준지도 학습 기반 선박충돌 예측에 대한 연구

석호준* · 심 승* · 우정훈* · 조준래* · 조득재** · 백종화** · † 정재룡

*,† 슈어소프트테크㈜ DX 신사업실, **선박해양플랜트연구소 해사디지털서비스연구센터

A Study on the Prediction of Ship Collision Based on Semi-Supervised Learning

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

**,† Suresoft Technologies Inc., Seongnam 13453, Korea*
***Korean Research Institute of Ships & Ocean engineering, Daejeon, 34103, Korea*

요 약 : 본 연구는 준지도학습(SSL)을 기반한 소형 어선의 충돌 경보 송출 예측 모델에 관한 연구이다. 지도학습(SL) 방법은 레이블링된 다수의 데이터가 필요하지만 레이블링 과정에서 많은 자원과 시간이 소요된다. 본 연구는 '지능형 해상교통정보 서비스'와 연계한 데이터 파이프라인을 통해 수집된 서비스 데이터와 실해역 시험에서 수집한 데이터를 사용하였다. 실제 사용자 만족도 기반으로 레이블이 결정된 실해역 시험 데이터만 아니라 레이블이 결정되지 않은 서비스 데이터를 함께 학습시킨 결과, 모델 정확도가 향상되었다.

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

Abstract : *This study studied a prediction model for sending collision alarms for small fishing boats based on semi-supervised learning (SSL). The supervised learning (SL) method requires a large number of labeled data, but the labeling process takes a lot of resources and time. This study used service data collected through a data pipeline linked to 'intelligent maritime traffic information service' and data collected from real-sea experiment. The model accuracy was improved as a result of learning not only real-sea experiment data with labeling determined based on actual user satisfaction but also service data without label determined together.*

Key words : *Maritime digital, Big-Data, Data Pipeline, Data Preprocessing, Semi-Supervised Learning, Machine Learning*

1. 서 론

해양수산부는 해양사고를 예방하기 위해 지능형 해상교통정보 서비스(이하 e-Nav서비스)를 구축하여 제공하고 있다.

더불어, KRISO 통합시험센터에 데이터 파이프라인을 설계하여 서비스 데이터를 분석할 수 있는 환경이 구축되었다. (백종화, 2022). ····(중략)····

선박의 충돌 알고리즘 모델 선행 연구의 대부분은 시뮬레이션과 실해역 시험 등 레이블이 결정된 데이터(Labeled Data)만으로 지도학습(Supervised Learning 이하 SL) 방법을 수행해왔지만, 본 연구에서는 레이블이 결정되지 않은 데이터(UnLabled Data)들을 포함한 준지도학습(Semi-Supervised Learning 이하 SSL) 모델을 연구하였다.

2. 데이터 파이프라인

수집된 서비스 데이터는 1일 분량을 정제 후 서버에 업로드 하고 있어, 데이터 연계 구성도(Fig. 1)와 같은 1일 단위의 배치처리 방식을 사용한다. ····(중략)····

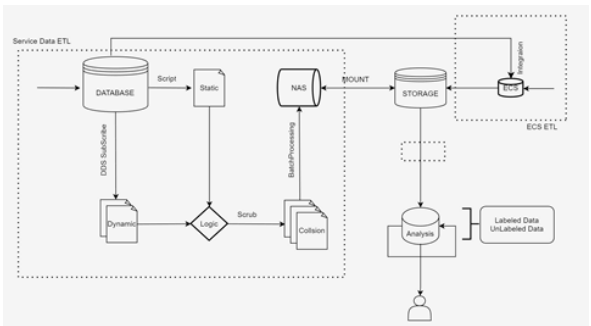

Fig 1. Data Link Diagram

실해역 시험 데이터의 경우 로컬 내부에 수집되는 데이터와 파이프라인을 연계하여 Raw Data를 혼합 전처리 후 분석 서버로 전송된다.(중략)·····..

3. 데이터 전처리

실해역 시험의 상황과 유사한 서비스 데이터를 추출하고 준지도 학습에 사용하기 위하여 다음과 같은 2차 정제 과정을 거친다.(Fig. 2)····(중략)····

Fig 2. Data Preprocessing Flow

4. 준지도 학습

준지도 학습은 Labeled Data의 독립 변수(X_i) 및 종속 변수(Y_i)와 UnLabeled Data의 독립 변수(X_j)를 함께 사용하여 존재하지 않는 종속 변수($\hat{Y}$)를 예측하는 방법론이다.

