Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 2, pp. 148-162, April 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.2.148

전기 이륜차 배터리 교환 수요 예측 개선을 위한 삼중 선택 전략 기반 의사 라벨링 기법

유선호[1] · 주승돈[2] · 고광종[1†] · 정태수[1†]

[1]고려대학교 산업경영공학과 / [2]젠트로피

TriMa: Tri-strategy Margin-filtered Pseudo-labeling for Battery Swapping Demand Forecasting

Seon-Ho Yoo[1] · Seungdon Zu[2] · Gwang-Jong Ko[1] · Taesu Cheong[1]

[1]Department of Industrial and Management Engineering, Korea University
[2]Zentropy Co., Ltd.

Accurate prediction of battery-swapping demand is crucial for minimizing operational costs and maintaining service quality in electric motorcycle Battery-as-a-Service (BaaS) deployment. However, early-stage BaaS providers often face severe data scarcity and bias, especially in newly expanded regions. In these areas, newly deployed battery swapping stations (BSS) encounter cold-start problems due to limited historical labels, and existing regional models degrade under domain shift caused by distributional changes in contextual covariates. Although conventional data augmentation techniques like SMOTE are often employed to mitigate these challenges, they perform poorly in multi-class settings by exacerbating inter-class overlap near complex decision boundaries. To address this limitation, we introduce TriMa (Tri-strategy Margin-filtered Pseudo-labeling), an advanced pseudo-labeling framework designed to enhance predictive performance under sparse and imbalanced multi-class data distributions. TriMa constructs augmented samples exclusively from high-confidence pseudo-labels, which are derived through three complementary labeling strategies combined with margin-based filtering. Experiments on UCI datasets under simulated sparsity and bias, as well as on real-world operational data from a BaaS provider in Seoul, demonstrate the efficacy of our proposed method. Results show that TriMa consistently improves the F1 score and reduces performance variability relative to existing techniques, while remaining robust to label noise.

Keywords: Battery Swapping Electric Motorcycle, Pseudo-labeling, Data Scarcity, Data Imbalance, Data Augmentation

1. 서 론

전기 이륜차는 대도시의 대기오염과 교통혼잡 문제를 동시에 완화할 수 있는 친환경 교통수단으로 주목받고 있다. 특히, BaaS(Battery-as-a-Service)로서 배터리 교환 방식의 인프라인 배터리 교환소(battery swapping stations, BSS)와 이를 이용하

이 논문은 2025년도 산업통상자원부 및 산업기술기획평가원(KETI) 연구비 지원에 의한 연구(20023616) 이며, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-00521940). 또한, 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의 연구결과임.
† 연락저자 : 고광종 박사수료, 서울특별시 성북구 안암로 145 고려대학교 서울캠퍼스(자연계) 신공학관 209호, Tel : 02-3290-3382, Fax : 02-3290-4550, E-mail : koptimizer@korea.ac.kr
† 연락저자 : 정태수 교수, 서울특별시 성북구 안암로 145 고려대학교 서울캠퍼스(자연계) 공학관 511호, Tel : 02-3290-3484, Fax : 02-3290-4550, E-mail : tcheong@korea.ac.kr
2025년 11월 21일 접수; 2025년 12월 31일 수정본 접수; 2026년 2월 13일 게재 확정.

는 배터리 교환형 전기 이륜차는 이용자가 배터리 교환을 통해 충전 시간을 기다리지 않고 즉시 주행을 재개할 수 있기 때문에 친환경 도심 물류에서 중요한 역할로 떠오르고 있다(Feng and Lu, 2022).

그러나 BaaS 제공자는 대부분 초기 시장 진입자이므로 신규 지역에 BSS를 설치 및 확장하는 과정에서 기술적 한계가 구조적으로 발생한다. 첫째, 신규 BSS 설치 초기에는 이용 이력이 거의 존재하지 않아 수요 라벨 데이터가 매우 제한되는 콜드 스타트(cold start) 문제가 발생한다. 이는 지도학습 기반 수요 예측 모델이 학습에 필요한 정답 라벨을 충분히 확보하지 못해 성능이 급격히 저하되는 원인으로 작용한다. 둘째, 기존에 운영 중인 지역에서 학습된 모델을 신규 지역에 그대로 적용하는 경우, 인구 구성, 토지 이용, 상업 시설 분포, 교통 접근성 등 지역 맥락이 달라 입력 특성의 분포가 변하는 도메인 이동(domain shift)이 나타난다. 이때 학습 데이터 분포와 적용 환경 분포 간의 불일치로 인해 동일 모델이라도 일반화 성능이 떨어지며, 특히 신규 지역은 콜드 스타트와 도메인 이동이 동시에 존재하므로 예측 오류가 더 크게 증폭될 수 있다.

한편, BSS 수요 예측은 라벨 데이터는 부족하지만, 인구 밀도, 상업 시설 분포, 토지 이용, 대중교통 접근성 등과 같은 라벨이 없는 주변 정황 데이터는 비교적 풍부하게 확보 가능한 전형적인 반지도 학습(semi-supervised learning) 환경에 해당한다. 따라서 제한된 라벨만으로 모델을 구축하는 접근은 가용 정보의 상당 부분을 활용하지 못한다는 한계를 갖는다. 이러한 맥락에서 의사 라벨링(pseudo-labeling)은 라벨이 없는 샘플에 대해 모델이 생성한 예측값을 임시 정답 라벨로 활용하여 추가적인 정답 라벨을 구성할 수 있어, 콜드 스타트 상황에서 가용한 라벨 정보를 보강하고 도메인 이동 환경에서 표현 공간을 정제하는 데 중요한 역할을 한다. 다만 의사 라벨은 오류가 포함될 수 있으므로, 신뢰도 관리 없이 단순히 활용하면 오류 전파로 성능이 악화될 위험이 존재한다.

또한 실 운영 환경에서 수요 분포는 본질적으로 클래스 불균형을 동반한다. 도심 상업지, 환승 거점 등 일부 지역의 스테이션에 수요가 집중되는 반면, 외곽, 주거 중심, 신규 설치 구간의 다수 스테이션은 낮은 이용 빈도를 보이는 경향이 강하다. 이로 인해 수요 수준을 이산 클래스로 구분할 경우 '저수요' 다수 클래스가 과대표집되고 '고수요' 소수 클래스는 극소수로 남아, 모델이 다수 클래스에 편향된 의사결정을 내리기 쉬워진다. 대표적인 데이터 증강 기법인 SMOTE는 소수 클래스의 학습 안정성을 높이고 클래스 간 불균형을 완화하는 데 효과적이지만, 데이터 분포의 경계 근처에서 생성된 합성 샘플이 다른 클래스의 영역을 침범할 위험이 존재한다. 이러한 문제는 다중 클래스 분류 문제에서 더욱 두드러지며, 클래스 수가 증가함에 따라 경계 또한 기하급수적으로 복잡해져 클래스 간 오버랩(overlap) 가능성이 급격히 증가하는 한계가 있다(Yang *et al.*, 2024).

본 연구에서는 세 가지의 의사 라벨 생성 전략을 병렬적으로 운영하여 신뢰도가 높은 의사 라벨만을 선별해 데이터 부족 및 편중 문제를 해결하는 데이터 증강 기법인 TriMa(Tri-strategy margin-filtered pseudo-labeling)와 이를 이용한 예측 프레임워크를 제안한다. 구체적으로는, 예측 확률 기반, 상위 퍼센타일 기반, 그리고 예측 불확실성 지표인 엔트로피 기반의 의사 라벨 생성 전략을 병렬적으로 사용하며, 신뢰도가 높은 의사 라벨을 선별하기 위해 margin 필터를 이용한다. 이를 통해 반지도 학습 환경에서 발생할 수 있는 의사 라벨 오류 전파를 억제하면서, 콜드 스타트 및 도메인 이동 조건에서 학습에 활용 가능한 라벨 정보를 안정적으로 보강할 수 있다. 이렇게 생성된 의사 라벨 데이터는 최종적으로 SMOTE를 통해 클래스 균형을 유지하도록 보완되어, 불균형 환경에서도 효율적인 학습을 가능하게 한다. 데이터 부족 및 편중이 반영된 UCI 공개 데이터셋과 서울에서 서비스를 제공하는 BaaS 사업자의 실제 데이터를 대상으로 제안 프레임워크의 성능 평가 결과, 기존 SMOTE 기반 증강 기법보다 예측 정확도와 모델 강건성 측면에서 더 우수한 성능을 보였다.

본 연구의 주요 기여점은 다음과 같다. 세 가지 상이한 의사 라벨링 기준과 margin 필터를 활용하여 새로운 데이터 증강 기반 예측 모델을 제안한다. 제안한 모델은 기존 데이터 증강 기법인 SMOTE와 비교하였을 때, 일정 불균형도 이상의 데이터셋에 대하여 유의미한 성능 향상을 확인하였으며, 데이터 증강에 있어 중요한 지표인 모델의 강건성 역시 기존 SMOTE 증강 방식보다 우수한 성능을 입증하였다. 추가적으로, 라벨 오류가 포함된 데이터 노이즈 환경에서도 앙상블 예측과 margin 기반 선택을 통해 오류 전파를 완화하며 성능 저하를 효과적으로 억제함을 정량적으로 검증하였다. 특히, 대한민국 서울에서 서비스되고 있는 BaaS 제공업체의 실 운영 데이터를 대상으로도 실험을 진행하였기에 실제 데이터에 기반한 수요 예측 분야에서 높은 효용성을 보일 수 있을 것으로 기대한다.

본 논문은 다음과 같은 구조를 가진다. 제2장에서는 데이터 부족 및 편중 환경에서의 수요 예측 연구와 의사 라벨링 기법의 선행 문헌들을 살펴보며 기존 방법의 한계와 보완 전략을 분석한다. 제3장에서는 제안 방법인 TriMa의 구체적인 메커니즘을 모식도와 함께 설명한다. 제4장에서는 UCI 공개 분류 데이터셋 및 실제 BaaS 제공업체의 운영 데이터를 기반으로 기존 증강기법과 TriMa의 성능을 비교 검증한다. 마지막으로 제5장에서 결론 및 향후 연구방향을 제시한다.

