머신러닝 기반 이상 항적 탐지 해안감시 모델 제안

강민주, 김민규, 조준범

성균관대학교

2023312782, 2024312548, 2022312957

요약 : 선박의 활동이 많은 시기 해안감시에 생길 수 있는 공백을 최소화하고 잠재적 위험을 식별하기 위해, 기존 선박들의 항적을 학습한 후 이상 항적을 나타내는 선박을 자동으로 감지하는 방법을 제안한다. 본 연구는 AIS 데이터를 활용해 지도학습의 일종인 XGBoost를 통해 일부 부족한 정보를 보간하고, K-Means 알고리즘을 적용하여 대표성을 띄는 항적들을 추출한다. 이후 추출된 정보들과 딥러닝을 기반으로 한 오토인코더 모델을 적절히 조합해 이상 항적을 실시간으로 검출하고자 한다. 이후 여러 가상 시나리오를 적용하여 본 모델에 대한 평가를 진행하고, 모델의 실효성을 보인다. 아울러 모델의 의의 및 한계점을 분석해보며 이를 개선할 수 있는 방법을 제시한다.

핵심용어 : 해안감시, 머신러닝, XGBoost, K-Means, 오토인코더, 이상 항적

1. 서론

바다를 통한 적의 침투를 막고, 해상에서의 돌발상황에 대처하기 위해 대한민국 국군은 해안감시 부대를 운용하고 있다. 각 부대 장병들은 책임 지역의 특징을 숙달하고 적의 예상 침투 전술을 고려하여 레이다, TOD 등의 장비를 다루며 국토의 최전선을 지킨다.

그러나 특정 계절과 시기에는 선박의 수가 평소보다 많고 그 이동량 또한 폭증하기에 사람의 동체시력과 순발력에 기대는 것만으로는 한계가 존재한다. 또한 권역 내 모든 선박을 추적해야 하기 때문에 매초 수많은 움직임을 동시에 관리하다 보면 주의를 덜 기울이게 되는 구역이 생길 수밖에 없다. 특히 '상대적으로 추적하는 대상의 수가 적은 곳'이 그러한 대상이 된다. 해당 구역에 대한 감시 공백을 줄이며 잠재적인 위험 요소를 사전에 식별 및 대응하기 위해, 머신러닝을 기반으로 선박들의 항적을 자동으로 분석하여 평소와 다른 행보를 띄는 선박에 대해 경고를 발생시키는 모델을 제안하고자 한다.

본 연구는 인천항과 인근 항구에서 출항하는 어선급의 선박(선외기급 선박도 포함)을 대상으로 AIS(Automatic Identification System) 신호를 활용한다. 먼저, 머신러닝 알고리즘을 적용하여 불규칙한 데이터 사이의 간격을 일정히 조정한 뒤, 선박들의 전반적인 항적을 분석한다. 이후, 속도와 방향 등의 요소를 고려해 대표성을 지닌 여러 군집으로 분류한 후 오토인코더 알고리즘을 함께 활용해 최종 이상치를 판단하는 방법을 제시한다.

2. 연구 방법

2-1. 학습 데이터

AIS는 선박의 위치를 포함한 여러 제원을 송수신하여 원활한 항해에 도움을 주고, 사고 발생 시 빠르게 대처하기 위해 개발된 국제적 신호체계이다. 대한민국 영해에서 발생하는 신호들은 가장 가까운 기지국을 거쳐 지역별 VTS(해상교통관제 시스템) 시설로 전송되어 관리된다. 이외에도 자동 입출항 신고 기능을 제공하는 V-PASS와 해상 내비게이션 시스템을 위한 LTE-M 등이 운용되고 있다. 이 중 가장 접근이 용이한 AIS 데이터를 활용해 모델의 실효성을 입증하고자 한다.

AIS 데이터는 선박의 명칭, 고유식별번호인 MMSI, 선종과 국적 등의 정적 데이터와 배가 움직임으로써 시간차를 두고 새롭게 생성되는 동적 데이터로 구분된다. 각 동적 데이터는 위도, 경도, 속도, 방향, 송신한 시간 등의 정보로 구성된다. 본 연구는 특정 위치에서의 선박의 동향을 분석하므로 "lat", "lon", "speed", "course", "last_position_UTC" 정보를 시간 순으로 나열하여 사용한다.

