2장. 이진 판단: 천체의 펄서 여부 판정 신경망

부산대학교 항공우주공학과

정대현 201527137

1. 개요

해당 실험에서는 데이터의 불균형을 내포하고 있을 때 문제점을 탐구하며 해결방안을 학습하고 있습니다. 이진 판단을 하는 문제에서 데이터가 불균형한 분포를 가지는 경우, 모델의 학습이 제대로 이루어지지 않을 확률이 높습니다. 이번 실험에서는 천체에서 펄서는 적은 비율로 존재하며 이들 모두 펄서가 아니라고 분류하면 펄서는 하나도 찾지 못하였지만 전체 정확도는 높게 나오게 됩니다. 비슷한 예로 금융 이상거래 탐지처럼 적은 빈도로 나타나는 문제가 중요한 경우가 있습니다. 이 문제를 해결하기 위해 다음과 같은 두가지 방법을 사용합니다.

2. 해결 방안

첫번째 방법은 데이터의 비율 조정입니다. 많은 쪽의 데이터를 줄여서 수를 맞추는 것을 언더샘플링, 적은 수의 데이터를 반복적으로 뽑는 것을 오버샘플링라고 합니다. 언더샘플링의 경우 중요한 데이터 손실 우려가 있으며 오버샘플링의 경우 과대적합 문제가 있습니다. 이번 실험에서는 펄스 데이터를 중복 사용하며 해당 샘플에 약간의 노이즈를 추가하는 방식으로 문제를 해소하고자 합니다.

두번째는 새로운 평가 지표의 도입입니다. 정밀도는 모델의 관점에서 신경망이 참으로 추정한 것 중 실제 참인 것을, 재현율은 실제 정답의 관점에서 참인 것 가운데 모델이 참으로 분류한 비율을 뜻합니다. 참이라고 많이 답할 수록 재현율을 높일 수 있고 거짓이라고 많이 답할수록 정밀도가 상승합니다. 하지만 어느 한 수치가 중요한 것이 아닌 모두 높은 값을 유지하는 것이 유리하기에 F1값(식1)을 도입하게 됩니다.

$$F1 = \frac{2}{recall^{-1}+precision^{-1}} = 2 * \frac{precison*recall}{precisoin+reacl} = \frac{tp}{tp+\frac{1}{2}(fp+fn)} \quad (식1)$$

3. 파라미터의 의미

배치사이즈(Batch size): 전체 학습 데이터 셋을 여러 작은 그룹을 나누었을 때 배치사이즈는 하나의 소그룹에 속하는 데이터 수를 의미합니다. 전체 트레이닝 셋을 작게 나누는 이유는 트레이닝 데이터를 통째로 신경망에 넣으면 비효율적이 리소스 사용으로 학습 시간이 오래 걸리기 때문입니다.

Epoch: 전체 트레이닝 셋이 신경망을 통과한 횟수 의미합니다. 예를 들어, 1-epoch는 전체 트레이닝 셋이 하나의 신경망에 적용되어 순전파와 역전파를 통해 신경망을 한 번 통과했다는 것을 의미합니다.

4. 파라미터별 학습결과

조건1) 원데이터 사용 epoch변경

4.1 모델 1(원데이터 사용)

구분	내용
조건	- 원데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 10
정확도	0.954
F1-Score	0.680
결과 화면	Epoch 1: loss=13.463, result=0.908,0.508,0.826,0.629 Epoch 2: loss=11.755, result=0.969,0.871,0.794,0.830 Epoch 3: loss=12.369, result=0.970,0.897,0.773,0.830 Epoch 4: loss=12.360, result=0.778,0.294,0.962,0.450 Epoch 5: loss=13.148, result=0.970,0.926,0.743,0.825 Epoch 6: loss=13.334, result=0.965,0.912,0.699,0.791 Epoch 7: loss=12.495, result=0.968,0.963,0.684,0.800 Epoch 8: loss=12.092, result=0.950,0.988,0.481,0.647 Epoch 9: loss=11.875, result=0.970,0.908,0.761,0.828 Epoch 10: loss=12.189, result=0.954,0.983,0.519,0.680 Final Test: final result = 0.954,0.983,0.519,0.680

4.2 모델 2(원데이터 사용)

