한국항해항만학회지 제47권 제6호
J. Navig. Port Res. Vol. 47, No. 6 : 358-366, December 2023 (ISSN:1598-5725(Print)/ISSN:2093-8470(Online))
DOI : http://dx.doi.org/10.5394/KINPR.2023.47.6.358

준지도 학습을 활용한 사용자 기반 소형 어선 충돌 경보 분류 모델에 대한 연구

석호준* · 심승** · 우정훈** · 조준래** · 정재룡** · 조득재*** · † 백종화

*,**슈어소프트테크(주) 연구원, ***선박해양플랜트연구소 책임연구원,
† 선박해양플랜트연구소 연구원

A Study on the User-Based Small Fishing Boat Collision Alarm Classification Model Using Semi-supervised Learning

Ho-June Seok · Seung Sim** · Jeong-Hun Woo** · Jun-Rae Cho** · Jaeyong Jung** · DeukJae Cho*** · † Jong-Hwa Baek*

*,**Researcher, Suresoft Technologies Inc., Seongnam 13453, Korea
***Principal Researcher, Korea Research Institute of Ships & Ocean Engineering, Daejeon 34103, Korea
† Researcher, Korea Research Institute of Ships & Ocean Engineering, Daejeon 34103, Korea*

요 약 : 본 연구는 해양수산부의 '지능형 해상교통정보시스템' 서비스 중 '사고취약선박 모니터링 서비스'의 선박 충돌 경보를 개선하기 위한 것으로, 현재의 선박 충돌 경보는 대형 선박 위주의 데이터와 그 운항자에 기반한 설문조사 레이블을 가지고 지도 학습(SL)한 모델을 사용하고 있다. 이로 인해, 소형선박 데이터 및 운항자의 의견이 현재 충돌 지도학습 모델에 반영되지 않아, 소형선박 운항자가 느끼는 체감보다 먼 거리에서 경보가 제공되기 때문에 그 효과가 미비하다. 또한, 지도학습(SL) 방법은 레이블링 된 다수의 데이터가 필요하지만, 레이블링 과정에서 많은 자원과 시간이 필요하다. 본 논문은 이러한 한계를 극복하기 위해 준지도학습(SSL)의 알고리즘인 Label Propagation과 TabNet을 사용하여 레이블이 결정되지 않은 데이터를 활용하여 소형선박을 위한 충돌 경보의 분류 모델을 연구하였다. 충돌 경보의 분류 모델을 활용하여 소형선박 운항자를 대상으로 실해역 시험을 수행한 결과 운항자의 만족도가 증가하는 결과를 확인하였다.

핵심용어 : 해상디지털, 데이터 파이프라인, 데이터 전처리, 준지도학습, 머신러닝, 딥러닝

Abstract : *This study aimed to provide a solution for improving ship collision alert of the 'accident vulnerable ship monitoring service' among the 'intelligent marine traffic information system' services of the Ministry of Oceans and Fisheries. The current ship collision alert uses a supervised learning (SL) model with survey labels based on large ship-oriented data and its operators. Consequently, the small ship data and the operator's opinion are not reflected in the current collision-supervised learning model, and the effect is insufficient because the alarm is provided from a longer distance than the small ship operator feels. In addition, the supervised learning (SL) method requires a large number of labeled data, and the labeling process requires a lot of resources and time. To overcome these limitations, in this paper, the classification model of collision alerts for small ships using unlabeled data with the semi-supervised learning (SSL) algorithms (Label Propagation and TabNet) was studied. Results of real-time experiments on small ship operators using the classification model of collision alerts showed that the satisfaction of operators increased.*

Key words : *maritime digital, data pipeline, data preprocessing, semi-supervised learning, machine learning, deep learning*

1. 서 론

해양수산부는 해양사고를 예방하기 위해 지능형 해상교통정보시스템(이하 바다 내비게이션)을 구축하여 국내 선박을 대상으로 다양한 서비스를 제공하고 있다(Yang et al., 2016). 현재 총 5,729척의 국내 선박에 바다 내비게이션 서비스를 위한 전용 단말기가 설치되어 있으며, 일일 평균 약 3,000척에 서비스가 제공되고 있다(Ministry of Oceans and Fisheries, 2023). 바다 내비게이션 관련 서비스 데이터는 선박해양플랜트연구소 내 e-Navigation 통합시험센터(이하 통합시험센터)에서 수집 및 저장되고 있으며, 데이터 파이프라인을 통해 해당 데이터에 접근할 수 있는 환경을 제공하고 있다(Baek et

† Corresponding Author : 정회원, jhbaek@kriso.re.kr 042)866-3636
* 정회원, hjseok@suresofttech.com 031)606-2356
(주) 이 논문은 "준지도 학습 기반 선박충돌 예측에 대한 연구"란 제목으로 "2023 공동학술대회 한국항만학회논문집(부산 벡스코(BEXCO), 2023.5.3.-4. pp.204-205)"에 발표되었음.

al., 2022).

사고취약선박 모니터링 서비스는 바다 내비게이션에서 제공하는 서비스 중 하나로, 선박의 위치 정보 등을 기반으로 충돌·좌초 위험도를 평가하여 선박이 위험 상황을 인식할 수 있도록 경보를 제공하는 서비스이다. 현재 사고취약선박 모니터링 서비스의 충돌 경보는 먼저 선박에 경보가 필요한 상황인지 판단하고, 두 번째로 경보 단계(관심-주의-위험)를 결정한 뒤 선박으로 제공되고 있다. 두 단계 모두 지도학습(Supervised Learning 이하 SL)을 활용한 모델을 적용하고 있으며, 경보 단계 판단은 대형 선박에 종사하는 운항자를 대상으로 실시한 설문조사 결과의 데이터들로 학습되었다(Yang, 2021). 즉, 현재 서비스의 경보 단계는 대형 선박 운항자 의견에 의존하고 있으며, 소형선박 운항자 관점에서 느끼는 경보 단계의 체감 경보와 다소 차이가 있다.

