논문 22-47-02-21 The Journal of Korean Institute of Communications and Information Sciences '22-02 Vol.47 No.02
https://doi.org/10.7840/kics.2022.47.2.387

최적 증강: IoST에서 Terrestrial-CubeSat 간 핸드오버를 위한 기계학습 기반 단말기 이동성 예측 알고리즘

오 준 석•, 이 동 현*, 하 태 윤*, 이 윤 성*, 조 성 래°

Optimal Augmentation: Machine Learning-Based Terminal Mobility Prediction Algorithm for Handover between Terrestrial-CubeSat in IoST

Junsuk Oh•, Donghyun Lee*, Taeyun Ha*, Yunseong Lee*, Sungrae Cho°

요 약

IoST(Internet of Space Things)에서는 단말기 이동성에 의해 Terrestrial-CubeSat 간의 핸드오버와 함께 링크를 재설정하기 위한 오버헤드가 발생한다. 이와 같은 오버헤드를 처리하는 과정에서 핸드오버가 실패할 경우, IoST는 링크 단절과 함께 통신의 안정성을 잃게 된다. 이에 따라 본 논문에서는 핸드오버를 효율적이고 안정적으로 지원하기 위해 단말기 이동성 예측 문제를 다중 분류 문제로 정의하는 그리드 레이블링(Grid Labeling) 알고리즘과 기계학습 모델의 성능을 최적화하는 데이터 최적 증강 알고리즘을 제안한다.

Key Words : Data Augmentation, IoST, CubeSat, Handover, Machine Learning

ABSTRACT

In IoST, terminal mobility causes overhead to reset links along with handover between terrestrial-cubesat. If handover fails in the process of processing such overhead, IoST loses the stability of communication along with link disconnection. Accordingly, in order to efficiently and stably support handover, this paper proposes grid labeling algorithm, which defines the terminal mobility prediction problem as a multiclass classification problem, and a data optimal augmentation algorithm that optimizes the performance of machine learning models.

※ 본 연구는 한국 전력 공사의 2019년 선정 기초연구개발 과제(R19XO01-41) 지원 및 중앙대학교 관리로 수행되었습니다.
※ 본 연구는 한국 전력 공사의 2019년 선정 기초연구개발 과제 연구비에 의해 지원되었음. (과제번호 : R19XO01-41)
◆ First Author : Chung-Ang University Department of Computer Science and Engineering, jsoh@uclab.re.kr, 학생회원
° Corresponding Author : Chung-Ang University Department of Computer Science and Engineering, srcho@cau.ac.kr, 종신회원
* Chung-Ang University Department of Computer Science and Engineering, dhlee@uclab.re.kr, 학생회원; tyha@uclab.re.kr, 학생회원; yslee@uclab.re.kr, 학생회원
논문번호 : 202110-260-B-RE, Received September 30, 2021; Revised November 30, 2021; Accepted November 30, 2021

Ⅰ. 서 론

IoT(Internet of Things)는 무선 통신으로 각종 사물을 연결하여 글로벌 유비쿼터스 플랫폼을 실현할 수 있는 차세대 무선 시스템의 주요한 일부분으로서 언제 어디서나 작동할 수 있는 특성, 무수한 물리적 지점 간의 연결을 제공할 수 있는 애플리케이션 지향적인 특성 등을 가진다[1]. 그러나 IoT의 실질적인 실현은 연결되는 단말기 수의 기하급수적인 증가와 함께 통신 서비스 커버리지를 제공할 수 없는 대기, 대양, 사막, 극지, 우주와 같은 영역의 존재로 인해 어려움을 겪는다. 이를 위해 지구 저궤도 위성 기반의 IoS(Internet of Space)가 제안되었으나, 긴 개발 일정과 높은 개발 비용 등으로 인해 IoT는 또 다른 어려움을 겪는다[2,3]. 이에 따라 지상, 대기, 우주에 걸친 새로운 유비쿼터스 사이버-물리적 시스템으로 정의되는 CubeSat 기반의 IoST가 실현 가능한 대안으로 인식되고 있다[4,5]. IoST는 그림 1과 같이 실시간으로 사용할 수 있는 위성 백홀 네트워크를 바탕으로 위성에서 수집한 데이터와 지상 데이터를 통합하여 새로운 애플리케이션을 제공함으로써 전통적인 IoT의 기능을 확장한다.

IoST에서 글로벌 위성 네트워크를 구축하기 위한 핵심 기술인 CubeSat은 부피 1ℓ(10㎝×10㎝×10㎝), 질량 1.33㎏을 초과하지 않는 입방형 구조의 초소형 인공위성을 의미한다[6]. 낮은 개발 비용, 낮은 전력 소비 등과 함께 CubeSat의 주요 장점 중 하나인 동시성은 할당된 지역만을 대상으로 임무를 수행하는 대형 인공위성과는 다르게 지구 전체 또는 우주 전체 방향을 대상으로 특정 임무를 동시다발적으로 수행할 수 있음을 의미한다. 그러나 이와 대조적으로 차량, 선박, 항공기와 같이 유동적인 이동성을 가진 단말기가 지상망을 거치지 않고 CubeSat과 직접 통신하면 잦은 핸드오버와 함께 링크를 재설정하기 위한 수많은 오버헤드가 발생하게 된다는 단점도 존재한다. IoST는 이러한 오버헤드를 처리하는 과정에서 적은 네트워크 자원으로 인해 예약 지연이 발생하게 되고, 이에 따라 핸드오버가 실패하게 되면 링크 단절과 함께 통신의 안정성을 잃게 된다. 이와 같은 불안정한 IoST의 안전성을 높이기 위해서는 단말기의 이동성을 예측함으로써 핸드오버를 효율적이고 안정적으로 지원해야 한다[7].