2-2. XGBoost를 이용한 항적 데이터 보간

선박 신호 체계는 일정한 시간을 기준으로 하여 정확한 제원을 수신받아 이를 관리하는 것을 목표로 한다. 하지만 선박과 기지국 사이의 거리, GPS 수신 장애, 송수신 장비의 고장 등의 여러 요인으로 인해 안정적인 흐름을 만들지 못하고 불규칙한 결과를 반환하기도 한다. 여러 선박의 항적을 동시에 파악하고 그 중 뚜렷한 특징을 보이는 군집을 분류해내기 위해서는 일정한 간격을 지닌 데이터를 구축해야 한다.

이를 위해 거리 혹은 시간을 기준으로 하여 기존의 흐름을 유지하는 방향으로 새로운 데이터를 예측하여 추가할 수 있다. 거리를 기준으로 할 시 4N 속도의 선박이 1분마다 움직이는 거리인 0.123km(0.066N/M) 간격으로 데이터를 보간할 수 있으나, 시간 당 움직이는 거리는 선박의 속도에 크게 종속적이기 때문에 향후 분석 과정에 있어 비효율성을 초래할 수 있다.

따라서 본 연구에서는 일정한 시간 간격의 경위도를 산출하기 위해 다양한 머신러닝 알고리즘을 활용해 적절한 시간 및 거리 차를 지닌 데이터를 학습한 뒤, 성능이 우수하고 손실이 최소화된 모델을 선정한다. 머신러닝 알고리즘을 항적 보간에 활용한 이유는 마지막 항적에서 선박의 속도와 방향을 고려해 가중치를 부여할 수 있기 때문이며, 이를 통해 보다 부드럽고 안정적인 항적 예측이 가능해진다.

$$\text{Input} = \{\ \text{Speed},\ \text{Course},\ \Delta\text{Time}\ \}$$
$$\text{Output} = \{\ \Delta\text{Latitude},\ \Delta\text{Longitude}\ \}$$

$$\text{where},\ 4 \le \text{Speed} \le 25 \text{ and}$$
$$\Delta\text{Time} \le 20$$

선정된 모델을 기반으로 기존 속도와 방향을 고려해 다음 항적의 경도와 위도를 예측하고 이를 원본 데이터에 추가한다. 그러나 직전 항적의 방향이 목표 항적을 정확히 가리키지 않는 경우, 최종 예측된 항적이 크게 왜곡될 가능성이 있다. 이를 보완하기 위해 현 위치에서 목표 위치까지의 각도와 기존 진행 각도를 3:2 비율로 조합하여 새로운 방향을 산출한다. 또한, 직전 항적이 내륙을 가리켜 섬을 가로지르는 등 잘못된 항적을 반환하는 경우에는 인위적인 가중치를 부여해 정상적인 항로를 유지할 수 있도록 조정한다.

보간 과정을 통해 평균 10분 간격으로 데이터를 나열할 수 있게 되었으며, 더욱 세밀하고 구체적인 군집을 형성하기 위해 연속된 두 항적 간의 시간 차이가 1분을 초과할 경우, 중간값을 계산하여 추가하는 방식으로 모든 데이터를 약 1분 간격으로 유지할 수 있게끔 한다.

해당 과정을 통해 선정된 XGBoost 모델은 대표적인 그래디언트 부스팅(Gradient Boosting) 알고리즘이다. 경사하강법을 기반으로 하는 그래디언트 부스팅은 여러 개의 결정 트리를 순차적으로 학습시키며, 각 트리의 잔차를 계산해 다음 단계에 반영하여 성능을 높이는 앙상블 기법 중 하나이다. XGBoost가 다른 알고리즘에 비해 좋은 성능을 보일 수 있던 것은 병렬적으로 학습함과 동시에 트리의 최대 깊이보다도 손실을 최적화하는 것에 중점을 두어 규모가 크고 복잡한 데이터에도 효과적으로 작용하기 때문이다.