본 논문에서는 이러한 한계를 극복하고자 소형 어선을 위한 충돌 경보 분류 모델에 대해 연구한 결과에 대해 기술한다. 소형 어선을 위한 충돌 경보 분류 모델 연구는 다음과 같이 수행하였다. 기존 사고취약선박 모니터링 서비스의 데이터과학 알고리즘 도입 과정을 참고하여, 소형 어선 운항자를 대상으로 실해역 시험을 통해 경보 단계 만족도 조사 및 레이블링을 진행하고, 이를 학습 데이터에 사용한다. 이렇게 수집한 학습 데이터(이하 Labeled data)와 통합시험센터에서 연계하는 데이터(이하 Unlabeled data)를 활용하여 준지도학습(Semi-Supervised Learning 이하 SSL)을 이용한 분류 모델을 적용했다. 그리고 이렇게 생성한 모델의 정확도를 확인하고 모델의 성능을 점검하였다.

2. 선행 연구 분석

선박 충돌경보는 선박의 크기, 거리, DCPA, TCPA 등 다양한 요소가 필요하며, 상황과 조건에 따라 경보가 달라질 수 있다. 가장 단순한 방법은 선박 사이의 일정한 거리 및 DCPA 기준에 따라 경보 등급을 나누는 것이다. 예를 들어, 소형선박 정면 상황의 경우 0.5mile 이하는 위험 상황의 경보를 제공한다. 하지만 이렇게 모든 조우 상황을 규칙으로 정의하기 어려운 경우가 많으며, 실제 사용자 만족도가 감소하는 상황이 발생한다.

최근에는 다양한 머신러닝 방법이 여러 산업에서 도입됨에 따라 선박 충돌과 관련한 연구가 지속되고 있다. 이러한 연구들 중 선박 충돌경보를 위해 Labeled data를 활용하여 SL 방법의 분류 모델인 Logistic regression 모델을 사용하여 RICM 모델을 제안하는 연구가 진행되기도 하였다(Yang, 2021).

SL을 위해서는 많은 Labeled data가 요구되고 Unlabeled data의 레이블링이 필요하다. 하지만 Unlabeled data를 실제 사용자 의견으로 레이블링 하거나 만족도를 계산하기에는 많은 자원과 시간이 소요된다. 따라서, Labeled data와 Unlabeled data를 동시에 사용하여 모델링할 수 있는 데이터 과학 학습 방법인 SSL모델을 통해 선박 충돌 경보 분류에 적용하는 연구도 진행되었다(Seok et al., 2023). 또한, 경보를 분류해서 제공하는 모델이 아니라 다양한 딥러닝 방법을 도입해 선박의 항적으로부터 이후 항적을 추론하여 충돌 위험을 식별하는 기술을 제안하기도 한다. 이 방법들은 LSTM(Long Short-Term Memory) 모델을 쌍방향으로 합성시킨 BiLSTM과 Attention 메커니즘을 병용하여 선박의 거동을 예측하고 위험을 미리 인지할 수 있도록 한다(J. MA et al., 2020).

한편, 충돌경보에서 사용자의 만족도를 끌어올리기 위한 연구들이 제안되기도 했다. 소형선박 충돌경보 알고리즘의 실용성과 사용자 만족도 향상을 위해 소형선박 운항자들을 대상으로 설문조사를 진행하여 알고리즘을 개선하였다(Park et al., 2021).

본 연구에서는 앞선 연구들을 바탕으로 SSL을 적용하고 소형 어선을 위한 충돌경보 분류 모델을 생성하였다. 그리고 실제 만족도를 위한 실해역 시험 데이터를 분리하여 만족도 및 거리 등의 변화를 조사하는 연구를 진행하였다(Fig. 1).

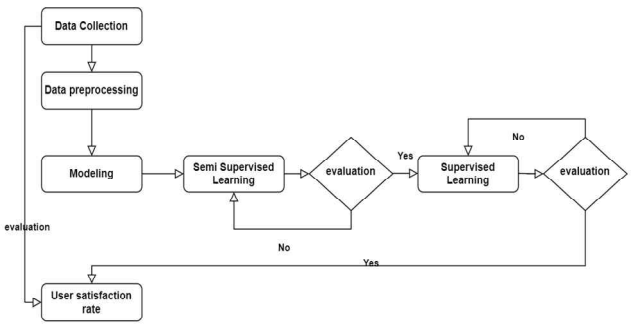

Fig. 1 Procedure of data modeling process

3. 데이터 수집 및 파이프라인

3.1 데이터 수집 도구

Fig. 2 ECS data collection tool

바다 내비게이션의 기존 사고취약선박 모니터링 시스템에서 소형 어선의 Labeled data를 수집하기 위하여 ECS(Electronic Chart System) 기반 효용성 평가도구를 개발 및 활용하였다

(Lee at al., 2022). ECS는 실해역 시험에서 선박 운항자의 경험으로 판단한 경보 단계를 새롭게 레이블링 할 수 있도록 개발된 평가 도구이다(Fig. 2). 또한, 기존 서비스와 데이터과학 기반 의사결정지원 모델의 충돌 경보를 단계별(관심-주의-위험)로 제공하고 사용자는 표출된 경보의 만족도를 선택하면 5초 간격으로 자동 저장할 수 있다.