핸드오버는 그림 2와 같이 단말기가 기존에 연결된 기지국의 서비스 공간에서 다른 기지국의 서비스 공간으로 이동할 경우, 단말기가 다른 기지국의 서비스 공간에 할당된 채널에 동조하여 서비스가 연결되는 기능을 의미한다. 최근 핸드오버는 신호 강도의 크기에 의존하는 전통적인 핸드오버와는 다르게 기계학습을 기반으로 각 단말기가 경험하는 QoE(Quality of Experience)를 학습하여 효율적이고 안정적인 핸드오버를 수행한다[8,9]. 이러한 기계학습을 바탕으로 단말기의 이동성을 예측하기 위해서는 모델의 학습과 높은 성능을 위해 수많은 양의 데이터 집합이 요구된다. 그러나 일반적으로 기계학습을 적용하기 위해 수많은 양의 데이터 집합을 수집하는 일에는 큰 어려움이 존재한다. 더하여 제한된 양의 데이터 집합으로 모델을 학습할 경우, 과소적합 또는 과대적합 현상이 발생하게 된다. 이를 위해 데이터 증강은 제한된 양의 데이터 집합에서 특정 샘플링 알고리즘으로 데이터 특성을 반영하거나 원본 샘플의 확률 등을 반영한 유사 샘플을 생성하여 수많은 양의 데이터 집합을 확보한다[10].

본 논문에서는 핸드오버를 효율적이고 안정적으로 지원하기 위해 제한된 양의 단말기 이동성 데이터 집합에 데이터 증강을 적용함으로써 기계학습 기반 단말기 이동성 예측 모델의 성능을 최적화하는 데이터 최적 증강 알고리즘을 제안한다. 제안하는 알고리즘은 데이터 집합에서 연속적인 수치형 샘플 간의 상관관계를 바탕으로 최적의 증강률을 결정하여 데이터 증강을 적용한다. 더하여 수집한 데이터 집합에 적합한

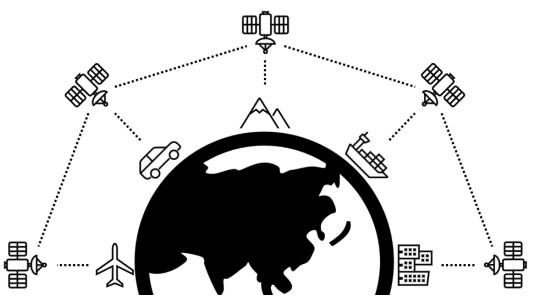


그림 1. CubeSat을 기반으로 하는 IoST
Fig. 1. IoST based on CubeSat

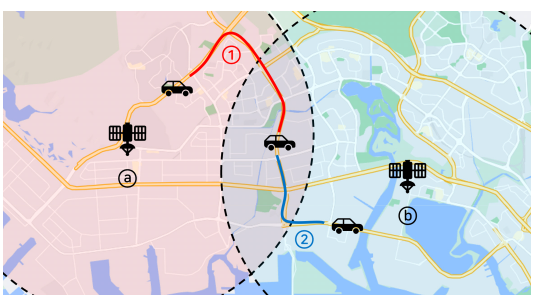


그림 2. 단말기 이동성에 의한 핸드오버
Fig. 2. Handover due to terminal mobility

기계학습 모델을 선택하기 위해 단말기 이동성 예측 문제를 다중 분류 문제로 정의하는 그리드 레이블링 알고리즘을 함께 제안한다. 제안하는 알고리즘은 특정 범위의 지리 좌표계를 그리드 좌표계로 변환한 후 각 그리드 셀의 클래스 레이블(Class Label)을 결정한다.

본 논문의 나머지는 다음과 같이 정리한다. Ⅱ는 그리드 레이블링 알고리즘을 설명한다. Ⅲ은 단말기 이동성 예측 모델을 선택한다. 데이터 최적 증강 알고리즘은 Ⅳ에서 설명한다. Ⅴ는 실험과 결과를 분석한다. 마지막으로 결론은 Ⅵ에서 제시한다.

Ⅱ. 그리드 레이블링 알고리즘

단말기 이동성 예측 모델의 학습과 성능 평가를 위한 데이터 집합으로 NUS(National University of Singapore) 셔틀버스의 이동성 데이터 집합을 수집한다. 데이터 집합에서 단말기의 일련번호, GPS 날짜, GPS 시간, 위도, 경도, 속도, 방향은 데이터 특성을 의미한다. 각 데이터 특성에 대한 매개변수는 표 1에서 정의한다.

단말기 이동성을 예측하기 위해서는 그림 3에 표시된 특성 x, y를 종속 변수로 사용하고, 나머지 특성을 독립 변수로 사용해야 한다. 그러나 특성 x, y는 $m \times 2$ 행렬이므로 $m \times 1$ 행렬의 형태를 가지는 종속 변수로 사용할 수 없다. 이를 위해 본 논문에서는 특정 범위의 지리 좌표계에 대해 레이블링을 수행하는 그리드 레이블링 알고리즘을 제안한다. 제안하는 알고리즘은 다음과 같이 정의하고, 각 매개변수는 표 2에서 정의한다.

$$0 \le x_a < x_b \le 180, \forall x_a, x_b \in R \tag{1}$$

$$0 \le y_a < y_b \le 90, \forall y_a, y_b \in R \tag{2}$$

표 1. 데이터 집합의 특성과 매개변수
Table 1. Features and parameters of the dataset

Feature	Parameter
Serial number of the terminal	n
GPS date of the terminal	d
GPS time of the terminal	t
The longitude of the terminal	x
The latitude of the terminal	y
Speed of the terminal	s
Heading of the terminal	h

