GAN을 이용한 데이터 불균형 문제에 관한 연구

A Study on Data Imbalance Problem Using GAN(Generative Adversarial Network)

저자 (Authors)
정성욱, 김이슬, 김일곤, 임경식
Jung Sung Wook, Kim I Seul, Kim Il Kon, Lim Kyung Shik

출처 (Source)
한국통신학회 학술대회논문집 , 2019.1, 1390-1391(2 pages)
Proceedings of Symposium of the Korean Institute of communications and Information Sciences , 2019.1, 1390-1391(2 pages)

발행처 (Publisher)
한국통신학회
Korea Institute Of Communication Sciences

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08003853

APA Style
정성욱, 김이슬, 김일곤, 임경식 (2019). GAN을 이용한 데이터 불균형 문제에 관한 연구. 한국통신학회 학술대회논문집, 1390-1391

GAN을 이용한 데이터 불균형 문제에 관한 연구

정성욱*, 김이슬, 김일곤, 임경식

*경북대학교

*squirrelswj@gmail.com, dltmf0689@gmail.com, ikkimgg@gmail.com, kslim@knu.ac.kr

A Study on Data Imbalance Problem Using GAN(Generative Adversarial Network)

Jung Sung Wook*, Kim I Seul, Kim Il Kon, Lim Kyung Shik

*Department of Computer Science & Engineering KyungPook National University.

요 약

본 논문은딥러닝 기술이 크게 성장하면서 방대한 양의 데이터를 활용하며 분류 및 예측에 대한 연구가 진행 중이다. 딥러닝은 충분한 양의 데이터를 학습시킬 때 기계학습 알고리즘보다 뛰어난 성능을 보이지만 학습 데이터의 각 범주간의 비율이 불균등하다면 성능이 크게 떨어지는 문제가 있다. 이 문제를 완화하는 방법으로 Under-Sampling 및 Over-Sampling 등이 있지만 이는 임의로 데이터를 제거하여 데이터의 손실을 일으키거나 Overfitting이 일어 날 수 있는 단점이 있다. 본 논문에서 제안하는 적대적 생성 인공신경망 샘플링법은 기존 데이터와 흡사한 새로운 가상의 데이터를 생성해 냄으로써 이러한 문제점을 완화하였다. 제안하는 기법을 94:6 비율의 범주를 갖는 당뇨관련 불균형 데이터에 적용하여 그 효용성을 입증한다.

Ⅰ. 서 론

데이터양의 증가와 딥러닝 기술이 크게 성장 하면서 데이터 분류 및 예측에 대한 많은 연구 활동이 이루어지고 있다. 딥러닝은 충분한 양의 데이터가 있을 때 기계학습 알고리즘보다 뛰어난 성능을 보이지만 데이터를 학습하는데 있어서 데이터 불균형(특정 라벨에 대한 데이터의 양이 고르지 못한 경우)이 심할 경우 성능이 크게 떨어지는 문제가 있다[1]. 데이터 불균형 문제를 완화시키기 위한 방법으로 여러가지 Resampling기법이 있다. 보편적인 Resampling방식은 현재 갖고 있는 데이터의 일부분을 재추출하는 방식이지만 본 논문에서 비교 관측할 적대적 생성 인공신경망(Ganerative Adversarial Network)기법은 데이터를 재추출 하는 것이 아니라 그와 흡사한 새로운 가상의 데이터를 만드는 방식이다. 본 논문에서는 GAN을 사용하여 국민건강보험에서 제공하는 당뇨 관련 불균형 데이터[2]문제를 완화시키는 방법을 제안한다.

Ⅱ. 본론

1. Resampling

일반적으로 데이터 불균형 문제를 완화시키기 위해 사용되는 Resampling 방법으로는 Under-Sampling, Over- Sampling, SMOTE 등이 있다. Under-Sampling은 다수범주의 관측치를 임의로 비복원추출하여 소수범주와의 비율을 조정하는 방식이다[3]. 이는 데이터가 임의로 제거되는 단점이 있다. Over-Sampling은 소수범주의 관측치를 임의로 복원추출하여 다수범주와의 비율을 조정하는 방식이다[4]. 이는 데이터의 손실은 없지만 소수범주를 복원추출함으로 인해 Overfitting 이 일어날 수 있다는 단점이 있다. SMOTE는 소수범주의 데이터의 최근접 데이터간의 차이에 0~1사이의 임의의 값을 곱하여 데이터에 추가 시키는 방식이다[5]. Under-Sampling의 데이터 손실과 Over-sampling의 Overfitting의 문제를 완화시켰지만 실제 데이터와 관련 없는 잠재적 노이즈 데이터를 생성하게 될 수 있다.

2 GAN(Generative Adversarial Network)

본 논문에서 사용할 GAN은 생성자(G, Generator)모델과 판별자(D, Discriminator)모델이 서로 적대적으로 경쟁하며 학습하여 학습 성능을 높이는 비지도 학습 기반의 학습 모델이다[6].

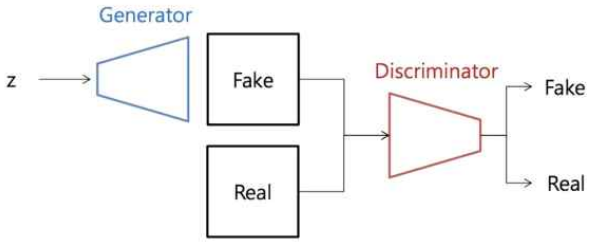

그림 1. GAN 개념도

G는 원본 데이터와 유사한 데이터를 생성 하는 모델이고 D는 원본 데이터와 생성된 데이터를 판별하는 모델이다. 학습이 진행 될수록 G는 D가 판별하기 어렵도록 원본 데이터와 유사한 데이터를 생성해 낸다. 충분한 학습을 거친 후에는 D가 원본 데이터와 가까운 데이터를 구분하기 어려울 정도로 생성해내게 된다. 이러한 방식은 데이터의 손실 및 Overfitting도 방지하며 원본 데이터와 유사한 데이터를 생성하는 것 이므로 잠재적 노이즈 데이터 생성문제도 해결할 수 있다.