2-3. K-Means 기반 군집화, 센트로이드 추출

1분 간격으로 정렬된 각 데이터는 특정 경위도에서의 속도, 방향, 송신된 시간 값을 담고 있다. 단순히 선박들이 많이 거치는 항로를 분석하고자 하면 경위도 값을 시간순으로 나열해 군집화하면 되겠으나, 본 연구는 전반적인 항로뿐만이 아닌 특정 위치에서 일반 선박과 다른 동향을 보이는 항적의 식별을 목표로 한다. 따라서 5분 동안의 위도, 경도, 속도, 방향, 이전 항적 대비 변화된 경도의 변화량과 위도의 변화량을 30차원의 벡터로 결합한 후 활용하고자 한다.

선박들이 자주 다니는 항로만을 추출하고자 하면 기준 시간을 10분 이상으로 지정하면 되지만, 본 연구는 특정 위치에서 다양한 요소를 세밀히 고려하는 것을 목표로 하며, 바삐 움직이는 실제 경계 근무 상황에서 빠르게 이상 징후를 감지할 수 있도록 5분을 기준으로 하였다. 서로 다른 선박, 다른 날짜의 데이터에 대해서는 철저히 분리해 두었으며 원해를 중점적으로 분석하고자 인천항 주변의 데이터는 일부 제거하였다.

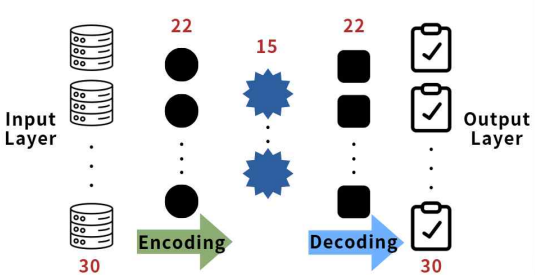
$$\text{Input} = \{\ T_1, T_2, T_3, T_4, T_5\ \}$$

$$T_n = \{\ \text{Latitude},\ \text{Longitude},\ \text{Speed},$$
$$\text{Course},\ \Delta\text{Latitude},\ \Delta\text{Longitude}\ \}$$

K-Means는 군집화를 위한 대표적인 비지도 학습 알고리즘으로, 초기화된 클러스터 중심과 데이터 간의 거리를 계산하고, 중심을 반복적으로 갱신하며 변화가 없을 때까지 군집을 형성하는 방식으로 동작한다. 클러스터 개수가 많으면 다양한 상태를 구분할 수 있지만, 각 군집에 속한 데이터 수가 줄어들어 이후 이상치를 검출하는 기준을 설정하는 데 어려움이 발생할 수 있다.

2-4. 데이터를 압축시키며 학습하는 오토인코더

오토인코더(AutoEncoder)는 딥러닝 기반 비지도 학습 기법으로, 입력 데이터를 중요한 요소 중심으로 압축한 뒤 이를 최대한 원래 형태로 복원하도록 학습하는 방법이다. 본 연구에서는 입력된 30차원 벡터를 22차원, 15차원으로 순차적으로 압축하는 인코더와, 이를 다시 22차원, 30차원으로 복원하는 디코더를 설계한다. 학습 데이터는 이전 단계와 동일한 구조를 사용했으며, 모든 요소를 균등하게 반영하기 위해 표준화를 수행하였다.



2-5. 이상 항적에 대한 판단 과정

새로운 항적 데이터의 이상 여부를 판단하기 위해, K-Means의 200개 군집과 오토인코더를 각각 활용한다.

군집을 통한 검출 과정은 다음과 같다. 먼저, 200개의 센트로이드와 각 샘플 간의 거리를 계산하여 거리의 평균과 표준편차를 구하고, 정규분포를 구성한다. 해당 정규분포에서 상위 5%와 하위 5% 지점을 임곗값으로 설정하고 저장한다. 이후, 새로운 항적 데이터가 입력되면 가장 가까운 센트로이드와의 거리를 계산하고, 사전에 구한 평균과 표준편차를 사용해 Z 점수로 표준화한다. 계산된 Z 점수가 해당 군집의 임곗값을 초과하거나 미달하는 경우, 해당 데이터를 이상치로 판단한다.