표 2. 그리드 레이블링의 매개변수와 정의
Table 2. Parameters and definitions of grid labeling

Definition	Parameter
Minimum and maximum longitude	x_a, x_b
Minimum and maximum latitude	y_a, y_b
Number of rows and columns	x_e, y_e
Index in grid coordinate system	i, j
Class label for each grid cell	l

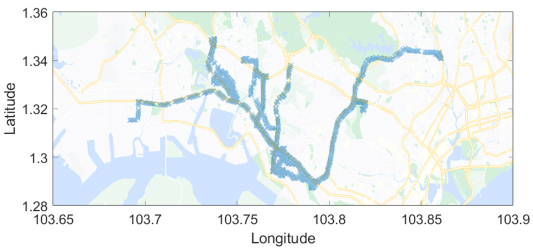


그림 3. 데이터 집합의 경도와 위도
Fig. 3. The longitude and latitude of the dataset

수식 (1)과 (2)는 x와 y가 각각 지리 좌표계에서 가질 수 있는 최솟값과 최댓값의 범위를 정의한다. 따라서 $[x_a, x_b]$과 $[y_a, y_b]$의 지리 좌표계를 특정 범위의 지리 좌표계로 정의한다.

$$1 \le x_e, \forall x_e \in \mathrm{N} \tag{3}$$

$$1 \le y_e, \forall y_e \in \mathrm{N} \tag{4}$$

수식 (3)과 (4)는 특정 범위의 지리 좌표계를 그리드 좌표계로 변환하기 위한 행의 개수 x_e와 열의 개수 y_e의 최솟값을 정의한다. 따라서 그리드 좌표계에서 그리드 셀의 개수는 1개 이상이다.

$$i = 0, 1, \cdots, x_e \tag{5}$$

$$j = 0, 1, \cdots, y_e \tag{6}$$

$$1 \le l \le x_e \times y_e \tag{7}$$

수식 (5)와 (6)이 그리드 좌표계에서 각 선분의 색인을 정의함에 따라 수식 (7)은 그리드 좌표계에서 각 그리드 셀이 가질 수 있는 클래스 레이블 l의 범위를 정의한다.

$$g(i,j) = x_e \times j + (i+1) = l \tag{8}$$

수식 (8)은 꼭짓점이 (i,j), $(i,j+1)$, $(i+1,j)$, $(i+1,j+1)$인 그리드 셀의 클래스 레이블을 결정하는 함수 $g(i,j)$를 정의한다. 따라서 그리드 레이블링 알고리즘은 클래스 레이블 l을 특성 x, y 대신에 종속 변수로 사용한다. 그림 4는 x_a, x_b, y_a, y_b, x_e, y_e가 순서대로 103.75, 103.80, 1.30, 1.32, 5, 2인 경우 각 그리드 셀의 클래스 레이블이 결정된 결과를 의미한다. 이에 따라 수집한 데이터 집합에 대해 그리드 레이블링 알고리즘을 적용한 결과와 각 클래스 레이블의 샘플 수는 그림 5와 같다.

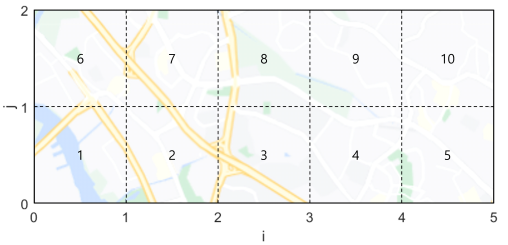


그림 4. 그리드 레이블링에 의한 그리드 셀의 클래스 레이블
Fig. 4. Class label of grid cell by grid labeling

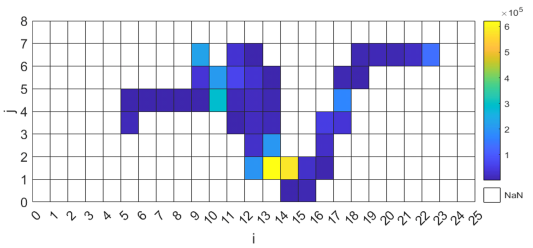


그림 5. 데이터 집합의 클래스 레이블과 샘플 수
Fig. 5. Class labels and number of samples in the dataset

Ⅲ. 단말기 이동성 예측 모델

일반화는 기계학습 모델이 학습 과정에서 사용한 데이터 집합을 제외하고 새로운 데이터 집합에 대해 성능 평가를 수행함으로써 측정한다. 이를 위해 본 논문에서는 단말기 이동성 데이터 집합을 표 3과 같이 70%의 훈련 데이터 집합과 30%의 테스트 데이터 집합으로 분리한다. 이에 따라 단말기 이동성 예측 문제는 단말기 수가 1개인 경우에도 클래스 레이블 수가 3개 이상이고, 종속 변수 l의 값이 불연속적인 형태이므로 다중 분류 문제로 정의한다.

3.1 평가 지표

오차 행렬(Confusion Matrix)은 양성과 음성 클래스에 대한 양성 예측과 음성 예측을 TP(True Positive), FP(False Positive), FN(False Negative), TN(True Negative)으로 정의한다.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{9}$$

정확도는 오차 행렬을 바탕으로 수식 (9)와 같이 정의한다. 그러나 양성 클래스와 음성 클래스 중 하나가 다른 하나보다 비중이 높은 불균형한 데이터 집합에서 정확도는 분류 모델에 대한 평가 지표로 적합하지 않다. 따라서 분류 문제는 정밀도와 재현율의 조화 평균인 F1-Score를 사용하고, 오차 행렬을 바탕으로 수식 (10)과 같이 정의한다.

$$\text{F1}-\text{Score} = \frac{\text{TP}}{\text{TP}+\frac{1}{2}(\text{FP}+\text{FN})} \tag{10}$$

다중 분류 문제의 F1-Score는 한 클래스를 양성 클래스로 간주하고, 나머지를 음성 클래스로 간주하여 각 클래스에서 산출한 모든 F1-Score의 평균으로 정의한다. 이에 따라 본 논문에서는 단말기 이동성 예측 모델의 평가 지표로 F1-Score를 사용한다.