Ⅲ. 실험 및 결과

1. 실험 환경

딥러닝 프레임워크로는 Keras를 사용하였다. Keras는 Python언어를 사용하며 수식 및 행렬 연산을 쉽게 만들어주는 Theano를 기반으로 이루

어져 있다. Resampling IDE로는 RStudio를 사용하였다. RStudio는 통계 계산과 수치해석 기법을 지원하는 R언어를 기반으로 하는 통합개발환경이다.

2. 데이터 셋

데이터로 사용한 데이터는 국민건강보험에서 제공하는 당뇨 관련 데이터 78만 3천명 중 당뇨에 걸리지 않은 사람(N) 74만명, 당뇨에 걸린 사람(D) 4만3천명 으로 두 범주의 데이터 비율은 N : D = 94.19 : 5.81 로 데이터 분포가 불균형 하게 구성되어 있다. 데이터셋은 성, 나이, 수축압력, 이완압력, 공복혈당, 체질량, 당뇨로 이루어져 있다. 데이터 분석을 위한 변수내역으로 남녀는 1, 2로 구분하였고 나이는 1~27로 20세부터 75세까지 분류하였다. 당뇨 유무는 각각 1, 0으로 구분하였다.

	sex	age	syspressure	diapressure	glucose	bodymass	diabetes
1	1	3	138	88	95	29.7	0
2	1	4	110	78	82	21.9	0
3	1	15	118	78	122	26.9	1
4	1	12	112	67	90	27.9	0
5	1	12	119	74	100	24.6	0
6	2	18	130	80	90	24.5	0
7	2	1	115	63	88	22.6	0
8	1	15	130	80	95	23.7	0
9	2	19	122	70	97	24.6	0
10	2	21	120	70	82	24.2	0
11	1	18	100	60	93	23.9	0
12	1	4	120	78	103	24.1	0
13	1	19	111	59	92	20.8	0

그림 2. diabetes_data.csv 예시

먼저 다수범주의 관측치를 임의로 비복원추출하여 소수범주와의 비율을 조정하는 Under-Sampling의 데이터셋이다. D의 데이터양은 43000 이므로 다수범주인 N에서 랜덤으로 43000개를 뽑아 50 : 50 비율로 맞추어 학습을 진행하였고 학습 딥러닝 모델은 MLP(Multi Layer Perceptron)을 사용하였다. 86000개의 데이터중 70%는 학습에, 30%는 테스트에 사용하였다.

다음으로 소수범주의 데이터를 임의로 복원추출하는 Over-Sampling의 데이터셋이다. N의 데이터양은 740000 이므로 소수범주인 D의 데이터를 랜덤으로 740000개 복원추출하여 50: 50 비율로 맞추었다. 학습 딥러닝 모델은 역시 MLP이며 1480000개의 데이터중 70%는 학습에, 30%는 테스트에 사용하였다.

마지막으로 소수범주의 데이터로 GAN학습을 반복하여가상의 데이터를 697000개 생성하였다. GAN구조는 [6]과 비슷하며 총 1480000개의데이터 중 70%는 학습에, 30%는 테스트에 사용하였다.

Ⅳ. 결 론

본 논문에서는 94:6의 당뇨관련 불균형 데이터로 Under-Sampling, Over-Sampling, GAN-Sampling을 비교 분석 해 보았다. 그 결과, GAN-Sampling은 Under-Sampling에 비해 정확도는 낮았고 Over-Sampling에 비해 중복된 데이터 없이 상대적으로 높은 정확도를 보였으며 또한 Overfitting 횟수가 가장 적음을 알 수 있었다. 하지만 정확도 만으로는 성능을 측정하는 Metric으로 부적절하며 각 Sampling의 올바른 평가를 위해서는 ROC Curve를 통한 평가가 필요하다고 판단하고 진행중이다.

ACKNOWLEDGMENT

"본 논문은 과학기술정보통신부 및 정보통신기술진흥센터의 SW 중심대학지원사업(2015-0-00912) 및 교육부와 한국연구재단의 BK21 플러스 사업 (경북대학교 컴퓨터학부 Smart Life 실현을 위한 SW 인력양성사업단)으로 지원된 연구임(21A20131600005)"

참 고 문 헌

[1] S. Barua, Md. M. Islam, X. Yao, "MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning," IEEE Transaction on Knowledgement and Data Engineering, vol. 26, no. 2, p. 405-424, 2014.

[2] https://nhiss.nhis.or.kr/bd/ab/bdabf003cv.do

[3] Wang, D., and M. Shi, "Density Weighted Region Growing Method for Imbalanced Data SVM Classification in Under-sampling Approaches," Journal of Information & Computational Science, Vol.11, No.18(2014), 6673~6680.

[4] Kubat, M. and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 179-186, Nashville, Tennesse. Morgan Kaufmann.

[5] Chawla, N. V., Boyer, K. W., Hall, L. O. and Kegelmeyer, W. P.(2002. SMOTE:Synthetic Minority oversampling Technique, Journal of Artificial Intelligence Research, 16, 321-357.

[6] I. J. Goodfellow, J. Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in Preceedings of the Neural Information Processing systems, pp.2672-2680,2014.