2. 선행 연구

2.1 수요 예측에서의 데이터 부족 및 편중 문제

신산업의 초기 서비스처럼 충분한 학습 데이터가 축적되지 않은 환경에서는 전통적 머신러닝 및 딥러닝 모델의 예측 정확도와 신뢰성을 확보하기 어렵다. 이러한 제약은 항공 운송 수요 예측 연구에서 대량의 과거 데이터를 요구하는 시계열

딥러닝 모델의 적용 한계로 보고되었으며(Wang *et al.*, 2020), 소매업에서도 소규모 매장의 판매 이력 부족이 재고 수요 예측 성능을 저하시킨다는 결과가 제시되었다(Lei *et al.*, 2025). 물류 배송 분야에서는 불균형한 데이터 분포가 모델 편향을 강화시켜 예측 품질을 악화시키는 문제가 관찰되었다(Leeuw *et al.*, 2023). 한편, 국방 분야에서는 제한된 과거 데이터와 불규칙한 수요 패턴 하에서도 머신러닝 기반 모델을 적용하여 기존 방법 대비 예측 정확도를 개선한 연구가 제안되었고(Son *et al.*, 2022), 자동차 예비 부품 장기 수요 예측에서는 계절성과 간헐적 수요로 인한 데이터 부족을 완화하기 위해 계절조정과 전이학습을 결합한 접근이 활용되었다(Lee *et al.*, 2021). 이러한 선행연구는 데이터가 충분하지 않은 산업 환경에서 단순한 데이터 중심 접근만으로는 한계가 존재하며, 데이터 제약을 직접적으로 보완하는 방법론이 필요함을 시사한다.

BSS 수요 예측 관점에서도 초기 인프라 구축 단계의 데이터 부족은 핵심 난제로 나타난다. 배터리 교환 수요를 스테이션 단위의 단기 이벤트 수로 정의하고 실제 운영 로그에 대해 딥러닝 기반 예측 모델을 적용한 연구는 제한된 데이터에 의존하는 현실을 보여주며 초기 구축 단계에서의 일반화가 여전히 과제로 남음을 강조한다(Wang *et al.*, 2023). 또한 BSS 간 공간적 결합과 시간 패턴을 함께 학습하기 위해 시공간 모델을 적용한 연구도 제안되었으나, 이러한 모델 역시 안정적인 패턴 학습을 위해 일정 수준 이상의 관측 데이터가 요구되어 초기 구축 단계에서는 성능 저하가 발생할 수 있다(Hu *et al.*, 2024). 더 나아가 콜드 스타트 환경에서는 다른 영역에서 관측된 수요 정보를 이전하는 접근도 제안되었는데, 대표적으로 자전거 공유와 대중교통 수요 간 관계를 활용한 전이학습은 단일 모달 수요 예측 대비 성능 개선을 보고함으로써, BSS에서도 외부 정황 데이터를 활용해 데이터 제약을 완화할 가능성을 제시한다(Hua *et al.*, 2025). 종합하면 BSS 수요 예측은 시공간적 불확실성과 초기 데이터 제약이 결합된 문제로서, 제한된 라벨을 보완할 수 있는 학습 전략이 요구된다.

2.2 의사 라벨링 기법의 연구 동향

의사 라벨링은 라벨이 부족한 상황에서 비라벨 데이터를 활용하기 위한 대표적인 반지도 학습 전략으로, 모델이 비라벨 데이터에 대해 예측한 결과를 임시 라벨로 간주해 학습 데이터를 확장하는 방식이다(Lee, 2013). 초기 연구들은 이러한 self-training 계열 접근이 라벨 비용을 줄이면서도 성능을 개선할 수 있음을 보여주었고, 이후 이미지 분류와 텍스트 분류 등 다양한 분류 문제에서 의사 라벨링이 실용적 방법론으로 자리잡았다(Yang *et al.*, 2023). 연구가 확장되면서 의사 라벨링의 핵심 난점도 함께 정리되었다. 특히 초기 모델이 편향되거나 불확실성이 큰 상황에서는 오분류된 의사 라벨이 반복적으로 주입됨에 따라 학습이 왜곡되는 오류 전파가 발생할 수 있음이 지적되었다(Arazo *et al.*, 2020). 이를 완화하기 위해 일정 신뢰도 이상의 예측만 선택하는 임곗값 기반 선별적 의사 라벨링이 제안되었고(Li *et al.*, 2024), 의사 라벨 품질을 높이기 위해 선택 기준을 강화하는 방향이 일반적으로 채택되었다. 다만 실제 수요 예측과 같이 외부 요인 변동, 관측 노이즈, 지역별 이질성 등 불확실성이 복합적으로 존재하는 환경에서는 단일 임곗값으로 의사 라벨 품질을 안정적으로 보장하기 어렵다는 한계가 존재한다.

최근 연구들은 이러한 한계를 보완하기 위해 임곗값 기반 선택의 경직성을 완화하거나, 의사 라벨의 신뢰도 추정을 정교화하는 방향으로 발전하고 있다. FreeMatch는 학습 상태에 따라 신뢰도 임곗값을 자동으로 조정하고 클래스 불균형을 완화하는 정규화를 도입함으로써, 극소수 라벨 및 불균형 환경에서의 안정적인 의사 라벨 활용을 목표로 한다(Wang *et al.*, 2022). 이후 SoftMatch는 의사 라벨의 수량과 품질 사이의 트레이드오프에 주목하여, 신뢰도를 연속적 가중치로 반영하여 더 많은 비라벨 샘플을 활용하면서도 노이즈를 억제하는 방식을 제안하였다(Chen *et al.*, 2023). 가장 최근에는 TrustMatch가 의사 라벨 편향이 학습 성능 저하로 이어지는 문제에 주목하여, 신뢰 기반 정제를 통해 오류 전파를 완화하는 접근을 제시함으로써 의사 라벨의 편향과 품질을 동시에 관리하는 관점의 중요성을 보여주었다(He & Hong, 2025). 종합하면 의사 라벨링 연구는 임곗값 선택의 자동화, 신뢰도 기반 가중 학습, 편향 완화 및 정제 메커니즘의 도입 순으로 고도화되어 왔으며, 이는 불확실성이 큰 수요 예측 문제에서도 의사 라벨 품질 관리가 핵심이라는 점을 보여준다.

2.3 불균형 데이터에서의 데이터 증강 기법 한계

데이터 불균형 문제를 완화하기 위한 대표적 기법으로 SMOTE(Synthetic Minority Over-sampling Technique)가 널리 사용된다. SMOTE는 소수 클래스 샘플의 k-최근접 이웃(k-NN)을 기반으로 샘플 간 선형 보간을 수행하여 합성 샘플을 생성함으로써, 클래스 간 샘플 수 차이를 줄이는 방식이다(Chawla *et al.*, 2002). 단순 복제 기반 오버샘플링보다 과적합 위험을 낮추고 소수 클래스의 지역적 분포를 보강할 수 있다는 점에서 다양한 도메인으로 확장되어 왔다. 이러한 맥락에서 수요 예측 및 시계열 분류 문제에서도 SMOTE 계열 증강을 적용해 소수 클래스의 학습 신호를 강화하려는 연구가 지속적으로 보고되었다. 예를 들어 시계열 예측 연구에서는 오버샘플링을 통해 희소 구간 또는 소수 패턴의 데이터 수를 보강하여 예측 성능을 개선하는 접근이 제시되었으며(Cerqueira *et al.*, 2024), 희귀 이벤트 예측을 포함한 시계열 예측 문제에서 다양한 리샘플링 전략을 체계적으로 비교한 연구는 오버샘플링이 적절한 설정과 결합될 경우 희소 구간 예측 성능을 유의하게 향상시킬 수 있음을 보였다(Moniz *et al.*, 2017). 더 나아가 간헐적 예비 부품 수요와 같이 극심한 불균형이 존재하는 환경에서는 SMOTE와 손실 함수 설계를 결합한 앙상블 접근이

예측 성능과 민감도를 동시에 개선할 수 있음을 제시되었다(Kenaka *et al.*, 2025). 이러한 연구들은 수요 예측 문제에서 불균형이 빈번히 나타나며, 데이터 증강이 실질적인 성능 개선에 기여할 수 있음을 보여준다.

그러나 SMOTE는 합성 샘플이 데이터 분포의 경계 근처에서 생성될 경우 인접 클래스 영역으로 침투하여 클래스 간 중첩을 증가시킬 수 있다는 한계를 지닌다. 특히 다중 클래스 분류에서는 클래스 수가 증가할수록 결정 경계가 복잡해지고, 그 결과 경계 주변에서의 중첩 가능성이 크게 높아진다(Yang *et al.*, 2024). 이러한 문제를 완화하기 위한 변형으로 경계 근처의 소수 표본을 선택적으로 증강하는 Borderline-SMOTE가 제안되었으나(Han *et al.*, 2005), 다중 클래스 환경에서는 경계가 다면적으로 복잡해져 경계 기반 증강 자체가 중첩을 확대할 가능성 또한 존재한다. 더 나아가 불균형 환경에서의 성능 저하는 단순한 클래스 비율뿐 아니라 클래스 간 중첩의 정도에 의해 크게 좌우될 수 있다는 논의도 제기되어 왔다. 불균형과 중첩이 동시에 존재할 때 학습기가 보이는 거동을 분석한 연구는 중첩이 클수록 단순 리샘플링만으로는 성능 개선이 제한될 수 있음을 보여주었고(Prati *et al.*, 2004), 불균형 분류에서의 클래스 오버랩 문제를 정리한 연구는 중첩 자체가 소수 클래스의 결정 영역을 붕괴시키거나 오분류를 증가시키는 핵심 요인이 될 수 있음을 강조하였다(Vuttipittayamongkol and Elyan, 2021). 따라서 불균형 데이터 증강을 적용할 때에는 단순히 샘플 수를 맞추는 접근을 넘어, 합성 샘플이 결정 경계 및 클래스 중첩 구조에 미치는 영향을 함께 고려할 필요가 있다.