구분	내용
조건	- 원데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 50
정확도	0.972
F1-Score	0.834
결과 화면	Epoch 5: loss=11.890, result=0.974,0.885,0.836,0.860 Epoch 10: loss=10.459, result=0.973,0.879,0.833,0.855 Epoch 15: loss=10.171, result=0.972,0.954,0.736,0.831 Epoch 20: loss=10.217, result=0.961,0.752,0.883,0.812 Epoch 25: loss=10.792, result=0.966,0.970,0.660,0.785 Epoch 30: loss=9.717, result=0.975,0.928,0.795,0.856 Epoch 35: loss=11.532, result=0.972,0.866,0.836,0.851 Epoch 40: loss=11.628, result=0.974,0.884,0.830,0.856 Epoch 45: loss=11.198, result=0.973,0.865,0.845,0.855 Epoch 50: loss=11.791, result=0.972,0.951,0.742,0.834 Final Test: final result = 0.972,0.951,0.742,0.834

4.3 모델 3(원데이터 사용)

구분	내용
조건	- 원데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 100
정확도	0.952
F1-Score	0.759
결과 화면	Epoch 10: loss=11.225, result=0.971,0.981,0.673,0.798 Epoch 20: loss=12.231, result=0.976,0.929,0.776,0.845 Epoch 30: loss=11.499, result=0.974,0.865,0.825,0.845 Epoch 40: loss=12.027, result=0.972,0.860,0.792,0.825 Epoch 50: loss=10.837, result=0.977,0.955,0.762,0.848 Epoch 60: loss=12.035, result=0.974,0.945,0.739,0.830 Epoch 70: loss=12.393, result=0.973,0.881,0.785,0.831 Epoch 80: loss=11.343, result=0.971,0.811,0.861,0.835 Epoch 90: loss=12.498, result=0.972,0.837,0.828,0.833 Epoch 100: loss=11.216, result=0.952,0.657,0.898,0.759 Final Test: final result = 0.952,0.657,0.898,0.759

조건2) 펄스 중복 데이터 사용, epoch변경

4.4 모델 4(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 10
정확도	0.926
F1-Score	0.921
결과 화면	Epoch 1: loss=41.066, result=0.930,0.970,0.886,0.926 Epoch 2: loss=35.770, result=0.882,0.994,0.767,0.865 Epoch 3: loss=35.326, result=0.920,0.918,0.921,0.919 Epoch 4: loss=36.711, result=0.895,0.996,0.792,0.882 Epoch 5: loss=35.160, result=0.923,0.981,0.863,0.918 Epoch 6: loss=34.956, result=0.916,0.918,0.913,0.916 Epoch 7: loss=34.506, result=0.912,0.982,0.838,0.904 Epoch 8: loss=36.573, result=0.905,0.981,0.825,0.896 Epoch 9: loss=35.877, result=0.922,0.932,0.909,0.920 Epoch 10: loss=35.139, result=0.926,0.979,0.870,0.921

4.5 모델 5(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 50
정확도	0.934
F1-Score	0.932
결과 화면	Epoch 5: loss=37.027, result=0.917,0.908,0.928,0.918 Epoch 10: loss=33.819, result=0.928,0.944,0.910,0.927 Epoch 15: loss=37.238, result=0.927,0.972,0.879,0.923 Epoch 20: loss=35.487, result=0.849,0.786,0.960,0.864 Epoch 25: loss=34.084, result=0.913,0.987,0.837,0.906 Epoch 30: loss=36.157, result=0.921,0.984,0.856,0.916 Epoch 35: loss=33.999, result=0.927,0.958,0.892,0.924 Epoch 40: loss=34.858, result=0.606,0.561,0.982,0.714 Epoch 45: loss=35.703, result=0.773,0.702,0.947,0.806 Epoch 50: loss=34.569, result=0.934,0.954,0.911,0.932 Final Test: final result = 0.934,0.954,0.911,0.932

4.6 모델 6(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 100
정확도	0.924
F1-Score	0.918
결과 화면	Epoch 10: loss=37.447, result=0.906,0.885,0.931,0.907 Epoch 20: loss=34.470, result=0.820,0.752,0.948,0.838 Epoch 30: loss=33.347, result=0.780,0.696,0.981,0.814 Epoch 40: loss=35.034, result=0.923,0.931,0.911,0.921 Epoch 50: loss=34.904, result=0.907,0.883,0.936,0.909 Epoch 60: loss=34.848, result=0.930,0.951,0.904,0.927 Epoch 70: loss=32.286, result=0.935,0.966,0.899,0.931 Epoch 80: loss=31.634, result=0.910,0.891,0.931,0.911 Epoch 90: loss=32.679, result=0.868,0.997,0.735,0.846 Epoch 100: loss=34.402, result=0.924,0.976,0.867,0.918 Final Test: final result = 0.924,0.976,0.867,0.918