표 3. 단말기 수에 따른 클래스 레이블과 샘플 수
Table 3. The number of class labels and samples based on the number of terminals.

The number of terminals	The number of samples in the training dataset	The number of samples in the test dataset	The number of class labels
1	4143	1776	13
3	12466	5345	23
5	20718	8884	34
7	28937	12408	41
9	37117	15914	41

3.2 손실 함수

다중 분류 문제는 모델이 테스트 샘플에 대해 예측한 클래스가 정확한 클래스임을 확신하는 정도가 중요하다. 이를 위해 다중 분류 문제는 손실 함수로 Log-Loss를 사용하고, 수식 (11)과 같이 정의한다.

$$\text{Log}-\text{Loss}=-\frac{1}{\Phi}\sum_{\phi=1}^{\Phi}\sum_{\psi=1}^{\Psi}log\left(p_{\phi,\psi}\right)\times q_{\phi,\psi} \quad (11)$$

수식 (11)에서 Φ, Ψ, $p_{\phi,\psi}$, $q_{\phi,\psi}$는 순서대로 테스트 샘플 수, 클래스 레이블 수, ϕ번째 테스트 샘플에 대해 클래스 레이블 ψ로 예측할 확률, ϕ번째 테스트 샘플에 대해 클래스 레이블 ψ가 올바른지를 의미한다. 이에 따라 본 논문에서는 단말기 이동성 예측 모델의 불확실성을 추정하기 위한 손실 함수로 Log-Loss를 사용한다.

3.3 교차 검증

교차 검증은 일반화 성능을 측정하기 위한 안정적이고 효율적인 통계적 평가 방법이다. 일반적인 교차 검증은 그림 6의 (a)와 같이 데이터 집합을 반복해서 분리하고 여러 모델을 학습하는 k-겹 교차 검증이 존재한다. 그러나 k-겹 교차 검증은 데이터 집합을 순서대로 k개의 폴드로 분리함에 따라 일부 폴드에는 클래스 레이블 비율이 불균형할 수 있다. 이에 따라 본 논문에서는 그림 6의 (b)와 같이 모든 폴드에서 클래스 레이블 비율이 데이터 집합의 클래스 레이블 비율과 같도록 분리하는 계층별 k-겹 교차 검증을 사용한다.

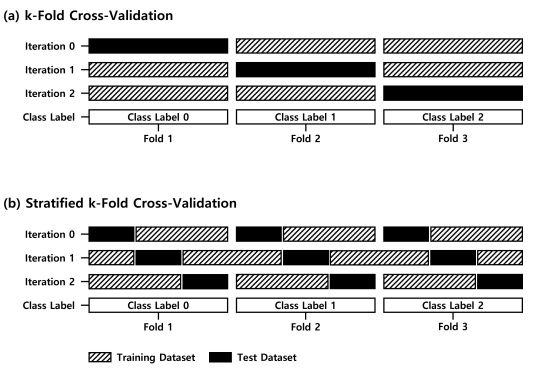


그림 6. 다양한 교차 검증 방법
Fig. 6. Various cross-validation methods

3.4 모델 선택

본 논문에서는 가능한 모든 다중 분류 모델과 데이터 스케일링 쌍을 비교하고, 단말기 이동성 데이터 집합에 가장 적합한 모델과 스케일링을 선택한다. 이를 위해 다중 분류 모델 후보군과 각 모델의 특징을 표 4에 정리하고, 데이터 스케일링 후보군과 각 스케일링의 특징을 표 5에 정리한다.

단말기 수가 서로 다른 각 데이터 집합에 대해 각 다중 분류 모델에서 F1-Score가 가장 높기 위한 최적의 스케일링과 Log-Loss가 가장 낮기 위한 최적의 스케일링을 표 6에 정리한다.

F1-Score에서 스케일링은 데이터 집합과 모델에 상관없이 정규화의 빈도가 가장 높고, 로버스트 스케일링의 빈도가 가장 낮은 것을 알 수 있다. 더하여 모든 데이터 집합에 대해 랜덤 포레스트가 평균 0.9147로 가장 높고, 가우시안 나이브 베이즈가 평균 0.3877로 가장 낮은 것을 알 수 있다.

Log-Loss에서 스케일링은 데이터 집합과 모델에 상관없이 스탠다드 스케일링의 빈도가 가장 높고, 최소-최대 스케일링의 빈도가 가장 낮은 것을 알 수 있다. 더하여 모든 데이터 집합에 대해 랜덤 포레스트가

표 4. 다양한 다중 분류 모델
Table 4. Various multiclass classification models

Model	Characteristics
k-NN	• Basic model and good for small dataset.
Logistic Regression	• Available for large dataset. • Available for high-dimensional data.
Gaussian NB	• Faster but less accurate than linear models. • Available for large dataset. • Available for high-dimensional data.
Random Forest	• Stable and powerful. • Data scaling is not required. • Impossible for high-dimensional sparse data.
Kernel-SVM	• Good for medium-sized datasets consisting of similar meaning features. • Sensitivity to parameters.

표 5. 다양한 데이터 스케일링
Table 5. Various data scaling

Data Scaling	Characteristics
Standard Scaling	• The mean is 0, and the variance is 1. • The entire features have the same scale. • No limitations on the minimum and maximum values.
Robust Scaling	• The entire features have the same scale. • Medium and quartile are used. • Not affected by outliers.
Min-Max Scaling	• The entire feature value is located between 0 and 1.
Normalization	• The Euclidean length of the feature vector becomes 1. • Direction is more important than the length of the feature vector.