3. 제안 방법론

3.1 제안 프레임워크 개요

앞서 선행연구 분석에서 논의한 바와 같이, 초기 단계의 BSS 수요 예측 정확도를 개선하기 위한 데이터 증강 기법에서는 고신뢰 의사 라벨을 선택적으로 선별함으로써 결정경계 근처의 불확실 샘플의 유입을 억제하는 전략이 필요함을 확인하였다. 따라서 본 연구에서는 기존 self-training 기법을 확장한 데이터 증강 기법인 TriMa와 이를 이용한 반지도 학습 프레임워크를 제안한다. Self-training은 라벨 데이터는 부족하나 비라벨 데이터가 풍부한 상황에서, 모델이 예측에 대해 일정 수준 이상의 신뢰도를 보이는 비라벨 샘플에 의사 라벨을 부여하여 학습 데이터를 점진적으로 확장하는 반지도 학습 기법이다. 그러나 단일 임곗값만을 기준으로 삼을 경우, 특정 환경에서 부적절한 샘플이 선택될 위험이 있으며, 대상 클래스 간의 불균형이 심화되어 예측 성능이 오히려 저하될 수 있다(Guo and Li, 2022). 이에 본 연구는 상호 보완적인 세 가지 self-training 전략을 병렬로 결합하고, margin 기반 필터링을 통해 신뢰도가 높은 샘플만을 선별하여 의사 라벨을 부여하는 TriMa 모델을 설계하였다. 제안하는 전체 프레임워크의 구조는 Figure 1에 제시하였다.

3.2 삼중 전략 기반 의사 라벨 생성 (TriMa)

TriMa는 라벨 데이터 D_L과 비라벨 데이터 D_U를 입력으로 받으며, *Static, Dynamic, Entropy*의 세 가지 self-training 모델이 비라벨 샘플 $x_i \in D_U$에 대한 클래스별 예측 확률 벡터를 병렬로 산출하는 구조를 지닌다. 세 전략은 동일한 입력과 형태의 확률 출력을 사용하되, 의사 라벨 후보를 채택하는 기준을 서로 다르게 설정함으로써 단일 기준 선택의 편향과 불안정성을 상호 보완한다.

첫 번째로 *Static Model*은 초기 학습 단계의 기본 모델을 정적 기준으로 활용한다. 즉, 사전에 학습된 기준 모델을 통해 비라벨 샘플의 예측 확률이 사전에 정의된 고정 임곗값 τ_s 이상인 경우에만 의사 라벨을 부여한다. 이를 통해, 학습 초기 단계에서 부정확한 예측을 의사 라벨로 포함하지 않도록 제어한다. 두 번째로 *Dynamic Model*은 학습 과정이 진행됨에 따라 모

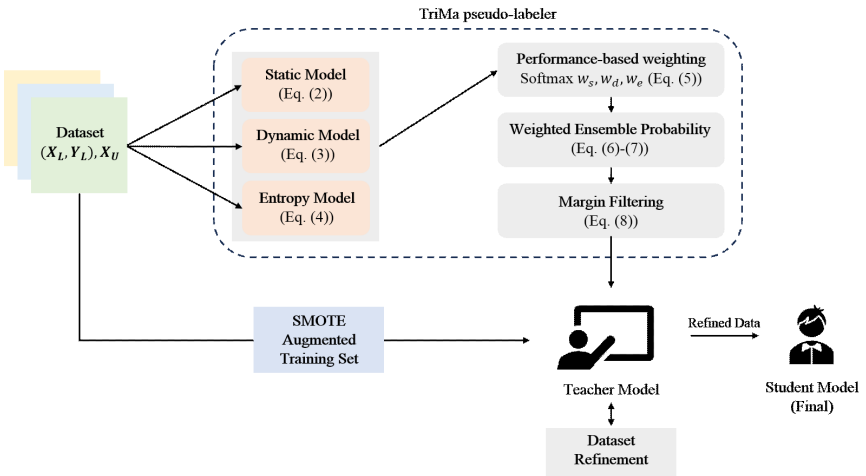


Figure 1. Overall framework of the proposed TriMa-based data augmentation and teacher-student learning scheme

델이 업데이트되는 동적 기준을 활용한다. 즉, 매 반복마다 현재 모델의 예측 신뢰도가 상위 N% 이내에 포함되는 샘플만 의사 라벨 대상에 포함함으로써, 학습이 진행될수록 예측의 신뢰도를 더 엄격하게 관리한다. 이로 인해, 모델이 개선됨에 따라 더욱 정밀한 비라벨 샘플이 선택되도록 유도한다. 마지막으로 *Entropy Model*은 예측 결과의 불확실성을 기준으로 삼는 엔트로피 기반 모델을 사용한다. 예측 확률 분포의 엔트로피가 낮을수록 모델의 신뢰도가 높다고 판단하고, 엔트로피가 특정 임곗값 미만인 샘플을 선별하여 의사 라벨을 부여한다. 이를 통해, 예측 분포가 편향되지 않고 신뢰도가 높은 샘플만을 학습 데이터로 선택한다. 다음 세 가지 방법론에 대한 수식은 다음과 같다.

$$\textit{Static Model}: \ \hat{y_i} = \arg\max_j p_{ij}, \ \text{if} \ \max_j p_{ij} \geq \tau_s \tag{2}$$

$$\textit{Dynamic Model}: \hat{y_i} = \arg\max_j p_{ij}, \text{if} \ \max_j p_{ij} \geq P_\alpha(p_{\max}) \tag{3}$$

$$\begin{aligned}\textit{Entropy Model}: \ \hat{y}_i &= \arg\max_j p_{ij}, \ H(p_i) \\ &= -\sum_j p_{ij}\log(p_{ij}), \ \textit{select if} \ H(p_i) \leq \tau_e\end{aligned} \tag{4}$$

앞서 식 (2)-(4)에서 $P_i = (p_{i1}, \cdots, p_{iK})$는 i번째 샘플에 대해 기본 분류기가 출력한 K차원 클래스별 예측 확률 벡터를 의미하며, p_{ij}는 그 중 j번째 클래스에 대한 예측 확률이다. 또한 τ_s와 τ_e는 각각 *Static* 모델과 *Entropy* 모델에서 사용되는 신뢰도 임곗값이다. $p_{\max}$는 파라미터가 아닌 검증 데이터에서 각 샘플별 최대 예측 확률들을 모은 집합을 의미하며, $P_\alpha(p_{\max})$는 이 분포에 대해 상위 α 백분위수를 반환하는 함수로 정의한다. 즉 *Dynamic* 모델에서는 최대 예측 확률이 $P_\alpha(p_{\max})$ 이상인 경우에만 해당 샘플을 신뢰 가능한 의사 라벨 후보로 채택한다.

식 (2)-(4)에서 $\hat{y}_i$는 모든 비라벨 샘플에 대해 항상 정의되는 값이 아니라, 선택 기준을 통과한 샘플에 대해서만 부여되는 의사 라벨이다. 구체적으로 *Static* 모델에서 식(2)의 조건을 만족하지 않거나, *Dynamic* 모델에서 식(3) 조건을 만족하지 않는 경우, 해당 샘플은 신뢰 가능한 의사 라벨 후보로 채택되지 않으므로 $\hat{y}_i$를 부여하지 않는다. 즉, 위 조건을 만족하지 못하는 샘플에 대해서는 $\hat{y}_i$를 임의의 클래스로 선택하거나 강제 할당하지 않고, 비라벨 데이터셋에 그대로 유지하도록 한다. 동일하게 *Entropy* 모델에서도 식 (4) 조건을 만족하지 않으면 해당 샘플은 후보에서 제외되며 $\hat{y}_i$를 부여하지 않는다. 반면, 각 기준을 통과하여 후보로 채택된 샘플에 대해서는 일관되게 $\hat{y}_i = \arg\max_j p_{ij}$ 규칙에 따라 의사 라벨을 확정한다.

3.3 성능 가중 결합 및 마진 필터링 기반 학습 프레임워크

각 모델의 채택에 대한 가중치를 결정하기 위해 검증 데이터에서 얻은 F1 점수를 각각 도출하고, *Static, Dynamic, Entropy* 세 모델의 상대적 성능을 반영하는 softmax 가중치를 다음 수식 (5)와 같이 정의한다. 즉, 각 모델의 F1 점수를 지수화하여 모델 간 성능 차이를 지수 함수의 곡선 형태로 설정하고, 세 모델의 지수화된 F1 점수의 합으로 나누어 정규화함으로써 다음의 가중치 값을 얻는다. 이 과정을 통해 검증 F1 점수가 높은 모델일수록 분자 기여도가 커지고, 결과적으로 높은 가중치 값을 갖게 된다. 반대로 검증 성능이 상대적으로 낮은 모델은 지수화된 값이 작아 분자 기여도가 낮아지므로, 가중치가 작아진다. 따라서 세 모델 중 가장 높은 성능을 보인 모델은 가장 큰 가중치를 얻게 되고, 가장 낮은 성능을 보인 모델은 가장 작은 가중치를 얻게 되어, 모델 간 결합 시 성능이 좋은 모델의 예측이 상대적으로 더 큰 비중을 차지하도록 한다.

$$w_m = \frac{e^{F1_m}}{e^{F1_s} + e^{F1_d} + e^{F1_e}}, \ m \in s, d, e \tag{5}$$