3.2 데이터 수집

Labeled data는 총 9 회 ECS를 활용한 실해역 시험을 진행하여 데이터 수집을 진행하였다(Table 1). 이때 대상은 소형 어선 운항자로 실제 해상환경에서 이루어졌다. 데이터 수집 시나리오는 2대의 소형 어선을 활용하여 충돌 경보를 발생시키기 위한 총 10가지 상황으로 구성하였다. 실해역 시험 시나리오는 국제해상충돌방지규칙(COLREGs)과 해사안전법을 참고한 정면(Head-On) 상황 1가지, 횡단(Crossing) 상황 6가지, 추월(Overtaking) 상황 2가지 등을 구분하여 선정하였다. 더불어, 20m 소형 어선의 경우 정박하여 어로작업에 종사하는 상황이 많으며 이때 접근하는 선박을 발견하지 못하고 충돌하는 사례가 다수 있는 것으로 분석되어 접근(Approaching) 상황 1가지를 추가하여 시나리오를 계획하였다(Fig. 3).

선박의 속력과 방위는 실해역 시험 장소와 조류 그리고 실제 소형선박 운항자가 평소 운항하는 속력을 기반으로 10 Knot에서 15Knot 범위로 설정하고 시험하였다.

Table 1 Real time experiment data collection list

No	Date	L.O.A	Breath	GT	Region
1	2022.08.17	16.50m	3.80m	9.77t	Gunsan
		16.17m	3.80m	9.77t	
2	2022.09.28	13.53m	3.53m	9.77t	Yeongheungdo
		16.68m	3.80m	9.77t	
3	2022.09.29	13.53m	3.53m	9.77t	Yeongheungdo
		16.68m	3.80m	9.77t	
4	2022.10.26	17.59m	3.80m	9.77t	Yeosu
		16.45m	3.82m	9.77t	
5	2022.10.27	12.87m	3.42m	9.77t	Yeosu
		13.31m	3.42m	9.77t	
6	2022.11.10	16.68m	4.09m	9.77t	Jeju
		15.86m	4.08m	9.77t	
7	2022.11.17	14.90m	3.32m	7.93t	Yeongmok
		15.75m	3.80m	9.77t	
8	2022.11.18	14.90m	3.32m	7.93t	Yeongmok
		15.75m	3.80m	9.77t	
9	2023.04.12	16.17m	3.80m	9.77t	Gunsan
		16.50m	3.80m	9.77t	

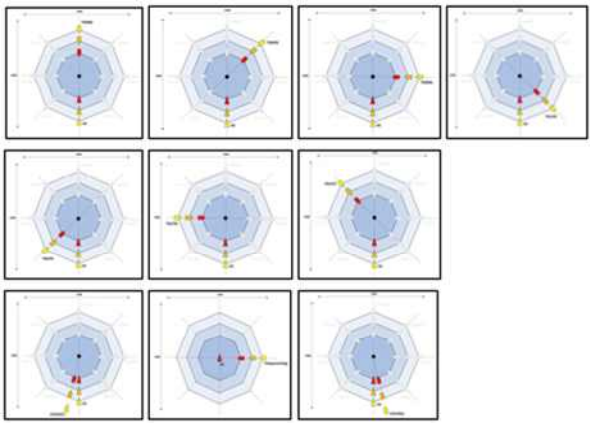

Fig. 3 Real time experiment Scenario

Unlabeled data 수집은 통합시험센터에 구축된 사고취약선박 모니터링 서비스에서 제공되는 충돌 경보를 활용하였다. 현재 서비스에서 통신 인터페이스로 사용 중인 Data Distribution Service(이하 DDS)는 Topic 구조를 Subscribe 함으로써 데이터를 수집 할 수 있었다. 따라서, 통합시험센터 내 충돌 경보 수집 프로그램 모듈을 구축하여 Unlabeled data를 수집하였다.

3.3 데이터 수집 결과

ECS 도구를 사용한 Labeled data 수집은 2022년 8월을 시작으로 2023년 4월까지 이루어졌다. 그 결과 학습 데이터는 수집한 Raw data를 처리하여 최종적으로 12,084 row Dataset을 구축하였다. 또한, 2023년 8월 새롭게 생성한 모델을 바탕으로 만족도 평가를 위한 데이터 수집을 진행하였다(Table 2). DDS Topic을 활용한 Unlabeled data는 2022년 하반기를 시작으로 현재까지 총 1,016 Gigabytes를 수집하고 있다.

Table 2 Real time data collection for user satisfaction rate evaluation

No	Date	L.O.A	Breath	GT	Region
1	2023.08.16	12.87m	3.42m	9.77t	Yeosu
		13.31m	3.42m	9.77t	
2	2023.08.17	12.87m	3.40m	9.77t	Yeosu
		16.81m	3.80m	9.77t	

3.4 데이터 파이프라인

수집된 Unlabeled Data는 1일 분량을 정제 후 서버에 업로드 하고 있어, 데이터 연계 구성도(Fig. 4)와 같은 1일 단위의 배치처리 방식을 사용한다. Service Data ETL(Extract·Transform·Load)에 따라 기존 서비스에서 발생하는 경보를 포함한 자선과 타선의 정적 정보와 동적 정보를 수집한다. 이때, 선박의 MMSI(Maritime Mobile Service Identity)나 MRN(Maritime Resource Name) 등의 개인정보를 익명화 하기 위하여 새로운 선박 고유 번호를 부여하고 선박 제원 정보

를 활용해 연결하였다.

Labeled data의 경우 로컬 내부에 수집되는 데이터와 파이프라인을 연계하여 Raw data를 결합 후 분석 서버로 전송된다. 이때, 기존 서비스에서 발생하는 경보와 사용자 경험으로 판단한 종속 변수인 경보 레이블을 같이 저장한다.