4.7 모델 7(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 500
정확도	0.939
F1-Score	0.938
결과 화면	Epoch 50: loss=37.082, result=0.898,0.994,0.802,0.887 Epoch 100: loss=34.339, result=0.899,0.876,0.931,0.903 Epoch 150: loss=30.610, result=0.910,0.892,0.934,0.912 Epoch 200: loss=32.828, result=0.915,0.989,0.840,0.909 Epoch 250: loss=32.053, result=0.892,0.850,0.953,0.899 Epoch 300: loss=32.746, result=0.936,0.967,0.903,0.934 Epoch 350: loss=31.970, result=0.916,0.897,0.940,0.918 Epoch 400: loss=30.596, result=0.923,0.924,0.922,0.923 Epoch 450: loss=31.277, result=0.919,0.909,0.932,0.920 Epoch 500: loss=29.638, result=0.939,0.966,0.911,0.938` Final Test: final result = 0.939,0.966,0.911,0.938

4.8 모델 8(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 750
정확도	0.933
F1-Score	0.931
결과 화면	Epoch 75: loss=32.195, result=0.916,0.973,0.855,0.910 Epoch 150: loss=32.300, result=0.909,0.986,0.828,0.900 Epoch 225: loss=31.017, result=0.847,0.779,0.966,0.862 Epoch 300: loss=30.114, result=0.935,0.963,0.903,0.932 Epoch 375: loss=30.920, result=0.933,0.953,0.910,0.931 Epoch 450: loss=29.373, result=0.934,0.978,0.886,0.930 Epoch 525: loss=30.474, result=0.919,0.987,0.847,0.912 Epoch 600: loss=29.512, result=0.933,0.945,0.918,0.931 Epoch 675: loss=29.895, result=0.919,0.907,0.931,0.919 Epoch 750: loss=29.262, result=0.933,0.954,0.909,0.931 Final Test: final result = 0.933,0.954,0.909,0.931

4.9 모델 9(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 10 - 학습량(epoch) : 1000
정확도	0.921
F1-Score	0.922
결과 화면	Epoch 100: loss=31.770, result=0.571,0.540,0.997,0.701 Epoch 200: loss=30.952, result=0.933,0.964,0.901,0.932 Epoch 300: loss=31.110, result=0.939,0.959,0.918,0.938 Epoch 400: loss=31.575, result=0.934,0.985,0.882,0.931 Epoch 500: loss=30.052, result=0.915,0.908,0.924,0.916 Epoch 600: loss=29.070, result=0.890,0.859,0.935,0.895 Epoch 700: loss=28.448, result=0.937,0.977,0.896,0.934 Epoch 800: loss=29.215, result=0.915,0.992,0.839,0.909 Epoch 900: loss=27.420, result=0.921,0.991,0.851,0.916 Epoch 1000: loss=27.833, result=0.921,0.908,0.937,0.922 Final Test: final result = 0.921,0.908,0.937,0.922

조건3) epoch는 500으로 고정, 매개변수 인수값(mb_size) 변경

4.10 모델 10(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 3 - 학습량(epoch) : 500
정확도	0.911
F1-Score	0.915
결과 화면	Epoch 50: loss=72.148, result=0.928,0.933,0.924,0.928 Epoch 100: loss=72.262, result=0.909,0.886,0.939,0.912 Epoch 150: loss=70.896, result=0.800,0.725,0.969,0.829 Epoch 200: loss=62.221, result=0.928,0.948,0.906,0.926 Epoch 250: loss=64.295, result=0.921,0.968,0.872,0.917 Epoch 300: loss=65.970, result=0.939,0.953,0.925,0.939 Epoch 350: loss=69.092, result=0.932,0.990,0.874,0.928 Epoch 400: loss=69.017, result=0.928,0.936,0.920,0.928 Epoch 450: loss=65.741, result=0.916,0.899,0.939,0.918 Epoch 500: loss=69.304, result=0.911,0.878,0.955,0.915 Final Test: final result = 0.911,0.878,0.955,0.915 time : 272.85s

4.11 모델 11(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 5 - 학습량(epoch) : 500
정확도	0.942
F1-Score	0.941
결과 화면	Epoch 50: loss=53.650, result=0.904,0.991,0.816,0.895 Epoch 100: loss=49.234, result=0.933,0.941,0.926,0.933 Epoch 150: loss=47.092, result=0.902,0.869,0.947,0.907 Epoch 200: loss=49.637, result=0.924,0.986,0.861,0.919 Epoch 250: loss=44.966, result=0.937,0.954,0.918,0.936 Epoch 300: loss=49.608, result=0.921,0.940,0.900,0.920 Epoch 350: loss=47.401, result=0.912,0.901,0.926,0.914 Epoch 400: loss=48.050, result=0.928,0.985,0.870,0.924 Epoch 450: loss=45.344, result=0.871,0.834,0.928,0.878 Epoch 500: loss=44.703, result=0.942,0.969,0.915,0.941 Final Test: final result = 0.942,0.969,0.915,0.941 time : 176.30s