각 모델이 비라벨 데이터에 대해 예측한 클래스별 확률 벡터를 수식 (6)과 같이 정의한다. 여기서 C는 전체 클래스의 개수이며, $p_{i,c}^{(m)}$는 모델 m이 샘플 i를 클래스 c로 예측한 확률을 의미한다. 각 벡터 $p_i^{(m)}$는 모델 m이 해당 샘플 i에 대해 산출한 확률 분포이며, 이 값들은 softmax를 통해 합이 1이 되도록 정규화된 상태이다. 단일 모델이 예측한 확률 분포만 사용할 경우, 특정 모델의 편향이 결과에 지나치게 반영될 수 있으므로, 본 연구에서는 세 모델의 예측 분포를 성능 기반 가중치로 결합하여 안정적인 확률 분포를 얻도록 하였다. 따라서 가중치를 이용하여 세 모델의 예측 확률 벡터를 선형 결합하면, 비라벨 샘플 i의 결합 확률 벡터 $\hat{p}_i$는 다음 수식 (7)과 같다.

$$p_i^{(m)} = [p_{i,1}^{(m)}, p_{i,2}^{(m)}, \cdots, p_{i,C}^{(m)}] \tag{6}$$

$$\hat{p}_{i,c} = w_s p_{i,c}^{(s)} + w_d p_{i,c}^{(d)} + w_e p_{i,c}^{(e)} \tag{7}$$

여기서 $\hat{p}_{i,c}$는 샘플 i가 클래스 c에 속할 최종 결합 확률이며, 세 모델의 예측 확률에 각각의 가중치를 곱해 더한 값이다. 이 단계는 모델별 예측 분포를 산술 평균하거나 다수결 투표 방식으로 합산하는 것보다 성능이 우수한 모델이 결합 확률에 상대적으로 더 큰 영향력을 행사하도록 한다. 또한, 모델이 극도로 불균형한 확률 분포를 예측하더라도 다른 모델의 예측 분포가 이를 보완해줌으로써, 단일 모델이 가진 과적합과 편향 문제가 완화된다.

다음으로 결합 확률에서 최댓값과 두 번째로 큰 값 사이의 차이를 구해 해당 샘플의 신뢰도를 평가한다. 이때, $\hat{p}_{i,(1)} = \max_c \hat{p}_{i,c}$이며, $\hat{p}_{i,(2)}$는 그 다음으로 큰 확률값이다. 두 값 간 차이를 측정하는 이유는 결합된 확률 분포 상에서 모델이 얼마나 명확하게 하나의 클래스를 지지하는지를 정량화하기 위해서이다. 만약, 두 값 사이의 차이가 크다면, 모델이 클

래스에 대해 높은 신뢰를 갖는다고 간주할 수 있다. 반대로 두 값 사이의 차이가 작다면, 가장 높은 확률과 두 번째 확률 간의 차이가 미미하여, 결합 과정 이후에도 모델이 충분한 확신을 가지지 못한 상황으로 해석할 수 있다.

따라서 미리 설정한 margin 기반 필터링 임곗값 γ를 기준으로 최종 의사 라벨을 확정한다. γ는 실험적으로 조정한 하이퍼파라미터로, 기존 분류 및 반지도 학습 연구들에서도 확률 margin 또는 Confidence threshold를 별도의 하이퍼파라미터로 두고 0.05 ~ 0.2 수준의 작은 상수 범위에서 설정하는 것이 일반적이며, multi-label 분류에서 제안된 Asymmetric Loss 계열 연구(Zhang *et al.*, 2021) 역시 확률 margin을 0.05, 0.1, 0.2 값들에 대해 검증 셋에서 튜닝하여 사용한다. 본 연구에서 γ는 의사 라벨의 품질과 양의 균형을 조절하는 핵심 하이퍼파라미터이므로, γ 후보 집합을 {0.1, 0.3, 0.5, 0.7, 0.9}로 설정하고, 각 후보에 대해 동일한 TriMa 절차를 수행한 뒤 검증 데이터에서의 F1 점수가 가장 높은 γ^*를 선택하였다. γ가 커질수록 최종 채택되는 의사 라벨 수는 감소하지만 라벨 신뢰도는 증가하고, γ가 작아질수록 더 많은 샘플을 확보할 수 있으나 오라벨 유입 위험이 증가한다. 따라서 본 연구는 후보 γ들에 대한 검증 성능 비교를 통해 데이터 특성에 가장 적합한 γ^*를 선택하였다.

$$\hat{y}_i = \arg\max_j \hat{p}_{ij}, \quad \text{if } \hat{p}_{i,(1)} - \hat{p}_{i,(2)} \geq \gamma \tag{8}$$

마지막으로 식 (8)을 통과하여 최종 의사 라벨이 확정된 샘플들만 모아 새로운 의사 라벨 데이터 집합을 구성한다. 이러한 γ 후보 선택과 조건을 만족하지 못한 샘플에 대해서는 $\hat{y}$를 부여하지 않고 비라벨로 유지한다. 이를 포함한 TriMa의 전체 절차는 의사 코드로 <Figure 2>에 요약하였다.

실험 환경에 따라 추가적인 데이터 증강이 필요하다면, 생성된 집합을 원래의 라벨 데이터와 통합한 뒤, SMOTE를 추가 적용하도록 하였다. SMOTE는 소수 클래스(minority class)에 속하는 데이터 포인트의 주변 영역에서 새로운 합성 샘플을 생성함으로써, 클래스 간 샘플 수 차이를 줄인다. 구체적으로, 각 소수 클래스 샘플 x에 대해 유클리드 거리 기반으로 k-최근접 이웃을 구하고, 이들 이웃 샘플 중 하나를 무작위로 선택하여 두 벡터 사이 선형 보간(linear interpolation)을 수행한다. 이렇게 생성된 합성 샘플은 기존 소수 클래스 데이터에 추가되어, 학습 시 소수 클래스 데이터가 충분히 반영되도록 한다. 이 과정을 통해 클래스별 데이터 분포가 보다 균형을 갖추게 되어, 모델이 특정 클래스에 과도하게 편향되지 않고 모든 클래스에 대해 골고루 학습할 수 있는 환경이 조성된다.

종합적으로, 검증 데이터 기반 softmax 가중치 산출 수식 (5)와 결합 확률 계산, 이어서 margin 기반 필터링 수식 (8)을 활용함으로써, TriMa는 세 가지 상호 보완적인 self-training 전략(*Static, Dynamic, Entropy*)의 장점을 모두 결합하여 성능이 우수한 모델에 더 큰 가중치를 부여하고, 결합된 확률 분포에서

Input: labeled set D_L, unlabeled set D_U, validation set D_V
$\Gamma = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, τ_s, τ_e, $P_\alpha(\cdot)$
Output: final pseudo-labeled set D_{P^*}

Train three models m ∈ {static, dynamic, entropy} using (D_L, D_U, τ_s, τ_e, P_α)
Evaluate each model on D_V to get $F1_{static}$, $F1_{dynamic}$, $F1_{entropy}$
Compute weights w_s, w_d, w_e by Softmax ($F1_{static}$, $F1_{dynamic}$, $F1_{entropy}$)
best F1 ← $-\infty$; $D_{P^*} \leftarrow \varnothing$
for γ in Γ do
  $D_P \leftarrow \varnothing$
  for each x_i in D_U do
    Get $p_i^{(s)}, p_i^{(d)}, p_i^{(e)}$; $\hat{p}_{i,c} \leftarrow w_s p_{i,c}^{(s)} + w_d p_{i,c}^{(d)} + w_e p_{i,c}^{(e)}$
    Let $P_i^{(1)} = \max_j \hat{p}_{ij}$, $P_i^{(2)} = 2nd \max_j \hat{p}_{ij}$, $M_i = P_i^{(1)} - P_i^{(2)}$
    if ($M_i \geq \gamma$) then
      $\hat{y}_i \leftarrow \arg\max_j \hat{p}_{ij}$; $D_P \leftarrow D_P \cup (x_i, \hat{y}_i)$
    end if // else: keep x_i unlabeled
  end for
  Train classifier on D_{train} and compute F1(γ) on D_V
  if F1(γ) > bestF1 then bestF1 ← F1(γ) ; $D_{P^*} \leftarrow D_P$ end if
end for
return D_{P^*}

Figure 2. Pseudo-code of the TriMa algorithm

높은 신뢰도를 보이는 샘플만을 선택하는 메커니즘을 구현한다. 이로써 잘못된 의사 라벨링으로 인한 학습 불안정성과 초기 단계의 과소적합 문제를 완화하며, 동시에 SMOTE를 통해 클래스 불균형을 해소함으로써 예측 모델의 일반화 성능을 최대화한다. 이러한 일련의 과정을 반복함으로써 모델은 점진적으로 더 많은 고품질 의사 라벨을 확보하며, 결과적으로 제한된 라벨 데이터만으로도 강건하고 정확한 수요 예측 모델을 구축할 수 있다. 본 연구에서는 TriMa로 생성된 의사 라벨과 SMOTE로 보강된 데이터를 Teacher-Student 프레임워크에 통합한다. Teacher-Student 구조는 반지도 학습에서 널리 사용되는 방법으로, Teacher 모델이 검증된 의사 라벨을 생성하고 Student 모델이 이를 모사하도록 학습함으로써 초기 의사 라벨의 노이즈를 점진적으로 감소시키는 것을 목표로 한다(Xie *et al.*, 2020).