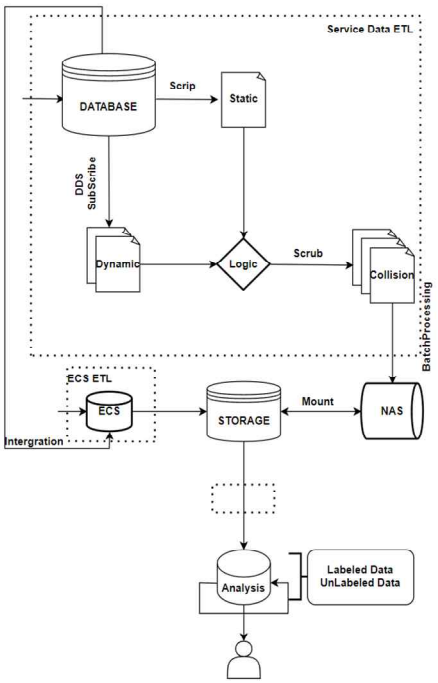

Fig. 4 Data pipeline diagram

4. 데이터 전처리

실해역 시험의 상황과 유사한 서비스 데이터를 추출하고 준지도 학습에 사용하기 위하여 다음과 같은 2차 정제 과정을 거쳤다.

4.1 파생 변수 생성

현재 자□타선의 총합 위험도(dRisk)는 기존 데이터과학 모델로 계산한 위험도가 저장된다. 하지만 자□타선의 각 위험도(dOsRisk, dTsRisk)는 Fuzzy 기반의 위험도 값이 산출된다. 따라서, 기존 데이터과학 모델의 결과를 반영하지 않는 Fuzzy 기반 위험도의 총합 위험도인 MaxRisk 변수를 생성하였다.

더불어, 선박 간의 거리 변수(RNG)를 생성하기 위하여 각각 자□타선의 위도, 경도를 사용하여 Haversine 계산을 위한 수식 (1), (2), (3) 공식을 적용하였다.

$$distance = 2 * R * \frac{\Delta\lambda}{2} \quad (1)$$

$$R = 6371(EarthRadius \in km) \quad (2)$$

$$\Delta\lambda = \sqrt{\frac{\Delta}{2}} * \Pi/180 \quad (3)$$

4.2 소형선박 추출

선박 코드별 선박의 정보를 가지고 있는 Table 3 규칙을 생성하여 데이터를 저장하였다. 추가로, 이러한 선박 크기 테이블과 선종 및 상세 선종을 연결하여 AB01010e와 같은 고유번호는 길이 20m 이하의 일반어선 등으로 식별하였다. 이때, 전장 20m 이하의 소형선박에 해당하는 고유 번호는 총 3,012개로 식별하였다.

Table 3 L.O.A mapping table

Code(0)	L.O.A
N	L.O.A ≤ 0m
A	1m < L.O.A ≤ 20m
B	20m < L.O.A ≤ 40m
C	40m < L.O.A ≤ 80m
D	80m < L.O.A ≤ 200m
E	200m ≤ L.O.A

4.3 오류 값 제거

자□타선의 위도 및 경도가 각각 180°와 90°가 넘는 값을 오류 값으로 판단하고 제거하였다. 또한, 자□타선의 SOG가 0 knot인 값은 선박이 정지 상태일 때의 SOG가 아닌 오류 값으로 판단하고 제거하였다.

4.4 유효 데이터 식별

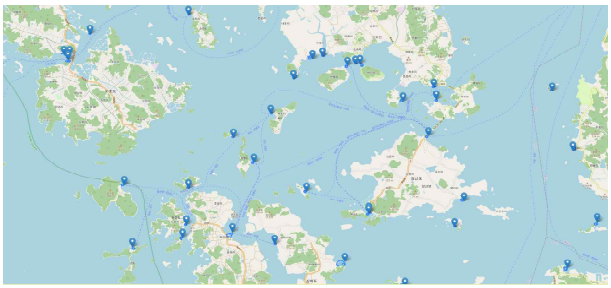

Fig. 5 Korean port polygon data using national land data of 'geoservice'

Source : http://www.gisdeveloper.co.kr/

한국 국토 데이터와 선박의 위치 데이터를 활용하여 국토의 경계와 555m 이하로 근접한 선박들의 위치를 추적하였다. 이렇게 추적한 위치를 바탕으로 자체 구축한 1,550개의 항구 데이터를 활용하여(Fig. 5), 항구 외부에서 발생한 경보를 학습 대상으로 식별하였다. 이때, 항구 외부 결정 판단은 자타선의 위도, 경도, CPA 지점을 기준으로 결정하였다.

사용자 경보의 극단적 의견을 제거하기 위하여, 각 경보 미송출(None-Alarm), 관심(Information), 주의(Caution), 위험(Alarm)의 선박 사이의 거리(RNG) 값에서 사분위수 범위(IQR)를 계산하여 $Q1 - 1.5 \times IQR$ 이하 값과 $Q3 + 1.5 \times IQR$

이상 값을 이상치로 판단하여 제거하였다.

4.5 상황 분류

현 데이터는 정면(Head-On), 횡단(Crossing), 추월(Overtaking) 각각의 상황을 판단할 수 있는 변수(IzCategory)를 활용하여 조우상황을 판단한다. 하지만 실제 서비스 데이터인 Unlabeled data는 모든 선박들이 미리 설정된 조우상황 내에서 정의되지 않아 정면(Head-On), 횡단(Crossing) 등 그 값이 변경된다. 따라서, 유효 길이 설정을 통해 일정한 조우 상황을 생성하였다. 더불어, 모든 경보 단계를 균일하게 제공받는 데이터로 처리하기 위하여 위험도의 범위를 제한하였다.

이렇게 Labeled data를 수집한 상황과의 비율을 조절하여 Unlabeled data를 분류하였다. 다음처럼 식별한 데이터들을 준지도 학습을 위한 Unlabeled dataset으로 구축하고 Random sampling 하여 사용하였다.

5. 모델 학습 및 평가

5.1 레이블 프로파게이션(Label Propagation)

5.1.1 모델 설명

준지도 학습은 Labeled data의 독립 변수(Xi) 및 종속 변수(Yi)와 Unlabeled data의 독립 변수(Xj)를 함께 사용하여 존재하지 않는 종속 변수($\hat{Y}$)를 예측하는 방법론이다.

