of data to be acquired has significantly increased Recently, deep learning has been used to replace the existing machine learning algorithms with improved predictive performances. With sufficient amount of data, solving classification problems using the deep learning model can approximate any underlying probability. However, a typical class imbalance in manufacturing field harms performance of the deep learning models. Although methods such as resampling or reweighting data have been proposed to solve the data imbalance issue, such methods are not considered as a fundamental solution to the issue. In this paper, we propose the Generative Adversarial Networks (GAN) which can learn the distributions of a real data set, and generate a virtual dataset that will be utilized for numerous manufacturing applications. To verify the proposed approach, we compared the performances of a model trained with the imbalanced dataset and another model trained with data generated by GAN.

기호설명

G : Generator
D : Discriminator

1. 서 론

스마트 팩토리의 등장 이후, 생산성이 향상됨과 함께 취득되는 데이터의 양이 크게 증가하고 있다. 설비 데이터를 이용한 유지 보수 방법론들이 특히 중요해졌으며, 특히 기계학습 알고리즘을 통한 설비의 상태 진단이 주로 이용되고 있다.

최근에는 딥러닝 알고리즘이 기계학습 알고리즘을 대체하고 있는 추세이다. 이론적으로 딥러닝 알고리즘은 충분한 양의 데이터가 있을 때, 기계학습 알고리즘보다 뛰어난 성능을 보인다. 하지만 딥러닝 알고리즘은 데이터 불균형(특정 라벨에 대한 데이터가 고르게 분포하지 않은 경우, 예를 들어 정상 데이터 99%, 고장 데이터 1%)이 심할 경우 성능이 크게 떨어지는 문제가 있다. 실제 현장에서 취득되는 데이터 중 상당수가 이런 데이터 불균형 문제를 가지고 있다.

일반적으로 데이터 불균형 문제를 해결하기 위해서 사용되는 방법으로는 Resampling (부족한 데이터를 여러 번 학습하는 방법) 등이 있으나, 적은 데이터에 대해 파라미터를 과도하게 학습할 경우, 학습된 모델이 새로운 데이터를 분석하는 데 성능이 크게 떨어지게 된다 (Overfitting).

따라서 본 논문에서는 데이터를 생성하는 딥러닝 알고리즘인 GAN 을 이용해 부족한 클래스의 데이터를 합성한 뒤, 그것을 학습 데이터로 사용해, 데이터 불균형을 해결하는 방법을 제시한다.

†회원, 논문발표자의 소속
E-mail : hojin12312@unist.ac kr
TEL : (010) 9366-0718
* UNIST

2. 데이터 취득 및 알고리즘 구축

GAN 기반 알고리즘으로 데이터 불균형을 해

결하고 그 유효성을 검증하기 위해, 데이터 취득과 알고리즘 구축으로 범위를 나눠 연구를 진행했다.

2.1 데이터 취득

본 논문에서는 시그널링크사의 회전체 테스트 베드에 변위 센서 및 가속도 센서를 부착하여 데이터를 취득한다 (Figure 1). 테스트 베드에서 정상, 불평형 1, 2, 3, 축정렬 불량 1, 2, rubbing 1, 2, 3, 4, 총 10 개 클래스의 데이터를 취득했다(1).

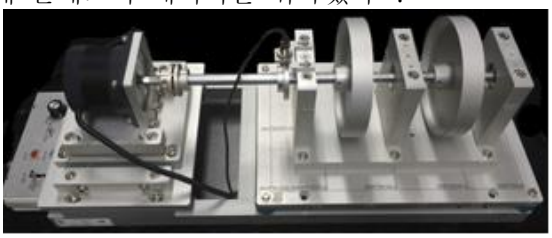

Figure 1 Signallink rotor kit

2.2 Generative Adversarial Networks (GAN)

GAN 은 데이터 생성을 목적으로 하는 딥러닝 알고리즘이다. Figure 2 와 같이, 모델 구조는 Generator 와 Discriminator 로 구성되어 있다. Generator 는 임의의 Z (Latent)를 입력값으로 받아 데이터를 합성하고, Discriminator 는 데이터를 입력값으로 받아 그 값이 Generator 가 생성한 데이터인지 실제 데이터인지 구분하는 역할을 한다.