4.12 모델 12(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 20 - 학습량(epoch) : 500
정확도	0.907
F1-Score	0.899
결과 화면	Epoch 50: loss=25.077, result=0.924,0.971,0.876,0.921 Epoch 100: loss=24.383, result=0.926,0.935,0.915,0.925 Epoch 150: loss=26.054, result=0.912,0.911,0.914,0.913 Epoch 200: loss=24.166, result=0.783,0.704,0.979,0.819 Epoch 250: loss=24.662, result=0.932,0.971,0.891,0.929 Epoch 300: loss=24.302, result=0.907,0.991,0.822,0.898 Epoch 350: loss=22.448, result=0.916,0.931,0.899,0.915 Epoch 400: loss=21.860, result=0.931,0.974,0.887,0.928 Epoch 450: loss=22.600, result=0.816,0.746,0.962,0.840 Epoch 500: loss=22.597, result=0.907,0.992,0.821,0.899 Final Test: final result = 0.907,0.992,0.821,0.899 time : 42.38s

4.13 모델 13(펄스 중복 데이터 사용)

구분	내용
조건	- 펄스 중복 데이터 사용 - 매개변수 인수값(mb_size) : 30 - 학습량(epoch) : 500
정확도	0.923
F1-Score	0.921
결과 화면	Epoch 50: loss=23.371, result=0.917,0.978,0.849,0.909 Epoch 100: loss=22.402, result=0.921,0.983,0.851,0.912 Epoch 150: loss=22.615, result=0.928,0.961,0.887,0.923 Epoch 200: loss=20.884, result=0.915,0.899,0.928,0.914 Epoch 250: loss=21.145, result=0.907,0.884,0.930,0.907 Epoch 300: loss=22.463, result=0.890,0.841,0.954,0.894 Epoch 350: loss=20.222, result=0.846,0.776,0.957,0.857 Epoch 400: loss=21.419, result=0.934,0.973,0.888,0.929 Epoch 450: loss=19.929, result=0.909,0.994,0.817,0.897 Epoch 500: loss=20.711, result=0.923,0.912,0.931,0.921 Final Test: final result = 0.923,0.912,0.931,0.921 time : 28.73s

4. 결론

구분	1	2	3	4	5	6	7
정확도	0.954	0.972	0.952	0.926	0.934	0.924	0.939
F1	0.680	0.834	0.759	0.921	0.932	0.918	0.938
순위(F1)	12	10	11	6	3	7	2
구분	8	9	10	11	12	13	
정확도	0.933	0.921	0.911	0.942	0.907	0.923	
F1	0.931	0.922	0.915	0.941	0.899	0.921	
순위(F1)	4	5	9	1	9	6	

표1 모델 별 정확도, F1-score

표1과 같이 총 12번의 실험과 3가지 다른 조건을 가정으로 실험을 진행했습니다. 모델 1부터 3까지는 편향된 데이터 수정없이 그대로 사용하였을 때 데이터의 쏠림으로 인해서 정확도는 90%이상 나왔으나 F1-score는 정확도에 비해서 크게 낮은 값을 보여주고 있습니다. 모델 4번부터는 펄스 중복 데이터를 추가해서 사용하자 F1-score가 90%이상으로 개선된 것을 보여주고 있습니다. 모델 4부터 9까지는 펄스중복데이터를 사용한 상태에서 epoch를 10부터 1000까지 여러 값을 주었을 때 F1-score와 정확도는 비례해서 개선되지 않았으며 때로는 저하되는 모습을 보입니다. 모델 10~13는 앞선 모델 중 가장 점수가 높은 epoch 500을 기준으로 배치사이즈를 변경해서 학습했습니다. 배치사이즈의 크기는 학습 시간에 영향을 크게 끼치며 배치 사이즈가 클수록 시간이 짧게 걸리고 반대로 배치사이즈를 줄일수록 시간이 오래 걸렸습니다. 전체적으로 이번 실험에서는 펄스데이터를 중복 사용, epoch 500회, mb_size는 5일 때 가장 우수한 정확도와 F1-score를 보여주고 있습니다.