표 6. 단말기 수와 다중 분류 모델에 대한 최적의 스케일링
Table 6. Optimal scaling for the number of terminals and multiclass classification models

Model	Optimal Scaling, F1-Score, Log-Loss	The number of terminals				
		1	3	5	7	9
k-NN	Optimal Scaling	Normalization	Robust Scaling	Min-Max Scaling	Normalization	Normalization
	F1-Score	0.7173	0.7384	0.7896	0.7951	0.7847
	Optimal Scaling	Normalization	Robust Scaling	Robust Scaling	Min-Max Scaling	Robust Scaling
	Log-Loss	3.7548	2.7304	2.1660	2.3639	2.4196
Logistic Regression	Optimal Scaling	Standard Scaling	Standard Scaling	Min-Max Scaling	Min-Max Scaling	Min-Max Scaling
	F1-Score	0.4454	0.5091	0.5581	0.3720	0.3661
	Optimal Scaling	Standard Scaling	Standard Scaling	Standard Scaling	Standard Scaling	Standard Scaling
	Log-Loss	1.7167	1.3270	1.2779	1.8048	1.8779
Gaussian NB	Optimal Scaling	Normalization	Normalization	Normalization	Normalization	Normalization
	F1-Score	0.4566	0.4221	0.5033	0.2596	0.2969
	Optimal Scaling	Normalization	Normalization	Normalization	Normalization	Normalization
	Log-Loss	2.0568	1.9960	2.1560	3.1644	2.8897
Random Forest	Optimal Scaling	-	-	-	-	-
	F1-Score	0.8975	0.9033	0.9260	0.9273	0.9194
	Optimal Scaling	-	-	-	-	-
	Log-Loss	0.4415	0.3495	0.2481	0.2401	0.2781
Kernel-SVM	Optimal Scaling	Standard Scaling	Standard Scaling	Standard Scaling	Standard Scaling	Standard Scaling
	F1-Score	0.5355	0.6073	0.6386	0.5957	0.5092
	Optimal Scaling	Standard Scaling	Standard Scaling	Standard Scaling	Standard Scaling	Standard Scaling
	Log-Loss	1.3493	1.0615	0.9694	1.0570	1.3051

평균 0.3115로 가장 낮고, k-최근접 이웃이 평균 2.6869로 가장 높은 것을 알 수 있다.

본 논문에서는 모든 다중 분류 모델과 데이터 스케일링 쌍에서 F1-Score가 가장 높고, Log-Loss가 가장 낮은 스케일링이 없는 랜덤 포레스트를 단말기 이동성 예측 모델로 사용한다.

Ⅳ. 데이터 최적 증강 알고리즘

단말기의 수가 다른 각 데이터 집합에 대해 랜덤 포레스트를 사용한 단말기 이동성 예측 모델은 평균 0.9147의 평가 지표와 평균 0.3115의 손실 함수를 보인다. 그러나 IoST에서 제한된 양의 데이터 집합으로 인해 단말기 이동성 예측이 잘못되면 핸드오버가 실패하고, IoST는 통신의 안정성을 잃게 된다. 이를 위해 본 논문에서는 그림 7과 같이 연속적인 샘플 간의 상관관계를 바탕으로 연속적인 원본 샘플 사이에 유사 샘플을 생성함으로써 제한된 양의 데이터 집합에서 수많은 양의 데이터 집합을 확보하는 데이터 최적 증강 알고리즘을 제안한다. 제안하는 알고리즘은 다음과 같이 정의하고, 원본 샘플과 유사 샘플에 대해 각각 아래 첨자 o와 s를 표기한다. 각 매개변수는 표 7에서 정의한다.

$$n_s = n_o \tag{12}$$

수식 (12)는 제안하는 알고리즘이 같은 단말기에 대해서만 적용됨을 의미한다. 따라서 생성하는 유사

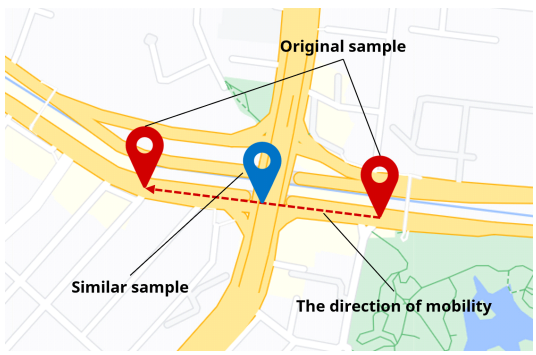


그림 7. 단말기 이동성 데이터 샘플에 대한 증강
Fig. 7. Augmentation for terminal mobility data samples

샘플의 n_s는 원본 샘플의 n_o로 정의한다.

$$2 \le w_a \le w \le w_b, \forall w \in \mathrm{N} \tag{13}$$

$$u = 1,2,\cdots,w-1 \tag{14}$$

수식 (13)은 o번째와 $o+1$번째 원본 샘플 사이에 생성할 유사 샘플의 개수를 결정하는 w의 범위를 정의한다. 따라서 생성하는 유사 샘플의 개수는 1개 이상이다. 더하여 특정 원본 샘플 사이에 생성하는 유사 샘플의 색인을 의미하는 u는 수식 (13)으로부터 수식 (14)와 같이 정의한다.

$$d = 0,1,\cdots \tag{15}$$

$$0 \le t \le 1, \forall t \in \mathrm{R} \tag{16}$$

$$z_o = d_o + t_o \tag{17}$$

$$z_{o+1} = d_{o+1} + t_{o+1} \tag{18}$$

수식 (15)부터 (18)은 원본 샘플의 d와 t를 z로 변환하고, z의 정수부와 소수부가 각각 d와 t임을 정의한다. 따라서 z_{o+1}는 z_o보다 크거나 같다.