TriMa로 생성된 의사 라벨 데이터와 SMOTE로 보강된 데이터를 합친 학습용 데이터셋을 먼저 Teacher 모델에 학습시킨다. Teacher 모델 학습 과정에서는 검증 데이터셋의 성능 변화를 모니터링하여, margin 필터링을 통해 한 차례 선별된 샘플 중에서도 실제 학습 과정에서 신뢰도가 낮게 평가되는 샘플을 추가로 제거하고, 필요할 경우 새로운 의사 라벨을 보완하여 학습용 데이터셋을 지속적으로 업데이트한다. 이러한 과

정을 통해 Teacher 모델은 반복적으로 재학습되며, 초기 단계에서 생성된 의사 라벨의 불확실성을 점차 줄여 나간다. 최종적으로, 충분히 정제된 학습용 데이터셋은 Student 모델에 전달되고, Student 모델은 고품질의 라벨 정보와 개선된 클래스 분포를 기반으로 최종 학습을 수행한다. 이 self-training 및 데이터 업데이트 과정은 새롭게 추가되는 고신뢰 의사 라벨의 수와 SMOTE로 보강된 데이터 수가 사전에 설정한 데이터 수에 도달할 때까지 반복되며, 이 시점을 알고리즘의 종료 조건으로 사용한다. 이를 통해, Teacher-Student 구조에서 Teacher 모델이 선별한 신뢰도 높은 의사 라벨과 SMOTE 증강을 통해 확보된 소수 클래스 데이터가 Student 모델로 효과적으로 전이되어, 예측 정확도와 모델 강건성을 동시에 극대화한다.

4. 실 험

4.1 실험 설계

본 연구에서는 실제 BaaS 운영 데이터뿐만 아니라, UCI 데이터인 Seed(Charytanowicz *et al.*, 2010)와 Satellite(Srinivasan, 1993) 데이터셋을 활용하여 TriMa의 성능을 평가하였다. 모든 실험에서는 XGBoost를 베이스라인 모델로 설정하였으며, 설정한 근거는 4.2.1에서 상세히 기술한다. XGBoost의 모델 하이퍼파라미터는 학습률은 0.05, n_estimators는 2000, max_depth는 5로 제한하였다. 또한 subsampling과 L2 정규화를 적용하였으며, 다중 클래스 확률 분포의 최적화를 위해 mlogloss를 사용하였다. 또한 γ는 {0.1, 0.3, 0.5, 0.7, 0.9} 후보 집합에서 검증 데이터의 F1 점수를 최대화하는 값 γ^*로 선택하였는데, 이는 데이터마다 최적의 신뢰도와 증강량 균형점이 달라질 수 있음을 고려하여, margin 민감도는 검증 기반으로 적응적으로 결정하도록 하였다. 각 데이터셋별로 10회 반복 실험을 수행한 후 평균 F1 점수와 분산을 산출하였다. 특히 UCI 데이터셋의 경우 원본이 균형적인 클래스 분포를 갖기 때문에, normalized entropy 값을 기준으로 일부 샘플을 무작위로 제거하여 인위적으로 불균형한 상태를 구성한 뒤 실험을 진행하였다. 이를 통해 제안 기법의 일관된 성능 개선 효과와 다양한 불균형 시나리오에서의 강건성을 검증하였다. 본 연구에서 데이터 클래스 간의 불균형 정도는 normalized entropy 지표로 정의했으며, 다음과 같이 계산된다.

$$H_{norm}(P) = \frac{H(P)}{H_{\max}} = \frac{-\sum_{i=1}^{K} p_i \log_b p_i}{\log_b K} \tag{9}$$

여기서 K는 클래스의 총 개수이며, p_i는 전체 샘플 중 클래스 i가 차지하는 비율을 의미한다. 분자는 클래스 비율에 대한 엔트로피 값을 나타내며, 이를 $\log(K)$로 나누어 정규화함으로써 클래스 수에 무관하게 0과 1 사이의 값으로 스케일링한다. 이때 normalized entropy 값이 0에 가까울수록 특정 클래스가 과도하게 편중된 상태를 나타내며, 값이 1에 가까울수록 각 클래스가 균등하게 분포된 상태임을 의미한다.

실험에 사용된 모델 구현 및 평가를 위해 다음과 같은 계산 환경을 구성하였다. 전체 구현은 Python 3.12.4 환경에서 프로그래밍되었으며, 주요 라이브러리로 Scikit-learn 1.4.2 버전을 사용하였다. 하드웨어는 12th Gen Intel(R) Core(TM) i7-1260P (2.10 GHz) 및 32GB RAM을 갖춘 환경에서 실험을 진행하였다.

(1) Seed 데이터셋

첫 번째 실험에는 3개의 클래스 레이블과 7개의 독립 변수를 갖는 총 210개의 실수형 샘플로 이루어진 Seed 데이터셋을 사용하였다. 원본 데이터는 클래스당 70개씩 균등 분포를 보이나, 본 연구에서는 전체 샘플의 30%를 비라벨 데이터로 설정하고 나머지 70%를 학습용 라벨 데이터로 활용함으로써 인위적인 불균형 시나리오를 조성하였다. 구체적으로, 학습용 라벨 데이터 중 일부를 무작위로 선정하여 제거하고 비라벨로 전환하여, 클래스 분포가 불균형해지도록 하였다. Seed 데이터셋에서의 클래스 불균형 정도를 정량화하기 위해 normalized entropy를 계산하였으며, 그 값을 기준으로 세 구간으로 구분하였다. 각 구간의 경곗값은 0.85와 0.70으로 설정하여, normalized entropy가 0.85 이상인 경우, 0.70~0.85 구간, 0.55~0.70 구간으로 나누었다. 이후 세 구간별로 각각 3가지의 다른 normalized entropy 값을 가진 실험 환경을 구축하여, 각 상황에서 TriMa를 적용했을 때의 예측 성능과 강건성을 10회 반복 실험으로 산출한 평균 F1 점수와 표준편차를 기준으로 평가하였다. 실험마다 XGBoost 모델을 베이스라인으로 사용하였으며, 총 10회 반복 실험을 통해 평균 F1 점수 및 표준편차를 기록하였다.

Seed 데이터 실험에서 TriMa의 하이퍼파라미터는 데이터셋의 상대적으로 작은 샘플 수와 인위적으로 조성된 클래스 불균형 환경에서의 안정적인 증강을 고려하여 설정하였다. 먼저 전체 학습 데이터 중 30%를 비라벨 데이터로 구성한 후, 나머지 라벨 데이터 내부에서 다시 20%를 검증 데이터로 분할하여 TriMa의 결합 가중치 산출 및 margin 임곗값 선택에 활용하였다.

Static 모델의 신뢰도 임곗값 τ_s는 0.95로 설정하여, 학습 초기 단계에서 예측 신뢰도가 충분히 확보되지 않은 샘플이 의사 라벨로 유입되는 것을 억제하였다. *Entropy* 모델의 엔트로피 임곗값 τ_e는 0.40으로 설정하여, 예측 확률 분포의 불확실성이 낮은 샘플만을 선별하도록 하였으며, 이를 통해 소규모 데이터 환경에서 발생하기 쉬운 오라벨 누적 문제를 완화하고자 하였다. *Dynamic* 모델은 최대 예측 확률 분포에 대한 상위 백분위수 기준을 사용하며, 초기 퍼센타일을 0.95로 설정하여 초기 반복에서는 특히 엄격한 기준으로 의사 라벨 후보를 제한하고, 반복이 진행됨에 따라 기준을 점진적으로 강화하도록 설계하였다. 이러한 설정은 Seed 데이터처럼 샘플 수가 제한된 상황에서, 모델이 충분히 안정화되기 이전에 과도한 의사

라벨 증강이 발생하는 것을 방지하기 위한 것이다. 반복 증강 횟수는 최대 5회로 제한하였고, 각 반복에서 전략별로 최종적으로 margin 필터를 통과한 샘플은 최대 100개까지만 반영하여 증강 규모의 상한을 두었다. 이는 Seed 데이터셋의 크기 대비 증강량이 과도해지는 것을 방지하고, 반복 실험 간 성능 변동성을 안정적으로 제어하기 위함이다.

(2) Satellite 데이터셋

두 번째 실험에서는 6개의 클래스 레이블과 36개의 독립변수를 갖는 6,434개의 정수형 샘플로 이루어진 Satellite 데이터셋을 활용하였다. 이 데이터는 클래스 간 샘플 수 차이가 최대 약 14%에 달하는 불균형 특성을 지니고 있으며, 특히 일부 소수 클래스가 전체 학습에 미치는 영향이 큰 편이다.

본 연구에서는 전체 클래스 비율을 그대로 유지하되, 라벨이 있는 데이터의 비율을 10%, 30%, 50%, 70%, 90%의 다섯 단계로 순차적으로 변경하였다. 예를 들어, 전체 샘플 중 10%만 라벨링을 유지하여, 라벨 데이터로 사용하고 나머지 90%를 비라벨 데이터로 설정한 뒤, 라벨 데이터 비율을 늘려가며 실험을 진행하였다. 각 비율 단계에서 학습용 라벨 데이터는 SMOTE 적용 모델과 TriMa 적용 모델을 각각 학습시키고, 검증 결과로부터 평균 F1 점수와 표준편차를 산출하여 두 기법 간의 성능 차이를 비교 분석하였다. 이를 통해, 데이터 규모 및 라벨 비율 변화에 따른 예측 정확도 및 모델 강건성을 종합적으로 검증하였다.

Satellite 데이터를 사용한 실험에서는 피처들의 수를 조절하기 위해 PCA를 사용하여 충분한 설명력을 지니는 주성분 수를 5로 고정하였으며, 학습 데이터의 20%를 검증 데이터로 사용하여 TriMa의 결합 가중치 산출 및 γ 값 선택에 사용하였다.

Static 모델의 신뢰도 임곗값 τ_s는 0.95로 설정하여 학습 초기에 저신뢰 예측이 의사 라벨로 유입되는 것을 억제하였고, *Entropy* 모델의 초기 엔트로피 임곗값 τ_e는 0.40으로 설정하여 불확실성이 충분히 낮은 샘플만 선별되도록 하였다. *Dynamic* 모델은 최대 예측확률 분포의 상위 백분위수 기준을 사용하며 초기 퍼센타일을 0.95로 두어 초기 단계에서 특히 엄격한 컷오프를 적용하도록 하였고, 반복이 진행되면서 기준을 강화하여 모델이 안정화될수록 더욱 정제된 샘플이 선택되도록 설계하였다. 반복 증강은 Satellite 데이터의 크기 대비 증강량이 과도해지지 않도록 최대 5회로 제한하였으며, 최종 결합 및 margin 필터를 통과한 샘플은 최대 500개까지만 반영하여 증강 규모의 상한을 두었다.