본 연구에서는 그래프 기반 준지도학습 Label Propagation을 적용하였다.

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

Table 1. Label Propagation

Label Propagation은 변수 간의 유사성을 기반으로, 레이블이 결정된 데이터에서 레이블이 결정되지 않은 데이터로 레이블을 전파함으로써 작동한다. 종속 변수 Y_i를 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. 유사성 Matrix가 계산되면 반복적으로 종속 변수 $\hat{Y}$에 레이블을 부여한다. 각 인접 데이터 포인트의 레이블 가중평균을 계산하고 수렴될 때까지 반복한다.(Table 1)····(중략)····

Learning Method		SL_1	SL_2	SSL_10	SSL_30	SSL_50	SSL_70
Train	Labeled	4,833	9.667	4,833	4,833	4,833	4,833
	UnLabled	-	-	483	1,450	2,417	3,384
Test	Labeled	2,417					
	Accuracy	0.613	0.628	0.618	0.620	0.623	0.623

Table 2. Labeled Data Validation-Test on SSL

준지도 학습의 유효성을 검증하기 위하여 Labeled Data를 학습, 테스트, 검증 데이터로 분리하였다. DecsionTree 분류기를 통한 SL 결과와 SSL의 비율(10%, 30%, 50%, 70%)의 결과를 다음과 같이 확인 하였다.(Table 2) 결과를 통해 Labeled Data를 충분하게 확보 할 수 없는 상황에서 SSL접근은 하나의 대안이 될 수 있음을 확인하였다. ····(중략)····

정제한 UnLabeled Data를 추가하여 XGB 분류기를 수행하였다. Cross Validation을 5회 수행하여 평균 Accuracy, Recall, Precision, F1-Score를 비교하였다. (Fig 3)····(중략)····

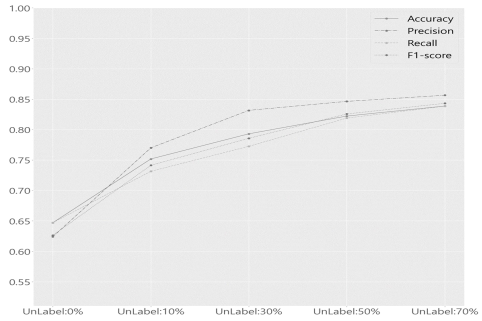

Fig 3. Comparing Cross Validation Score on XGB

5. 결 론

본 논문에서는 SSL 모델이 Labeled Data 수집이 한정되는 상황에서 효과적인 대안이 될 수 있음을 확인했다.

클래스 불균형 문제를 해결하기 위하여 Over Sampling SMOTE 기법을 사용하였고 준지도 학습을 위해 Label Propagation을 사용하였다. 이러한 접근은 유사도 거리에 기반하고 있기 때문에 고차원 데이터에 적합하지 않다는 점과 관측치 수가 증가함에 따라 제곱에 비례해서 연산량이 증가하기 때문에 효율성 관점에서 한계를 보인다.

따라서, 향후 연구 방향은 선박충돌 예측을 위하여 본 모델을 기반으로 다양한 Co-training 또는 Self-training 모델을 이용하는 방법을 적용해 볼 계획이다.

또한, UnLabeled Data의 데이터 마이닝을 통해 새로운 비지도 학습(USL: UnSupervised Learning)등을 적용하는 모델로의 확장도 고려할 수 있다.

감 사 의 글

본 논문은 해양수산부와 해양수산과학기술진흥원의 지원을 받아 수행하는 '지능형 해상교통정보 서비스 기반의 해상디지털 정보활용 기술개발'에 의해 수행되었습니다.

참 고 문 헌

[1] 슈어소프트테크㈜ (2022), 해상디지털 통합활용연계 기술개발 연구개발계획서

[2] 백종화(2022), 지능형 해상교통정보시스템 연계를 위한 데이터파이프라인설계, 2022 한국항해항만학회 춘계학술대회