$$z_s = \frac{u}{w} \times z_o + \frac{w-u}{w} \times z_{o+1} \tag{19}$$

$$d_s = \lfloor z_s \rfloor \tag{20}$$

$$t_s = z_s - \lfloor z_s \rfloor \tag{21}$$

수식 (19)는 o번째와 $o+1$번째 원본 샘플의 비율로 생성하는 유사 샘플의 z_s를 정의한다. 더하여 원본 샘플은 z가 아닌 d와 t가 존재하므로 수식 (20)과 (21)에서 z의 정수부와 소수부를 각각 d와 t로 역변환함을 정의한다. 이와 같은 변환 과정은 연속적인 원본 샘플 간의 실질적인 시간차를 계산하여 유사 샘플의 d와 t를 생성함을 의미한다.

$$x_s = \frac{u}{w} \times x_o + \frac{w-u}{w} \times x_{o+1} \tag{22}$$

$$y_s = \frac{u}{w} \times y_o + \frac{w-u}{w} \times y_{o+1} \tag{23}$$

표 7. 데이터 최적 증강의 매개변수와 정의
Table 7. Parameters and definitions of the data optimal augmentation

Definition	Parameter
Number of intervals between original samples to produce similar samples	w
Minimum and maximum values of w	w_a, w_b
Index of similar samples to be produced between specific original samples	u
The value obtained by converting d and t	z
The number of original samples	O
Constraints on the time difference	c_z
Constraints on the difference in direction angle	c_v
The angle between the two directional angles	θ

$$s_s = \frac{u}{w} \times s_o + \frac{w-u}{w} \times s_{o+1} \tag{24}$$

$$h_s = \frac{u}{w} \times h_o + \frac{w-u}{w} \times h_{o+1} \tag{25}$$

수식 (22)부터 (25)는 순서대로 o번째와 $o+1$번째 원본 샘플의 비율로 생성하는 유사 샘플의 x_s, y_s, s_s, h_s를 정의한다.

$$l_s = g(i,j) \tag{26}$$

수식 (26)은 생성하는 유사 샘플의 클래스 레이블이 존재하지 않으므로 그리드 레이블링 알고리즘에 따라 클래스 레이블을 결정함을 정의한다.

$$o = 1,2,\cdots,O-1 \tag{27}$$

$$s = O+1,\cdots,O+(O-1)\times(w-1) \tag{28}$$

수식 (27)은 수식 (12)부터 (26)이 사용하는 원본 샘플의 색인을 정의한다. 더하여 수식 (28)은 원본 샘플 사이에 생성하는 모든 유사 샘플의 색인을 수식 (27)로부터 정의한다.

제안하는 알고리즘은 F1-Score와 Log-Loss를 최적화하기 위해 $[w_a, w_b]$에서 최적의 증강률을 의미하는 매개변수 w를 결정한다. 그러나 그림 8의 (a), (b)와 같이 신뢰할 수 없는 증강은 잘못된 유사 샘플을 생성한다. 이를 위해 연속적인 원본 샘플 사이에 유사 샘플을 생성할지를 결정하는 제약조건을 다음과 같이 정의한다.

그림 8. 신뢰할 수 없는 증강
Fig. 8. Unreliable augmentation

$$z_{o+1} - z_o \le c_z \tag{29}$$

$$0 < c_z \le 1 \tag{30}$$

수식 (29)는 연속적인 원본 샘플의 시간차가 제약조건보다 작거나 같으면 유사 샘플을 생성함을 정의한다. 수식 (30)은 시간차가 큰 경우를 위한 제약조건의 범위를 정의한다. 따라서 연속적인 원본 샘플의 시간차가 하루보다 크면 제약조건에 따라 유사 샘플을 생성하지 못한다.

$$\overrightarrow{v_\alpha} = \overrightarrow{v_{o,o+1}} = [x_{o+1} - x_o, y_{o+1} - y_o] \tag{31}$$

$$\overrightarrow{v_\beta} = \overrightarrow{v_{o+1,o+2}} = [x_{o+2} - x_{o+1}, y_{o+2} - y_{o+1}] \tag{32}$$

수식 (31)과 (32)는 연속적인 원본 샘플과 그다음 연속적인 원본 샘플의 방향각을 정의한다.

$$c_v \le \left(\overrightarrow{v_\alpha} \cdot \overrightarrow{v_\beta}\right) / \left(\left|\overrightarrow{v_\alpha}\right| \times \left|\overrightarrow{v_\beta}\right|\right) \tag{33}$$

수식 (33)은 수식 (31)과 (32)에서 정의한 두 방향각의 $\cos(\theta)$가 제약조건보다 크거나 같으면 유사 샘플을 생성함을 정의한다.

$$-1 \le c_v \le 1 \tag{34}$$

$$o = 1, 2, \cdots, O-2 \tag{35}$$

수식 (34)는 방향각의 차가 큰 경우를 위한 제약조건의 범위를 정의한다. 수식 (35)는 수식 (31)부터 (34)가 사용하는 원본 샘플의 색인을 정의한다.

제안하는 알고리즘은 단말기 이동성 예측 모델의 F1-Score와 Log-Loss를 최적화하기 위해 매개변수 w, c_z, c_v와 최적의 증강률을 결정하고, 제한된 양의 단말기 이동성 데이터 집합에서 수많은 양의 데이터 집합을 확보한다.

V. 실 험

단말기 이동성 예측 모델의 일반화를 위해 데이터 최적 증강 알고리즘을 테스트 데이터 집합이 아닌 훈련 데이터 집합에 대해서만 적용한다. 그리드 레이블링 알고리즘을 적용한 단말기 이동성 데이터 집합에 대해 데이터 최적 증강 알고리즘을 적용하기 위한 매개변수의 값은 표 8에 정리한다.