본 연구에서는 그래프 기반 준지도학습 Label Propagation을 적용하였다.

Table 4 Label Propagation

Label Propagation
1. Compute Matrix D $=\sum$jWij
2. Initialize $\hat{Y}$(i)$=(Y_1,\cdots,0)$
3. Iterate $\hat{Y}$(i+1)$=D^{-1}W\hat{Y}$
4. Converge to $\hat{Y}(\infty)$

Label Propagation은 변수 간의 유사성을 기반으로, 레이블이 결정된 데이터에서 레이블이 결정되지 않은 데이터로 레이블을 전파함으로써 작동한다. 종속 변수 Yi를 가지는 데이터 점에 레이블을 할당하고 각 데이터 포인트 쌍의 변수 벡터 간의 유사성을 측정하는 Matrix를 사용하여 종속 변수 $\hat{Y}$의 데이터 포인트로 레이블을 전파한다(Yoshua et al., 2006; Olivier et al., 2005). 유사성 Matrix가 계산되면 반복적으로 종속 변수 $\hat{Y}$에 레이블을 부여한다. 각 인접 데이터 포인트의 레이블 가중평균을 계산하고 수렴될 때까지 반복한다(Table 4).

5.1.2 실험 과정

모델의 유효성을 검증하기 위하여 Labeled Data를 8:2의 비율로 학습 데이터와 테스트 데이터로 분리하였다. 또한 데이터의 스케일링이나 정규화 등의 영향이 적은 의사결정나무(Decision Tree)를 분류 모델의 기본적인 학습기로 선정하였다. 이때, 분리한 학습 데이터의 50%를 이용하여 학습한 SL_1과 100 %를 이용하여 학습한 SL_2를 분리하여 측정하였다. SSL을 위하여 SL_1에 Unlabeled data의 비율을 각기 달리한 4 개의 데이터셋을 구축하였다(Fig. 6). 이때, Unlabeled data의 비율은 각 10%, 30%, 50%, 70%이며 각 데이터 셋의 이름을 SSL_10, SSL_30, SSL_50, SSL_70로 구분하였다.

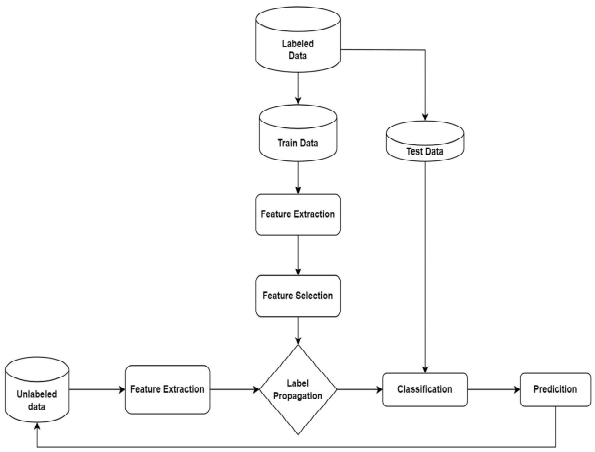

Fig. 6 SSL using DT model experiment process

처음 분리한 테스트 데이터와 SL_1, SL_2, SSL_10, SSL_30, SSL_50, SSL_70을 동일한 파라미터의 의사결정나무 모델로 학습하고 정확도를 측정하였다.

이때, 사용할 독립변수는 Labeled data의 종속 변수(Yi)와 상관관계가 가장 높은 변수를 선택하는 SelectKBest selection과 RandomForest 모델의 Tree-based selection 등을 수행하였다. 더불어, 다중공선성 판단을 위해 VIF(Variance Inflation Factor) 값을 확인하고 선행 연구들과 종합하여 변수를 각 선박의 DCPA, TCPA, RNG, MaxRisk로 선정하였다. 추가적으로, 이 모든 과정들을 총 5 회 반복 진행하여 평균값을 계산하였다.

5.1.3 실험 결과

처음 분리한 테스트 데이터와 SL_1, SL_2, SSL_10, SSL_30, SSL_50, SSL_70의 동일한 파라미터의 Decision Tree 모델로 학습한 결과를 다음과 같이 확인하였다(Table 5). 최대 SL_2의 Accuracy를 기준으로 SL_1의 Unlabeled data의 50% 비율인 SSL_50까지 정확도가 지속적으로 향상하는 것을 확인하였다. 따라서, Labeled data를 충분하게 확보할 수 없는 상황에서 SSL 접근이 모델의 정확도를 높여주는 접근이 될 수 있다.

Table 5 Prediction results of label propagation SL and SSL

Learning Method	Train		Test	Accuracy
	L	U	L	Test
SL_1	4,833	-	2,417	0.613
SL_2	9,667	-		0.628
SSL_10	4,833	483		0.618
SSL_30	4,833	1,450		0.620
SSL_50	4,833	2,417		0.623
SSL_70	4,833	3,384		0.623

*Labeled: L / Unlabeled: U

실험 결과에 따라, 가장 많은 Labeled data를 가진 SL_2의 데이터셋에 UnLabeled data를 추가하여 XGBoost Classifier 모델링을 수행하였다. 과적합 방지를 위하여 총 5회 Cross-Validation을 수행하고 각 모델의 최적 성능을 위해서 GridSearchCV Library를 사용하여 최적 파라미터를 조정해 주었다. 더불어, 평균 Accuracy, Recall, Precision, F1-Score를 계산하였다(Fig. 7). Table 5와 마찬가지로 UnLabeled data의 50%인 SSL_50을 SL_2와 학습한 결과(이하 SSL_50_XGB)를 최적의 모델로 확인하고 배포하였다.