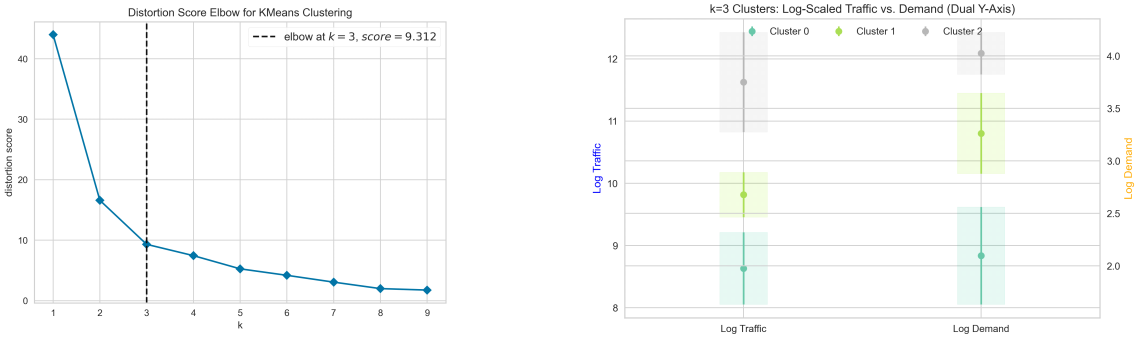
(3) 실제 BaaS 수요 데이터셋

초기 배터리 교환소는 설치 지점 수가 적고 일부 교환소에 수요가 집중되어 있어, 연속형 회귀 모델만으로는 안정적인 예측이 어렵다. 이를 보완하기 위해 본 연구에서는 일평균 배터리 교환 횟수와 전기 이륜차 교통량 데이터를 바탕으로 지역을 수요 수준별로 군집화한 후, 이를 분류 문제로 전환하여 접근하였다. 먼저 원시 데이터를 로그 스케일로 변환해 분산을 완화하고 이상치를 줄인 뒤, K-means 알고리즘을 적용하여 저, 중, 고 세 개의 수요 그룹으로 구분한다. K-means 클러스터링은 각 군집 C_j 내 데이터 포인트 x_i와 군집 중심 μ_j 간 거리 제곱합을 최소화하는 다음의 목적 함수를 최적화한다.

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \tag{10}$$

이 과정을 통해 얻은 클러스터 라벨은 이후 분류 모델의 학습 및 예측 대상 클래스로 활용되어, 데이터가 제한적이거나 불균형한 환경에서도 보다 견고한 수요 예측을 가능하게 한다.

Elbow method를 적용한 결과(<Figure 3> (a)), 군집 수 k의 최적 값은 3으로 결정되었다. 구체적으로는 k를 2부터 여러 값에 대해 왜곡도(distortion score)를 계산한 뒤, k가 증가함에 따라 왜곡도 감소 폭을 비교하여 급격히 감소하는 지점을 elbow point로 정의하였다. 본 데이터에서는 k가 3에서 곡선 기울기의 변화가 가장 크게 나타나, 이를 최적 군집 수로 채택하였다. 이러한 기준은 배터리 교환형 전기 이륜차 운전 패턴 분석에서 elbow method를 적용한 선행연구와 동일한 방식이다(Choi *et al.*, 2025). 이후 각 클러스터별로 일평균 배터리 교환 횟수와 통행량 수준을 시각화한 결과(<Figure 3> (b)), 세 개의 클러스터가 서로 뚜렷한 차이를 보였다. 저수요, 중수요, 고수요로 구



(a) Elbow method for K-means (b) Traffic vs. demand clusters

Figure 3. K-means clustering results

분된 이 군집 정보는 분류 모델 학습에 활용되며, 특히 고수요(Cluster 2)에 속하는 지역을 우선적으로 신규 BSS 설치 후보지로 선정하는 전략을 수립할 수 있다.

본 연구에서는 국토교통부의 이륜자동차 신고 현황 데이터(Ministry of Land, 2023), 행정구역별 국토이용 현황 데이터(National Data Office, 2024), KOSIS의 가구원수별 가구 데이터를 독립변수로 활용하였으며, 종속변수는 실제 BaaS 기업 배터리 교환소의 일평균 교환 횟수 데이터, 전기 이륜차 주행 데이터를 기반으로 한 교통량 데이터를 바탕으로, 클러스터링을 통해 얻은 저, 중, 고수요 수준의 클러스터 라벨을 사용하였다. 첫 번째 실험에서 사용할 수 있는 라벨 데이터는 전체 56개 구역 중 22개의 구역에 해당하는 데이터이며, 34개의 구역은 수요가 라벨링 되어있지 않아 의사 라벨링의 대상으로 활용하였다. 이와 같은 데이터 구성은 신규 BSS 확장 초기 단계에서 일반적으로 관찰되는 운영 현실을 반영한다. 즉, 실제 배터리 교환 로그로부터 수요 라벨을 확보할 수 있는 지역은 제한적이지만, 행정구역별 국토이용, 가구 구성, 이륜차 신고 현황과 같은 설명변수는 전 지역에 대해 상대적으로 풍부하게 존재한다. 따라서 본 문제는 라벨이 부족하고 비라벨 공변량이 풍부한 전형적인 반지도 학습 환경이며, 제한된 22개 라벨만으로 분류 경계를 학습하면 고수요 클래스에 대한 학습 라벨이 부족해 일반화가 불안정해질 수 있다. 결과적으로 비라벨 34개 구역을 활용해 추가 정답 라벨을 생성하는 의사 라벨링은 단순한 성능 향상 기법이 아니라, 초기 데이터 제약을 극복하기 위한 필수적인 학습 구성 요소로 볼 수 있다.

이러한 데이터 환경에서 TriMa와 단순 SMOTE 데이터 증강에 대해 각 기법이 불균형 데이터를 어떻게 보완하는지 비교하였다. SMOTE는 기존 라벨 데이터 내부에서만 합성 샘플을 생성하므로, 라벨이 극히 제한된 초기 단계에서는 새로운 지역의 정황 정보를 학습에 반영하기 어렵다. 반면 의사 라벨링은 비라벨 구역의 공변량을 학습 과정에 편입시켜, 데이터 규모뿐 아니라 적용 도메인의 범위를 함께 확장한다는 점에서 초기 확장 환경에 적합하다. 모델 성능 평가는 10회 반복 실험을 통해 산출한 평균 F1 점수와 표준편차를 사용하였으며, 학습곡선 하부 면적(AULC, area under the learning curve)을 추가 지표로 활용해 학습 속도와 안정성 측면에서 TriMa의 성능을 평가하였다. 본 연구에서 AULC는 TriMa를 비롯한 데이터 증강을 통해 훈련 데이터 크기를 점진적으로 증가시키며 각 시점에서 F1 점수를 측정해 얻은 학습곡선을 기반으로, 해당 곡선 아래 면적을 적분한 값으로 정의하였다. 따라서 AULC가 높다는 것은 학습 데이터 단계에서도 성능이 빠르게 상승했다는 점, 그리고 전체 데이터 증가 구간에서 비교적 안정적으로 높은 성능을 유지했다는 점을 모두 의미한다. 이러한 정의를 기반으로 AULC를 측정함으로써, TriMa가 초기 학습 단계에서 얼마나 빠르게 성능을 확보하며, 전체 실험 구간에서 얼마나 일관적인 성능을 보였는지를 정량적으로 평가하였다.

BaaS 데이터를 활용한 실험에서는 학습 데이터 규모 증강에 따른 성능 변화를 분석하기 위해, 각 반복에서 초기 라벨 학습 세트 크기를 기준으로 학습 데이터 크기를 10개씩 증가시키며 최대 50개까지 확장하는 학습 곡선 실험을 수행하였다. TriMa의 하이퍼파라미터는 데이터셋의 상대적으로 작은 샘플 수와 인위적으로 조성된 클래스 불균형 환경에서의 안정적인 증강을 고려하여 설정하였다. 먼저 전체 학습 데이터 중 30%를 비라벨 데이터로 구성한 후, 나머지 라벨 데이터 내부에서 다시 20%를 검증 데이터로 분할하여 TriMa의 결합 가중치 산출 및 margin 임곗값 선택에 활용하였다.

Static 모델의 신뢰도 임곗값 τ_s는 0.95로 설정하여, 학습 초기 단계에서 예측 신뢰도가 충분히 확보되지 않은 샘플이 의사 라벨로 유입되는 것을 억제하였다. *Entropy* 모델의 초기 엔트로피 임곗값 τ_e는 0.40으로 설정하여, 예측 확률 분포의 불확실성이 낮은 샘플만을 선별하도록 하였으며, 이를 통해 소규모 데이터 환경에서 발생하기 쉬운 오라벨 누적 문제를 완화하고자 하였다. *Dynamic* 모델은 최대 예측 확률 분포에 대한 상위 백분위수 기준을 사용하며, 초기 퍼센타일을 0.95로 설정하여 초기 반복에서는 특히 엄격한 기준으로 의사 라벨 후보를 제한하고, 반복이 진행됨에 따라 기준을 점진적으로 강화하도록 설계하였다. 이러한 설정은 BaaS 데이터처럼 샘플 수가 제한된 상황에서, 모델이 충분히 안정화되기 이전에 과도한 의사 라벨 증강이 발생하는 것을 방지하기 위한 것이다.