저자가 아는 한 수치형 데이터 집합에 대해 데이터 증강을 적용한 다른 알고리즘은 존재하지 않는다. 이에 따라 본 논문에서는 F1-Score의 최적화를 목표로 적용한 경우와 Log-Loss의 최적화를 목표로 적용한 경우를 서로에 대한 비교군으로 적용한다.

단말기 수가 1인 훈련 데이터 집합에 대해 표 8에 따른 매개변수의 범위에서 데이터 최적 증강 알고리즘을 적용한 결과는 그림 9와 같고, 매개변수 w의 값이 증가하면서 매개변수 c_v의 값이 감소할수록 샘플의 수가 증가하는 것을 알 수 있다. 이에 따라 단말기 수가 서로 다른 이동성 데이터 집합에 대해 각각 F1-Score의 최적화, Log-Loss의 최적화를 목표로 적용한 결과를 표 9에 정리한다. 단말기 수에 상관없이 매개변수 w, c_v의 값은 F1-Score의 최적화를 목표로 적용한 경우가 평균 7.8, -0.56이고, Log-Loss의 최적화를 목표로 적용한 경우가 평균 9.6, -0.32임을 알 수

표 8. 실험을 위한 매개변수의 값
Table 8. Values of parameters for experiments

Parameter	Value	Description
w_a	2	Minimum values of w
w_b	10	Maximum values of w
c_z	1.1574E-4	Time difference constraints
c_v	-1.0~1.0	Directional angle difference constraints

표 9. 단말기 수와 목표에 따른 최적의 매개변수와 증강률
Table 9. Optimal parameters and augmented rate according to the number of terminals, and objectives

The number of terminals	Objective of optimization	The optimal value of w	The optimal value of c_v	The number of samples in the training dataset		The optimal augmented rate
				Original	Augmented	
1	F1-Score	8	-1.0	4143	30260	7.3039
	Log-Loss	9	-0.4		33375	8.0558
3	F1-Score	6	-0.7	12466	67801	5.4389
	Log-Loss	10	0.0		110647	8.8759
5	F1-Score	7	-0.3	20718	129684	6.2595
	Log-Loss	10	0.0		183609	8.8623
7	F1-Score	9	-0.4	28937	231977	8.0166
	Log-Loss	10	-0.2		256619	8.8682
9	F1-Score	9	-0.4	37117	296965	8.0008
	Log-Loss	9	-1.0		302621	8.1532

있다. 이에 따라 w가 낮고, c_v가 높을수록 F1-Score가 최적화되며 Log-Loss는 이와 정반대임을 알 수 있다.

단말기 이동성 예측 모델에 단말기 수가 서로 다른 증강된 이동성 데이터 집합을 적용한 결과는 그림 10과 같다. F1-Score는 F1-Score의 최적화를 목표로 적용한 경우가 평균 0.9413이고, Log-Loss의 최적화를 목표로 적용한 경우가 평균 0.9358임에 따라 적용하지 않은 경우보다 모두 높은 것을 알 수 있다. 더하여 Log-Loss는 F1-Score의 최적화를 목표로 적용한 경우가 평균 0.2226이고, Log-Loss의 최적화를 목표로 적용한 경우가 평균 0.2123임에 따라 적용하지 않은 경우보다 모두 낮은 것을 알 수 있다. 이에 따라 F1-Score와 Log-Loss 중 하나를 최적화의 목표로 적용하는 경우 다른 하나도 함께 향상되는 것을 알 수 있다.

실험에 따라 제안하는 데이터 최적 증강 알고리즘의 매개변수 w, c_z, c_v와 최적의 증강률을 결정하면 F1-Score와 Log-Loss 중 하나를 최적화의 목표로 하더라도 F1-Score에서 적용하지 않은 경우보다 모두 높고, Log-Loss에서 적용하지 않은 경우보다 모두 낮음을 증명한다.

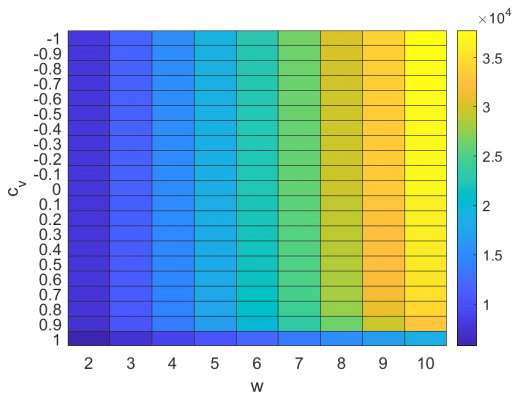


그림 9. 매개변수에 따른 훈련 데이터 집합의 샘플 수
Fig. 9. The number of samples of the training dataset according to the parameters

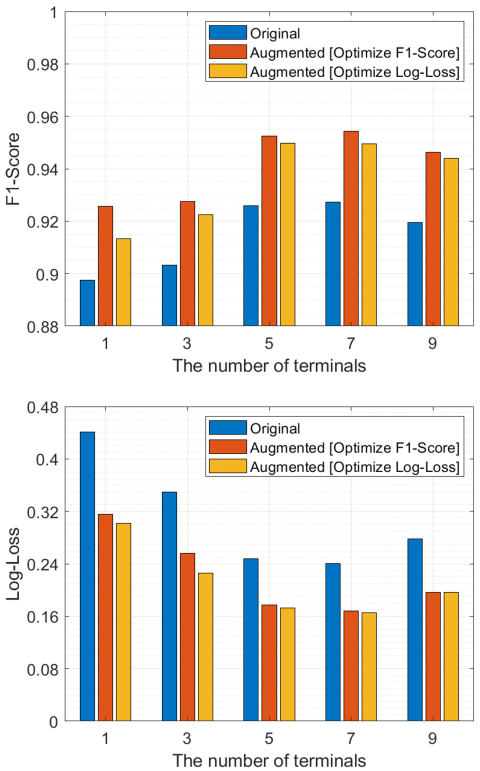


그림 10. 단말기 수에 따른 F1-Score와 Log-Loss
Fig. 10. F1-Score and Log-Loss according to the number of terminals