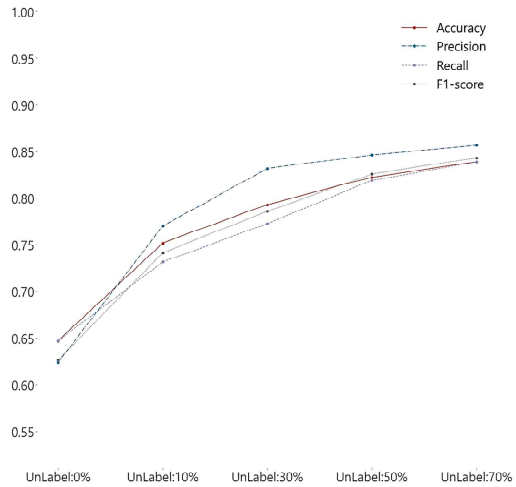

Fig. 7 Comparing cross validation score on XGB

5.2 탭넷(TabNet)

5.2.1 모델 설명

Tabular data를 학습하는 본 연구에서 Label Propagation보다 많은 양의 Unlabeled data를 활용하여 SSL을 진행하고 확인하기 위하여 TabNet을 적용하였다.

TabNet은 Google에서 연구한 Tabular data에서 성능이 좋은 딥러닝 알고리즘으로서 SL, SSL 모두에 활용할 수 있다(Arik et al., 2019). 더불어, AutoEncoder 구조를 통해 결측값들이 포함되어도 별도의 전처리 없이 대체할 수 있고 딥러닝 특성상 Feature Engineering과 같은 단계를 완화시킬 수 있다(Fig. 8).

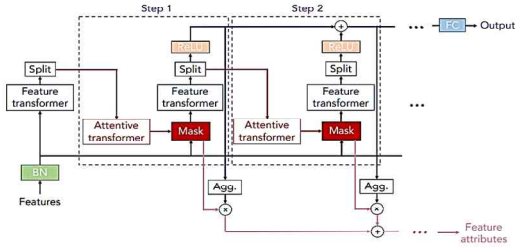

Fig. 8 TabNet encoder architecture

TabNet을 SSL로 사용하기 위한 단계는 비지도 학습을 사용하는 사전학습기인 TabNet-Pretrainer를 통해 Unlabeled data를 학습시키고 가중치 값을 Labeled data의 지도 학습 TabNet Classifier 모델에 반영하며 작동한다.

5.2.2 실험 과정

실험의 유효성 검증을 위하여 Labeled data를 8:2로 분리하여 학습 데이터와 시험 데이터를 준비하였다. 또한, 딥러닝의 Loss를 확인하기 위하여 학습 데이터를 데이터 증강(Data Augmentation) 한 후 다시 8:2로 분리하여 검증 데이터를 준비하였다.

Labeled data의 학습 데이터를 100% 사용한 SL_2와 50%를 사용한 SL_1을 기준으로 SL 데이터 셋을 분리하였다. Unlabeled data는 SL_2를 기준으로 각각 학습 데이터의 50% 크기를 Random sampling 하여 SSL_1, SSL_2, SSL_3의 데이터셋을 구축하였다. 또한 각 SSL_1, SSL_2, SSL_3에 사용한 Unlabeled data를 합친 총 학습 데이터의 150%인 SSL_concat 데이터 셋을 구축하였다.

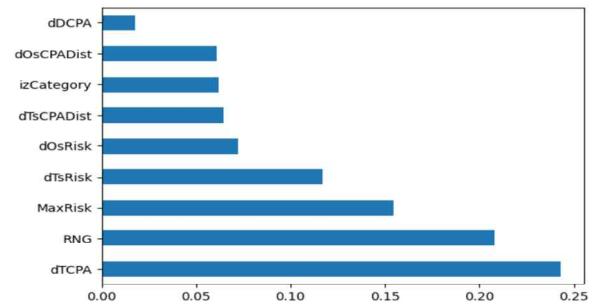

Fig. 9 TabNet feature importance graph

처음 분리한 Test data와 다음으로 분리한 Valid data의 정확도를 각각 구하기 위하여 SL_1, SL_2, SSL_1, SSL_2, SSL_3, SSL_concat을 동일한 파라미터로 고정하고 TabNet Classifier 모델을 학습하였다. 이때, 사용한 변수는 Labeled data의 양이 적은 관계로 SL_2를 기준으로 TabNet의 Feature Importance를 수행하여 선택하고 고정하였다(Fig. 9). 추가적으로, 이 모든 과정을 5회 반복하고 정확도를 측정하였다.

5.2.3 실험 결과

실험 결과를 통해 기존보다 많은 양의 Unlabeled data를 가지고 모델의 정확도를 올릴 수 있는 접근을 확인하였다. 또한, Unlabeled data의 양을 더욱 늘려 SSL 모델을 진행할 연구 가능성을 확인하였다(Table 6, Fig. 10).

Table 6 Prediction results of TabNet SL and SSL

Learning Method	Train		Test	Validation	Accuracy	
	L	U	L	L	Valid	Test
SL_1	4,744	-	1,186	1,482	0.71	0.64
SL_2	9,487	-			0.73	0.67
SSL_1	4,744	4,744			0.685	0.652
SSL_2	4,744	4,744			0.713	0.667
SSL_3	4,744	4,744			0.724	0.668
SSL_ Concat	4,744	14,232			0.716	0.656

* Labeled: L / Unlabeled: U

그러나 SSL_1, SSL_2, SSL_3의 경우 동일한 파라미터와 학습 방법 및 입력 변수를 사용했지만, Labeled data의 부족한 상황을 채워줄 수 있는 Unlabeled data의 품질에 따라 결과 간의 정확도가 두드러진 차이를 보였다.