4.2 실험 결과

(1) 하이퍼파라미터 및 베이스라인 모델

먼저 margin 필터링 임곗값 변화에 따른 성능 변화를 확인하기 위해 Satellite 데이터셋에서 데이터 스케일을 90%, 50%, 10%로 샘플링하여 사용하였고, 각 스케일에 대해 비라벨 비율을 각각 10%, 50%, 90%로 구성하였다. 이때 margin은 {0, 0.1, 0.3, 0.5, 0.7, 0.9}로 변화시키며 10개 시드에 대해 반복 실험을 수행하였다. 본 실험은 margin 값이 의사 라벨 선택에 미치는 영향만을 관찰하는 데 목적이 있으므로, 의사 라벨로 추가되는 데이터 수에 대한 별도의 절대적 상한을 두지 않고 margin 조건만으로 선택을 제한하였다. 또한 해당 설정에서는 비라벨 데이터가 충분하여 TriMa를 통해 선택되는 의사 라벨이 증강의 대부분을 차지하도록 구성하였다. 이러한 구성을 통해 margin 값 변화에 따른 TriMa의 선택 특성이 최종 성능에 반영되는 양상을 직접적으로 확인하였다.

<Table 1>에서 확인할 수 있듯이, 비라벨 데이터가 충분하여 의사 라벨 후보군이 넓게 형성되는 조건에서는 margin이 커질수록 성능이 상승하는 경향이 나타난다. 전체 데이터의 50% 비율 데이터 조건에서는 낮은 margin 대비 높은 margin에서 F1 점수가 전반적으로 개선되며, 특히 γ값이 0.7 부근에서 가장 높은 성능이 관찰된다. 반면 데이터가 충분하지 않거나 라벨 비율이 낮아 기본 분류기의 예측 불확실성이 상대적으로 커지는 조건에서는 margin 증가가 항상 성능 향상으로 이어지지 않

Table 1. Average F1 scores by margin across different data scales

Data Scale \ Margin	0.0	0.1	0.3	0.5	0.7	0.9
90% Data	0.7477	0.7514	0.7586	0.7613	**0.7624**	0.7616
50% Data	0.7439	0.7430	0.7497	0.7590	**0.7711**	0.7629
10% Data	0.7149	0.7133	0.7061	0.7204	0.7309	**0.7343**

는다. 90% 데이터 조건에서는 γ 증가에 따라 성능이 증가하다가 매우 큰 margin에서 다시 감소하는 형태가 나타나며, 10% 데이터 조건에서도 margin 변화에 따른 성능의 단조 증가 관계가 뚜렷하게 유지되지 않는다. 이는 margin을 크게 설정할수록 선택되는 의사 라벨이 더 보수적으로 제한되는 반면, 동시에 학습에 활용 가능한 추가 데이터의 양이 감소할 수 있기 때문에, 비라벨 후보 풀이 충분하지 않은 경우에는 오히려 성능 향상이 제한되거나 감소할 수 있음을 시사한다.

Table 2. Performance comparison of machine learning models with and without TriMa

	PLAIN	TriMa
RF	0.7665 ± 0.0124	0.8045 ± 0.0165
LGBM	0.8714 ± 0.0084	0.8753 ± 0.0070
XGB	0.8723 ± 0.0081	**0.8767** ± 0.0053

다음으로 정형 데이터 분류에서 널리 사용되는 머신러닝 모델 간 비교를 통해 베이스라인 모델 선정의 타당성을 확인하였다. Satellite 데이터셋 90% 샘플링 조건에서 의사 라벨 생성 수를 1500개로 제한한 뒤, Random Forest(RF), LightGBM (LGBM), XGBoost(XGB)를 각각 기본 분류기로 사용하여 증강을 사용하지 않는 PLAIN과 TriMa 적용의 성능을 비교하였다(<Table 2>). PLAIN 결과만 비교하면 부스팅 계열(LGBM, XGB)이 RF 대비 높은 기본 성능을 보이며, 이는 정형 데이터 환경에서 부스팅 계열이 강한 성능을 보이는 경향과 일치한다. 다만 TriMa 적용에 따른 성능 향상 폭은 모델별로 차이가 존재한다. RF는 0.7665에서 0.8045로 상승하여 개선 폭이 상대적으로 크게 나타나는 반면, LGBM과 XGB는 기본 성능이 이미 높은 상태에서 TriMa 적용 시 추가 개선 폭이 상대적으로 제한적이지만 성능이 안정적으로 개선되는 것이 확인된다. 특히 XGB는 TriMa 적용 후 평균 성능이 가장 높고 표준편차도 가장 낮게 관찰되어, 이후 본 논문의 주요 실험에서는 XGB를 기본 분류기로 채택하였다. 요약하면, 데이터가 충분한 조건에서 TriMa는 RF와 같이 상대적으로 보수적인 트리 앙상블 모델에서도 유의미한 성능 향상을 제공하며, 부스팅 계열 모델에서는 높은 기본 성능 위에서 추가 개선을 안정적으로 확보하는 경향을 보인다. 이러한 결과는 이후 실험에서 XGB 기반 설정을 일관되게 사용하는 근거로 활용한다.

(2) Seed 데이터

Seed 데이터셋에서 TriMa의 효과를 분석하기 위해, 클래스 불균형도를 normalized entropy 값으로 세 구간으로 나누어 실험을 수행하였다. 원본 Seed 데이터는 클래스당 샘플 수가 균등하였으나, 본 실험에서는 전체 샘플의 30%를 비라벨로 설정하여 의도적으로 불균형 시나리오를 조성하였다. 이렇게 만들어진 학습용 데이터에 대해, 클래스 비율을 각각 1:1:1, 1:0.6:0.4, 1:0.8:0.3(첫 구간), 1:0.3:0.3, 1:0.4:0.2, 1:0.5:0.1(두 번째 구간), 1:0.15:0.2, 1:0.15:0.15, 1:0.2:0.1(세 번째 구간)로 설정하고, 각 비율마다 10회 반복 실험을 수행하여 SMOTE만 적용한 모델과 TriMa를 적용한 모델의 평균 F1 점수 및 표준편차를 비교하였다.

Table 3. Average F1 scores by imbalance scenario on Seed dataset

Normalized Entropy		1 : 1 : 1	1 : 0.6 : 0.4	1 : 0.8 : 0.3
0.85 ~ 1.00	SMOTE	0.9101	0.8857	0.8812
	TriMa	**0.9225**	0.8887	**0.8973**
		1 : 0.3 : 0.3	1 : 0.4 : 0.2	1 : 0.5 : 0.1
0.70 ~ 0.85	SMOTE	0.8559	0.8790	0.8817
	TriMa	0.8591	**0.8928**	**0.8945**
		1 : 0.15 : 0.2	1 : 0.15 : 0.15	1 : 0.2 : 0.1
0.55 ~ 0.70	SMOTE	0.8250	0.8476	0.8578
	TriMa	0.8267	0.8496	**0.8692**

Table 4. Standard deviation of F1 scores by imbalance scenario on Seed dataset

Normalized Entropy		1 : 1 : 1	1 : 0.6 : 0.4	1 : 0.8 : 0.3
0.85 ~ 1.00	SMOTE	0.0421	0.0420	0.0505
	TriMa	**0.0311**	**0.0286**	**0.0345**
		1 : 0.3 : 0.3	1 : 0.4 : 0.2	1 : 0.5 : 0.1
0.70 ~ 0.85	SMOTE	0.0384	0.0477	0.0421
	TriMa	0.0318	0.0443	0.0339
		1 : 0.15 : 0.2	1 : 0.15 : 0.15	1 : 0.2 : 0.1
0.55 ~ 0.70	SMOTE	0.0535	0.0622	0.0484
	TriMa	0.0512	0.0604	**0.0261**

<Table 3>과 <Table 4>를 종합하면, TriMa는 normalized entropy가 감소하여 클래스 불균형이 심화되는 환경에서도 성능

저하를 효과적으로 완화하며 안정적인 예측 성능을 유지함을 확인할 수 있다. 특히 normalized entropy가 0.55-0.70에 해당하는 극심한 불균형 구간에서도 TriMa는 SMOTE 대비 평균 F1 점수가 급격히 붕괴되지 않고, 전반적으로 동등하거나 더 우수한 수준을 유지하였다. 구체적으로 1 : 0.15 : 0.2 및 1 : 0.15 : 0.15와 같이 소수 클래스 비중이 매우 낮은 조건에서는 SMOTE 대비 성능 저하 없이 안정적으로 유지되었다. 더욱이 1 : 0.2 : 0.1과 같이 불균형이 한쪽 클래스에 집중된 조건에서는 TriMa가 평균 F1 점수를 0.8578에서 0.8692로 향상시켜, 극심한 불균형 상황에서도 성능 개선 효과를 확인할 수 있었다. 이러한 경향은 TriMa가 불균형이 심해질수록 발생하기 쉬운 오라벨 누적과 학습 불안정성을 margin 기반 확률 결합과 의사 라벨 선별을 통해 억제함으로써, 성능 하락을 완만하게 만들고 예측 품질을 유지했기 때문으로 해석할 수 있다. 즉, TriMa는 극단적인 클래스 비율 조건에서도 단순히 평균 성능을 끌어올리는 데 그치기보다, 성능 붕괴를 방지하는 완충 장치로서 작동한다.

이러한 특성은 모델 안정성 측면에서 더욱 뚜렷하게 나타난다. 동일한 normalized entropy 0.55-0.70 구간에서 TriMa는 SMOTE 대비 표준편차를 일관되게 감소시켰으며, 특히 1 : 0.2 : 0.1 조건에서는 표준편차가 0.0484에서 0.0261로 약 46.1% 감소하여 반복 실험 간 변동성이 크게 완화되었다. 이는 극심한 불균형 환경에서 소수 클래스의 학습 결과가 실험마다 크게 흔들릴 수 있음에도 불구하고, TriMa가 고신뢰 의사 라벨만을 선택적으로 추가함으로써 모델 학습을 보다 안정적인 방향으로 유도했음을 의미한다.

요약하면, normalized entropy가 0.70 이하로 감소하는 극심한 클래스 불균형 환경에서도 TriMa는 SMOTE 대비 평균 F1 점수를 안정적으로 유지하거나 소폭 개선하였으며, 특히 반복 실험 간 변동성을 최대 약 46%까지 감소시켜 모델 강건성을 크게 향상시켰다. 이러한 결과는 TriMa가 클래스 불균형이 심화될수록 단순 데이터 증강 기법의 한계를 보완하며, 예측 성능 유지와 학습 안정성 확보라는 두 측면에서 모두 효과적으로 작동하는 방법론임을 보여준다.