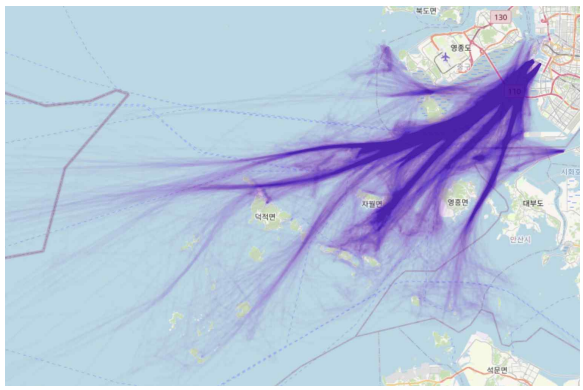
오토인코더는 정상 데이터로 학습되어 입력 벡터를 압축한 뒤 복원하는 과정에서 정상 항적에 가까운 방향으로 값을 재구성한다. 이상 항적이 입력되면 인코더와 디코더를 거치는 과정에서 원본과 복원된 값 사이에 큰 차이가 발생하며, 이를 통해 이상 여부를 검출한다. 이때, 이상치를 판단하는 임곗값은 정상 데이터를 학습하고 복원하는 과정에서 발생한 오차를 기준으로 설정한다.

이상 항적을 정상 항적으로 판별하는 오류를 최소화하기 위해 앞서 보인 두 가지 방법을 활용하였다. 두 기법 중 하나라도 이상치를 감지하면 해당 데이터를 이상 항적으로 판별하도록 결정했으며, 이로써 탐지 민감도를 높이고 신뢰성을 강화할 수 있게 된다.

3. 결과

3-1. 데이터 수집 및 분석

연구를 위해 인천 내 도서 (인천항~대무의도~덕적도) 인근에서 활동하는 52개의 선박들을 조사해 2022년부터 2024년 사이 활동이 가장 활발한 시기에 대한 일별 항적 데이터를 수집하였다. 수집 단계에서 출항 기록이 없거나 출항 후 바로 입항하는 등 개수가 적고 불필요한 부분은 배제하였다. 이후 추가적인 정제과정을 거쳐 합산 4,504일, 약 90,000개의 항적 데이터를 학습 데이터로 활용하였다.



3-2. 항적 보간

일부 항적 데이터 내 시간 간격이 넓어 발생하는 불규칙한 흐름을 줄이기 위해 여러 머신러닝 알고리즘을 적용시켜 중간 항적을 보간 하고자 했다. 연속적으로 이어지는 2개의 항적 데이터의 속도, 방향, 시간차를 학습해 경도 차이, 위도 차이를 찾아야 하므로 입력 데이터로 출력 데이터를 예측하는 지도학습을 사용하였다. 이때

실험을 위해 사용된 알고리즘들과 각 성능에 대한 결과는 다음과 같으며, 안정적인 성능과 낮은 손실을 보여준 XGBoost 알고리즘이 채택되었다.

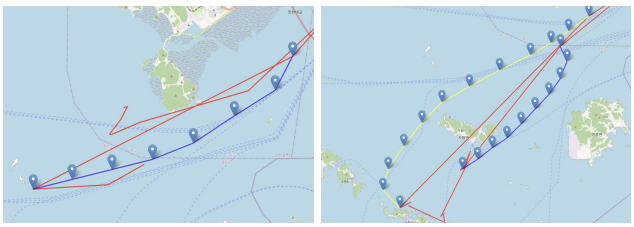
다중 회귀		
	훈련	테스트
경도	0.75	0.77
위도	0.79	0.78

XGBoost		
	훈련	테스트
경도	0.88	0.89
위도	0.76	0.80

랜덤 포레스트		
	훈련	테스트
경도	0.88	0.76
위도	0.86	0.81

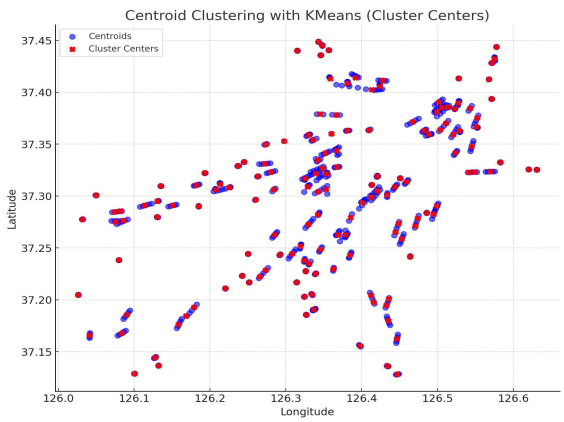
딥러닝		
	훈련	테스트
통합	0.76	0.77

XGBoost 모델을 통해 40분 이상의 시간 차이가 발생하는 두 항적 사이에서 10분 간격으로 새로운 항적을 예측할 수 있었으며, 직전 항적의 속도와 방향을 고려해 실제 선박의 흐름과 유사한 결과를 도출할 수 있었다.