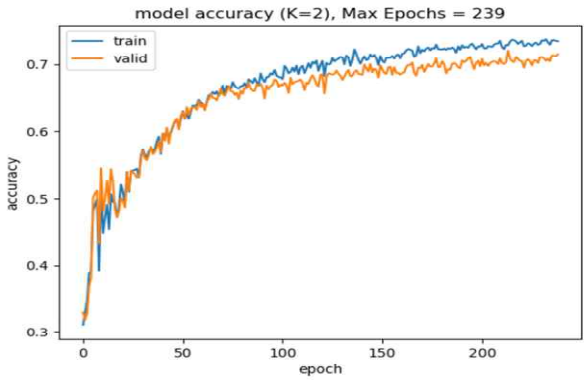

Fig. 10 Model learning curve graph

6. 모델 적용 결과

기존 서비스와의 비교를 위해서, 앞서 만든 모델 중 최적의 성능을 보여준 SSL_50_XGB 모델을 적용하였다.

모델 적용 비교는 2022년 10월 26일부터 27일까지 여수에서 수집한 기존 모델의 결과와 2023년 8월 16일부터 17일까지 여수에서 수집한 SSL_50_XGB모델의 결과이다. 이때, 실해역 시험 시나리오와 시험 지역은 Fig. 3 및 여수 돌산도 우측으로 동일하게 진행하였고 만족도 조사는 최대 5점부터 최소 1점까지 기록하였다. 2022년 실해역 시험 대상 선박은 Table 1의 No. 4 와 No. 5 이고 2023년 실해역 시험 대상 선박은 Table 2의 No. 1 과 No. 2 이다.

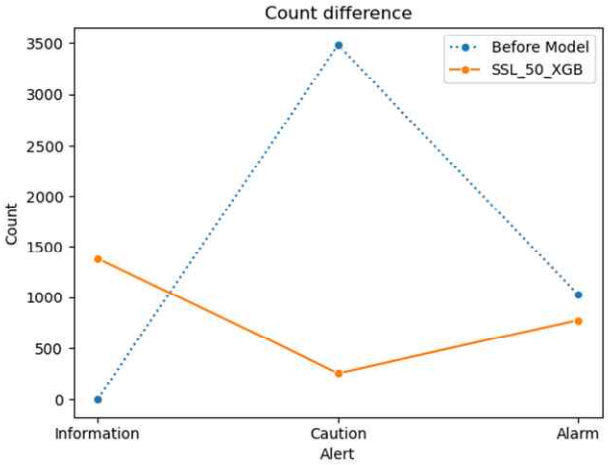

Fig. 11 Graph comparing the number of alerts in the before model and the SSL_50_XGB model for each alert

기존 모델의 만족도는 평균 2.2점으로 경보 발생 수는 총 4,513개가 발생하였다. 순서대로 관심(Information)은 2개, 주의(Caution)는 3,485개, 위험(Alarm)은 1,026개로 최대 3mile부터 경보 발생이 시작되었다. SSL_50_XGB 모델의 경우 평균 3.4점의 만족도를 보이고 경보 발생 수는 총 2,409개가 발생하였다. 순서대로 관심은 1,380개, 주의는 256개, 위험은 773개로 최대 2.5mile부터 경보 발생이 시작되었다(Fig. 11). 경보 발생 거리를 살펴보면, 단계 별 중앙값으로 관심이 1.78mile에서 0.8mile로 감소하였고, 주의는 0.96mile에서 0.4mile, 위험은 0.33mile에서 0.17mile로 감소한 결과를 확인하였다(Fig. 12).

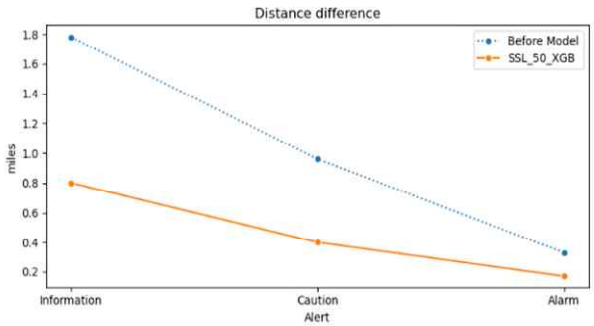

Fig. 12 Graph comparing the average of the first distance for each alert that occurred in the before model and SSL_50_XGB model

만족도 조사를 좀 더 자세하게 살펴보면, 조우상황별 정면(Head-On), 횡단(Crossing), 추월(Overtaking)의 만족도는 모델 적용 후 순서대로 1.2점, 1.8점, 1.2점 상승하였다(Fig. 13). 또한, 단계별 관심은 0.4점, 주의는 1.3점, 위험은 1.7점 상승하였다(Fig. 14).

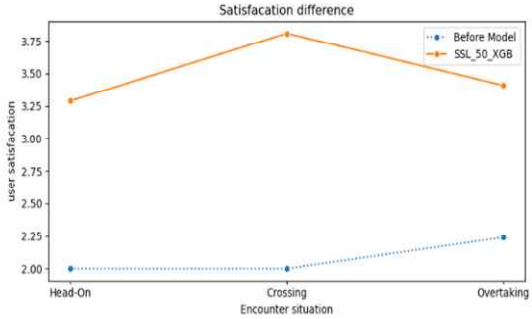

Fig. 13 Graph comparing user satisfaction in the before model and the SSL_50_XGB model for each encounter situation

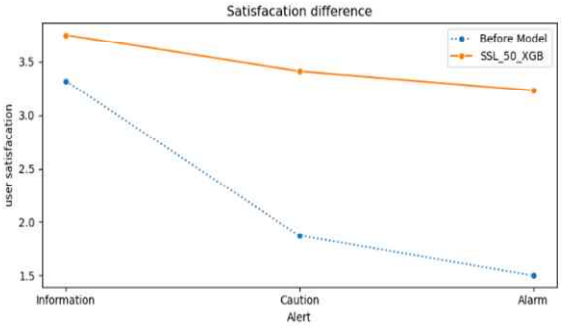

Fig. 14 Graph comparing user satisfaction in the before model and the SSL_50_XGB model for each alert

기존 모델에 비해서 SSL_50_XGB 모델 적용 후 주의에 집중되어 있던 경보가 분산되면서 총 알람의 수가 감소되고 사용자 만족도가 향상하는 것을 확인할 수 있었다. 하지만 관심의 경우 경보가 초기에 발생하다가 중지되고 다시 발생하는 상황들이 발생하였다. 이는 종속 변수가 0인 경우의 Label이 부족해서 나온 결과로 추정된다.