따라서 충분한 학습을 거치면 Generator 는 Discriminator 가 구분하기 어렵게 실제 데이터와 가까운 데이터를 생성해내게 된다(2).

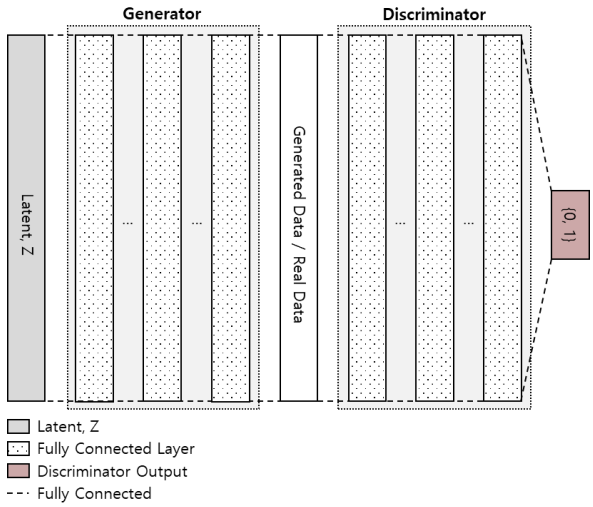

Figure 2 GAN Structure

3. 알고리즘 검증

취득한 데이터를 불균형이 존재하는 10 개의 클래스로 분리하여 각 클래스에 해당하는 가상의 데이터를 GAN 을 이용해 생성하였다. 생성된 가상의 데이터만을 이용하여 인공신경망을 학습하고 그 성능을 확인하였다. 검증을 위하여 전체 데이터의 80%를 사용하여 GAN 을 학습하였으며 남은 20%를 이용하여 성능을 검증하였다. 또한 비교 검증을 위하여 데이터 불균형이 있는 데이터만으로 인공신경망을 학습하였다(3).

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

4. 결 론

딥러닝 알고리즘은 높은 성능을 보여 최근 제조업 데이터에 적용하려는 시도가 증가하고 있다. 하지만 제조업 데이터 특성상 불균형 문제가 심각하기 때문에 딥러닝 활용 및 적용에 있어 성능 저하 문제가 지속적으로 이야기되었다. 본 논문에서는 딥러닝 알고리즘 중 GAN 을 사용하여 부족한 클래스의 데이터를 학습시킨 후 가상의 데이터를 합성하여 균형 데이터를 만들었다. 생성된 균형 데이터로 딥러닝 알고리즘을 학습하고 분류 성능을 검증함으로써 GAN 을 이용한 가상 데이터 합성의 유효성을 보였다.

후 기

본 연구는 미래창조과학부 지역신산업선도인력양성사업 (2016H1D5A1910285), 미래창조과학부 신산업창출을 위한 SW 융합기술고도화 기술개발사업 (S0177-16-1002), 중소기업청 중소기업기술혁신 개발 사업(S2439592) 의 지원을 받아 수행하였습니다.

참고문헌

(1) Jeong, H. D., Kim, S. H., Woo, S. H., Kim, S. H., Lee. S. C., 2017, “Real-time Monitoring System for Rotating Machinery with IoT-based Cloud Platform,” *Transactions of the Korean Society of Mechanical Engineers - A*, Vol. 41, No.5, pp. 137-145.

(2) Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y., 2014, “Generative adversarial nets,” *In Advances in neural information processing systems*, pp. 2672-2680.

(3) Schmidhuber, J. (2015). “Deep learning in neural networks: An overview,” *Neural networks*, *61*, pp. 85-117.