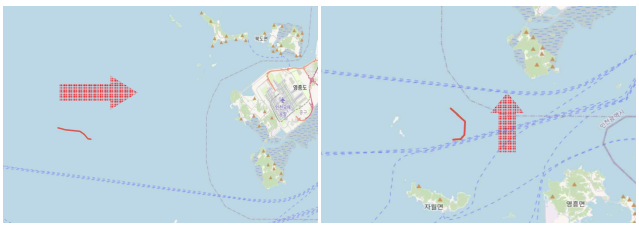
3-3. 데이터 군집화 및 센트로이드 추출

1분 간격으로 구성된 항적 데이터를 5분 단위로 묶어 새로운 벡터 집합을 생성한 뒤, K-Means를 활용하여 총 200개의 군집을 형성하였다. 군집과 중심점을 시각화한 결과, 선박들이 자주 지나는 주요 항로를 효과적으로 나타냄을 확인할 수 있었다. 또한 동일한 공간 내에서도 방향과 속도와 같은 요소의 차이에 따라 서로 다른 특성을 가진 2개 이상의 군집이 형성되는 것이 관찰되었다.



3-4. 가상 시나리오를 활용한 이상치 검출

본 연구는 가장 가까운 군집을 찾아 군집 내 데이터의 분포를 기반으로 중심으로부터의 거리와, 오토인코더를 통해 복원된 결과와 실제 값 간의 차이를 조합하여 최종 이상 여부를 반환하는 방법을 제안하였다. 해당 방법을 검증하기 위해 실제 발생할 수 있는 두 가지 가상 시나리오를 제작하고 이를 이용해 실험을 진행하였다.



첫 번째 시나리오는 일정한 속도로 서해 먼 바다에서 영종도 서방으로 접근하는 경우이고, 두 번째 시나리오는 인천항 정기 항로를 통해 내해로 접근하는 중 갑자기 속도를 높이며 반대 방향으로 도주하는 경우이다. 해당 시나리오들은 감시 자산에 포착되어 추적을 시작한 후 5분 동안의 항적으로 가정하며 그 결과는 다음과 같다.

이상 시나리오 1			
	이상치	임곗값	검출 여부
K-Means	-0.0093	-34.0329 492.0737	**비검출**
오토인코더	52.7219	29.0540	검출
최종 판단	검출		

이상 시나리오 2			
	이상치	임곗값	검출 여부
K-Means	1.2699	161.6649 392.4242	검출
오토인코더	1421.2792	29.0540	검출
최종 판단	검출		

K-Means 모델이 시나리오 1의 이상을 검출하지 못한 것은 해당 군집 내 데이터의 분포에 영향을 받은 것으로 추정된다. 센트로이드 중심과 거리가 먼 극단적인 데이터가 다른 군집에 비해 많아, 이상 여부를 판단하는 임곗값 설정에 영향을 미쳤을 가능성이 있다. 이를 해결하기 위해 군집의 총 개수를 늘려 극단적인 데이터가 서로 다른 군집으로 분리되도록 하거나, 임곗값을 조정하여 이상 탐지의 민감도를 높이는 방안을 고려할 수 있다.

4. 결론

4-1. 결론

서론에서 언급했듯이, 선박의 수가 증가하고 활동이 활발해질수록 인간의 능력만으로 모든 선박을 관리하며 수상한 선박을 분류하는 것은 어려워지고 있다. 특히 보트, 요트, 카약, 제트스키 등 해양 레저 산업의 확대는 관리해야 할 선박과 영역을 크게 늘리며, 이를 악용한 침투 가능성을 증가시키고 있다. 이러한 상황에서 항적을 자동으로 분석하고 이상 항적을 판단하는 모델은 해안감시 강화에 기여를 할 것으로 예상된다.