7. 결 론

본 논문에서는 소형선박 데이터 및 운항자의 의견이 충돌 경보 데이터에 반영되지 않아 운항자의 체감보다 먼 거리에서 경보가 발생하여 효용성이 떨어지는 문제를 해결하기 위해, 소형선박을 위한 충돌 경보 분류 모델을 연구하였다. 또한, 한정적인 수집 데이터에서 SL 방법의 한계를 극복하기 위해 SL과 SSL을 동일한 조건에서 비교하고 SSL을 위한 데이터 비율을 조절하여 Unlabeled data의 활용성을 확인하였다. 결과, SSL 모델들이 Labeled data를 수집하기 한정적인 상황에서 정확도를 올리기 위한 효과적인 대안이 될 수 있음을 확인했다.

배포 모델의 학습 과정에서 클래스 불균형 문제를 해결하기 위하여 Over Sampling SMOTE 기법을 사용하였고 준지도 학습을 위해 Label Propagation을 사용하였다. 다만, 이와 같은 방법들은 유사도 거리에 기반하고 있기 때문에 차원이 커질수록 한계성이 존재한다. 또한, 데이터가 증가함에 따라 제곱에 비례해서 연산량이 증가하기 때문에 효율성 관점에서 한계를 보인다.

딥러닝 모델인 TabNet은 지금과 같이 Labeled data가 적은 환경에서는 과적합의 위험을 고려하여 배포 모델로 사용하기에는 한계가 보인다. 모델의 Loss가 지속적이게 커지지 않으나, Labeled data의 양이 적어 Feature의 변화가 많고 Unlabeled data의 품질에 따라 결과 정확도의 차이가 많이 나는 것을 확인하였다. 따라서 지속적인 Labeled data 수집 환경과 반복적인 실험을 통해 Feature를 저장하고 데이터 및 Metric을 관리할 수 있는 모델 관리 체계가 필요하다.

방법론으론 소형선박 충돌 예측을 위하여 본 모델을 기반으로 다양한 Co-training 또는 Self-training 모델을 이용하는 방법을 적용해 볼 계획이다. 그리고 Unlabeled data가 지속적으로 업데이트됨에 따라, Unlabeled data의 데이터 마이닝을 통해 새로운 비지도학습(USL: UnSupervised Learning)등을 적용하는 모델로의 확장도 고려할 수 있다.

후 기

이 논문은 2023년도 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임(20210645, 지능형 해상교통정보 서비스 기반의 해상디지털 정보활용 기술개발).

References

[1] Baek, J. H., Cho, D. J. and Lim, K. H.(2022), “Design of Data Pipeline for Linkage the Intelligent Maritime Transport Information System”, J. Navig. Port Res, Vol. 2022, No. 1, pp. 315-316.

[2] Lee, J. H., Baek, S. B. and Lee, J. H.(2022), “ECS-based intelligent marine traffic information service usability and effectiveness assessment tool and method study”, J. Navig. Port Res, Vol. 2022, No. 1, pp. 310-312.

[3] Ma, J., Jia, C., Yang, X., Cheng, X., Li, W. and Zhang, C.(2020), “A Data-Driven Approach for Collision Risk Early Warning in Vessel Encounter Situations Using Attention-BiLSTM”, in IEEE Access, Vol. 8, pp. 188771-188783.

[4] Ministry of Oceans and Fisheries(2023),“ Implementation Plan for Intelligent Maritime Traffic Information Service”, pp. 3-5., https://www.mof.go.kr/

[5] Olivier, D., Yoshua, B. and Nicolas, L. R.(2005), “Efficient Non-Parametric Function Induction in Semi-Supervised Learning”, Proceedings of the Tenth International Workshop on AISTAT 2005, pp. 96-103.

[6] Park, M. J., Park, Y. S., Lee, M. K., Kim, D. W. and

Kim, N. E.(2021), “A Study on the Improvement of Collision Prevention Algorithm for Small Vessel Based on User Opinion”, Journal of the Korean Society of Marine Environment and Safety, Vol. 27, No. 2, pp. 238-246.

[7] Seok, H. J., Sim, S., Woo, J. H., Cho, J. R., Cho, D. J., Baek, J. H. and Jung, J.(2023), “A Study on the Prediction of Ship Collision Based on Semi-Supervised Learning”, J. Navig. Port Res, Vol. 2023, No. 1, pp. 204-205.

[8] Sercan O. Arik. and Tomas P.(2021), “TabNet: Attentive Interpretable Tabular Learning”, The 35th AAAI Conference on Artificial Intelligence, Vol. 35, No. 8, pp. 6679-6687.

[9] Yang, Y. H.(2021), “A Study on the Development of Ship Collision Risk Informed Classification Model for Using Logistic Regression”, Mokpo National Maritime University, Department of Maritime Transportation System, PhD Dissertation.

[10] Yang, Y. H. and Oh, S. W.(2016), “Implementation of core technology korean e-Navigation service for the prevention of maritime accidents”, The Korean Association of Maritime Police Science, Vol. 6 No. 2, pp. 63-76.

[11] Yoshua, B., Olivier, D. and Nicolas, L. R.(2006), “In Semi-Supervised Learning”, MIT Press, pp. 193-216.

Received 17 November 2023
Revised 27 November 2023
Accepted 15 December 2023