Ⅵ. 결 론

CubeSat 기반의 IoST에서 유동적인 이동성을 가진 단말기가 지상망을 거치지 않고 CubeSat과 직접 통신할 경우, Terrestrial-CubeSat 간의 핸드오버와 함께 링크를 재설정하기 위한 오버헤드가 발생한다. 이러한 오버헤드를 처리하는 과정에서 IoST의 적은 네트워크 자원으로 인해 예약 지연이 발생하여 핸드오버가 실패할 경우, IoST는 링크 단절과 함께 통신의 안정성을 잃게 된다. 이와 같은 불안정한 IoST의 안전성을 높이기 위해서는 단말기의 이동성을 예측함으로써 핸드오버를 효율적이고 안정적으로 지원해야 한다.

본 논문에서는 단말기 이동성 예측 문제를 다중 분류 문제로 정의하는 그리드 레이블링 알고리즘과 기계학습 기반 모델의 성능을 최적화하는 데이터 최적 증강 알고리즘을 제안한다. 이에 따라 실험은 제안하는 알고리즘이 IoST에서 효율적이고 안정적인 핸드오버를 지원할 수 있음을 증명한다.

References

[1] J. Höller, V. Tsiatsis, C. Mulligan, S. Karnouskos, S. Avesand, and D. Boyle, *From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence*, Academic Press, 2014.

[2] Z. Qu, G. Zhang, H. Cao, and J. Xie, "LEO satellite constellation for Internet of Things," *IEEE Access*, vol. 5, pp. 18391-18401, 2017.

[3] L. A. Davis and L. Filip, *How Long Does It Take to Develop and Launch Government Satellite Systems?*, Aerospace, 2015.

[4] I. F. Akyildiz and A. Kak, "The internet of space Things/CubeSats: A ubiquitous cyber-physical system for the connected world," *Comput. Netw.*, vol. 150, pp. 134- 149, Feb. 2019.

[5] I. F. Akyildiz and A. Kak, "The internet of space Things/CubeSats," *IEEE Netw.*, vol. 33, no. 5, pp. 212-218, Sep. 2019.

[6] N. Saeed, A. Elzanaty, H. Almorad, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "CubeSat communications: Recent advances and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1839-1862, 3rd Quart. 2020.

[7] H. Park, Y. Kwon, K. Lee, Y. Choi, Y. Cho, and B. Cho, "A Seamless Handover Scheme Based on Path-Prediction for Network Mobility," *J. KICS*, vol. 30, no. 7A, pp. 550-556, Jul. 2005.

[8] Z. Ali, N. Baldo, J. Mangues-Bafalluy, and L. Giupponi, "Machine learning based handover management for improved QoE in LTE," in *Proc. IEEE/IFIP Netw. Oper. Manag. Symp. (NOMS)*, pp. 794-798, Istanbul, Turkey, Apr. 2016.

[9] H. Xu, D. Li, M. Liu, G. Han, W. Huang, and C. Xu, "QoE-driven intelligent handover for user-centric mobile satellite networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10127-10139, Sep. 2020.

[10] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 113-123, Jun. 2019.

오 준 석 (Junsuk Oh)

2021년 2월 : 중앙대학교 컴퓨터공학부 학사 졸업

2021년 3월~현재 : 중앙대학교 컴퓨터공학과 석사과정

<관심분야> IoT, 5G, 무선통신, 빅데이터, 연합학습

[ORCID:0000-0001-7855-6461]

이 동 현 (Donghyun Lee)

2020년 2월 : 동국대학교 컴퓨터공학과 학사 졸업

2020년 3월~현재 : 중앙대학교 컴퓨터공학과 석사과정

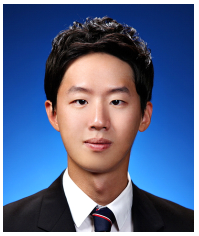
<관심분야> IoT, 5G, 무선 센서 네트워크

[ORCID:0000-0001-9117-5647]

하 태 윤 (Taeyun Ha)

2020년 2월 : 중앙대학교 컴퓨터공학부 학사 졸업
2020년 3월~현재 : 중앙대학교 컴퓨터공학과 석사과정
<관심분야> TCP, IoT, 무선통신, 강화학습
[ORCID:0000-0001-6001-4226]

이 윤 성 (Yunseong Lee)



2013년 2월 : 중앙대학교 컴퓨터공학부 학사 졸업
2015년 8월 : 중앙대학교 컴퓨터공학과 석사
2017년 3월~현재 : 중앙대학교 컴퓨터공학과 박사과정
<관심분야> Directional routing, Handover control, Heterogeneous networks
[ORCID:0000-0001-9245-2968]

조 성 래 (Sungrae Cho)

1992년 2월 : 고려대학교 전자전산공학과 학사
1994년 2월 : 고려대학교 전자공학과 석사
2002년 12월 : 미국 조지아공대 전기및컴퓨터공학과 박사
1994년 2월~1996년 8월 : 한국전자통신연구원 연구원
2003년 8월~2006년 7월 : 미국 조지아서던대학교 컴퓨터공학과 조교수
2006년 9월~현재 : 중앙대학교 컴퓨터공학부 교수
<관심분야> 무선네트워크, Ubiquitous Computing
[ORCID:0000-0003-1879-688X]