해당 모델을 개발하고 평가하기 위해 본 연구에서는 공개된 AIS 신호 체계를 활용하였고, 이상 시나리오를 정상 항적과 구분지어 이상을 검출하는 과정을 시연하였다. 하지만 실제 감시 임무 현장에서는 선명, 선종 등의 제원을 확인할 수 있는 선박보다 신호를 송수신하지 않아 피아식별이 되지 않는 소형 선박과 레져 선박 감시에 더 큰 관심이 필요하다. 다만 모든 감시 기지에는 추적한 선박에 대한 제원 및 항적을 면밀히 기록하고 있으므로, 이러한 데이터를 기반으로 모델을 적용하여 실질적인 감시 작업에 활용할 수 있을 것으로 기대된다.

4-2. 의의

이상 항적 탐지 모델이 가지는 구체적인 의의는 다음과 같다. 첫 번째로, 경제적 이익을 창출할 수 있다. 24시간 운영되는 해안감시 부대의 특성상, 장병들의 피로가 누적되며, 어선, 화물선 등의 생계형 선박을 관리하는 VTS 관제 시설에서도 업무 부하로 인해 위험을 감지하지 못하는 상황이 발생할 수도 있다. 본 모델은 감시 장비를 운용하는 장병과 관제사를 보조하여 추가적인 인력 없이도 위험 상황을 감지해 이를 알림으로써 인간의 한계를 보완할 수 있다.

두 번째로, 사고 발생 시 이를 조기에 파악하여 신속한 대응과 상황 처리가 가능해진다. 평시에 이상 항적을 띄는 선박은 보통 사고 위험이 높거나 사고가 발생한 선박을 의미하며, 이를 빠르게 식별하고 구조 작전에 돌입해야 인명피해를 줄일 수 있다. 특히 구조 신호 장치가 설치되지 않은 소형 선박의 경우, 사고 발생 시 자체적인 수습이 어려워 피해가 확대될 가능성이 크다. 이러한 상황에서 본 모델을 활용하면 사고를 조기에 감지하고 피해가 커지는 것을 예방할 수 있게 된다.

4-3. 한계점

본 모델이 가지는 여러 이점에도 불구하고 여전히 해결해야 할 과제들이 존재한다. 첫째로, 앞서 XGBoost를 이용해 데이터를 학습하는 과정에서 테스트 세트에 대한 정확도가 상대적으로 낮아 과대적합된 경향을 보인다는 것이다. 이러한 문제에 대해서는 보다 세밀하고 정확히 기록된 현장의 데이터를 활용하고, 선박의 움직임 데이터는 시계열 정보의 특성을 가진다는 것에 의거해 RNN과 같은 이전 상태를 반영하는 알고리즘을 적용하면 학습의 정확도와 안정성이 개선될 것으로 생각된다.

둘째로, 이상 항적에 대한 데이터가 부족하여 현재 모델은 정상 데이터와 차이가 큰 항적만을 이상으로 검출하게 된다. 이는 모델의 일반화 여부를 평가하기 어렵게 하며, 이상 거동 항적에 대한 명확한 기준이 부재하다는 문제로 이어진다. 이를 해결하기 위해, 모델을 사용하는 담당자의 경험을 바탕으로 인위적인 조정을 병행할 필요가 있다. 초기에는 민감도를 높게 설정해 가능한 많은 이상 항적을 검출한 뒤, 반환된 결과를 분석하여 임곗값을 점진적으로 조정함으로써 적절한 민감도와 기준을 찾아가는 과정이 요구된다.

5. 참고 문헌

[1] 오재용 외, 「AIS 데이터 분석을 통한 이상 거동 선박의 식별에 관한 연구」, 『한국항해항만학회지』, 제42권, 제4호, 2018.

[2] 이원희 외, 「보간기법을 활용한 AIS 데이터 기반 선박 경로 예측 딥러닝 연구」, 『한국컴퓨터정보학회논문지』, 제29권, 제3호, 2024.

[3] “4주차 - AutoEncoder 기반 이상치 탐지 알고리즘”, ToBigs1617.log, 2022년 05월 12일 수정, 2024년 12월 04일 접속, https://velog.io/@tobigsts1617/4주차-AutoEncoder-기반-이상치-탐지-알고리즘.

[4] 김화업 외, 「중소형 선박용 해양 네비게이션 시스템의 구현」, 『한국정보기술학회』, 제11권, 제3호, 2013.