(3) Satellite 데이터

Satellite 데이터셋의 클래스 분포는 1 : 0.46 : 0.89 : 0.41 : 0.46 : 0.98이고, 이로 인한 normalized entropy는 약 0.961로 매우 균등한 편에 속한다. 이러한 환경에서 다양한 라벨링 비율(라벨 데이터 크기 90%, 70%, 50%, 30%, 10%)마다 SMOTE만 적용한 경우와 TriMa를 적용한 경우를 비교 평가한 결과를 <Table 5>에 정리하였다.

<Table 5>의 평균 F1 점수를 살펴보면, 데이터 규모가 감소함에 따라 두 방법 모두 성능이 점진적으로 하락하는 경향을 보이지만, TriMa는 전 구간에서 SMOTE 대비 최소한 동등하거나 소폭 높은 성능을 유지하였다. 구체적으로 90% 라벨링에서는 SMOTE 0.8822, TriMa 0.8826으로 거의 동일한 수준이었고, 70%에서도 0.8752 대비 0.8763으로 유사한 성능을 보였다. 라벨 데이터가 더 줄어든 50%, 30%, 10%에서도 TriMa는 각각 0.8631, 0.8543, 0.8223으로 SMOTE 대비 소폭 높은 평균 성능을 유지하여, 데이터가 제한되는 상황에서 성능 저하를 완만하게 만드는 경향을 확인할 수 있다.

한편 모델 강건성을 나타내는 표준편차 측면에서는 TriMa가 전반적으로 뚜렷한 개선을 보였다. 90% 라벨링에서 SMOTE의 표준편차는 0.0054인 반면 TriMa는 0.0028로 약 48.1% 감소하였고, 70% 라벨링에서도 SMOTE 0.0073 대비 TriMa 0.0062로 약 15.1% 감소하였다. 특히 50% 라벨링에서는 SMOTE 0.0075에서 TriMa 0.0034로 약 54.7% 감소하였으며, 30% 라벨링에서는 0.0084에서 0.0034로 약 59.5% 감소하여 반복 실험 간 변동성이 크게 줄어드는 양상이 두드러졌다. 라벨이 극히 적은 10% 구간에서도 TriMa는 0.0096으로 SMOTE의 0.0130 대비 약 26.2% 낮은 변동성을 보여, 라벨 부족 상황에서 예측 결과가 더 일관되게 나타남을 확인하였다.

요약하면, Satellite 데이터셋의 normalized entropy가 0.961로 비교적 균등한 상황에서도 TriMa는 평균 F1 점수 측면에서 SMOTE와 거의 동등하거나 소폭 향상된 성능을 유지하였고, 특히 라벨 데이터 비율이 50%와 30%로 감소하는 구간에서 표준편차가 각각 약 54.7%, 59.5%까지 감소하여 모델 강건성을 크게 개선하였다. 이러한 결과는 데이터 규모가 줄어들수록 불확실성이 증가하는 조건에서도 TriMa가 SMOTE 대비 예측 결과의 일관성을 높여, 보다 안정적이고 신뢰 가능한 수요 예측을 가능하게 함을 보여준다.

(4) 실제 BaaS 수요 데이터

본 절에서는 실제 BaaS 수요 데이터처럼 표본 수가 매우 제한적이고 클래스 불균형이 동시에 존재하는 환경에서 TriMa의 데이터 증강 효과를 분석하였다. 본 실험에서는 초기 학습 데이터의 크기를 10으로 설정하였는데, 이 시점에서는 모든 방법이 동일한 라벨 데이터만을 사용하므로 평균 F1 점수가 동일하게 나타난다. 이는 아직 데이터 증강이나 의사 라벨링

Table 5. Performance comparison on Satellite dataset by data scale

Data Scale		90%	70%	50%	30%	10%
F1 Score	SMOTE	0.8822	0.8752	0.8606	0.8526	0.8191
	TriMa	0.8826	0.8763	0.8631	0.8543	0.8223
Std Dev	SMOTE	0.0054	0.0073	0.0075	0.0084	0.0130
	TriMa	**0.0028**	0.0062	**0.0034**	**0.0034**	**0.0096**

이 적용되지 않은 순수 초기 학습 단계로, 이후 관찰되는 성능 차이가 각 방법의 증강 전략에서 기인함을 명확히 해준다.

Table 6. Average F1 scores of SMOTE and TriMa on the BaaS dataset

	10	20	30	40	50
SMOTE	0.4601	0.4513	0.4202	0.4156	0.4448
TriMa	0.4601	**0.4616**	**0.4628**	**0.4904**	**0.5681**

학습 데이터 규모를 점진적으로 증가시키면서 비교한 결과(<Table 6>), BaaS 데이터와 같이 극히 소규모이면서 불균형한 환경에서는 단순 합성 기반 증강인 SMOTE가 항상 성능 향상으로 이어지지 않는다는 점이 확인된다. 실제로 학습 크기가 20에서 30으로 증가하는 구간에서 SMOTE의 평균 F1 점수는 0.4513에서 0.4202로 오히려 감소하였으며, 이후 40에서도 낮은 수준을 유지한다. 이는 데이터가 충분히 많지 않은 상황에서 합성 샘플이 실제 데이터 분포를 정밀하게 반영하지 못할 경우, 결정 경계를 왜곡하여 학습 성능을 저하시킬 수 있음을 시사한다. 즉, BaaS와 같은 소표본, 고불균형 데이터에서는 단순히 데이터의 양을 늘리는 방식이 반드시 성능 개선으로 이어지지 않는다.

반면 TriMa는 동일한 조건에서 성능이 급격히 붕괴되는 현상을 보이지 않고, 학습 데이터가 누적됨에 따라 점진적으로 성능이 개선되는 경향을 보인다. 특히 학습 크기 30 이후부터는 SMOTE와의 성능 격차가 뚜렷하게 확대되며, 최종적으로 학습 크기 50에서는 TriMa가 SMOTE 대비 큰 폭으로 높은 평균 F1 점수를 기록한다(<Table 6>). 이는 TriMa가 비라벨 데이터로부터 신뢰도가 충분히 높은 샘플만을 선별적으로 통합함으로써, 데이터 양의 증가가 곧바로 학습 품질의 저하로 이어지는 것을 방지하고, 제한된 데이터 환경에서도 안정적인 성능 향상을 달성했음을 의미한다.

Table 7. Standard deviation of F1 scores for SMOTE and TriMa on the BaaS dataset

	10	20	30	40	50
SMOTE	0.2139	0.1967	0.1784	0.1622	0.1918
TriMa	0.2139	0.2198	0.1900	0.1771	**0.1586**

이러한 차이는 변동성 측면에서도 확인된다(<Table 7>). 매우 작은 학습 크기에서는 두 방법 모두 시드에 따른 데이터 분할 영향이 커 표준편차가 크게 나타나는데, 이는 소표본 학습의 일반적인 특성이다. 그러나 학습 데이터가 증가함에 따라 SMOTE는 평균 성능이 하락하거나 정체되는 구간이 존재하는 반면, TriMa는 성능이 크게 개선되는 구간에서 표준편차가 함께 감소하는 경향을 보인다. 이는 TriMa가 단순히 평균 성능만을 높이는 것이 아니라, 반복 실험 간 결과를 보다 일관되게 유지함으로써 모델의 강건성을 동시에 확보하고 있음을 보여준다.

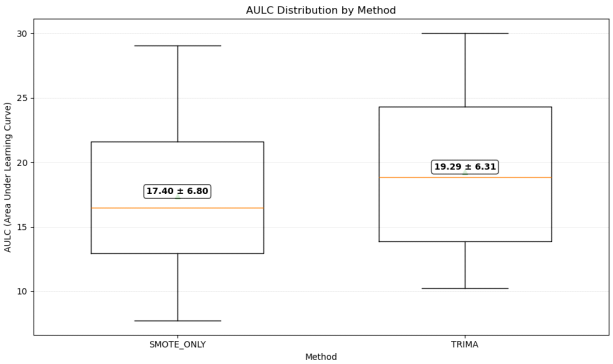


Figure 4. AULC comparison of SMOTE and TriMa on the BaaS Dataset

학습 전 구간에 걸친 성능을 종합적으로 반영한 AULC 비교 결과(<Figure 4>) 역시 이러한 해석을 뒷받침한다. AULC는 특정 학습 크기에서의 단일 성능이 아니라, 학습 데이터가 증가하는 전체 과정에서의 성능 궤적을 반영하는 지표로서, 실제 운영 환경과 같이 라벨 데이터가 점진적으로 확보되는 상황을 평가하는 데 적합하다. <Figure 4>에서 TriMa는 SMOTE 대비 더 높은 AULC 평균을 기록하였으며, 이는 학습 과정 전반에서 성능 저하 구간이 상대적으로 적고, 누적 관점에서 더 안정적인 학습 효율을 달성했음을 의미한다.

종합하면, 실제 BaaS 수요 데이터와 같이 데이터 수가 매우 적고 클래스 불균형이 심한 환경에서는 단순 합성 샘플 생성 방식이 오히려 성능 저하를 초래할 수 있음이 확인되었다. 이에 비해 TriMa는 초기 단계에서는 동일한 출발점을 가지면서도, 학습 데이터가 증가할수록 성능을 안정적으로 개선하고, 누적 성능과 변동성 측면에서도 우수한 특성을 보였다. 이러한 결과는 TriMa가 극심한 데이터 제약 환경에서도 성능 붕괴를 방지하며, 실제 서비스 환경에서 요구되는 안정성과 신뢰성을 동시에 제공할 수 있는 데이터 증강 프레임워크임을 시사한다.

(5) 데이터 노이즈에 대한 강건성 검증

본 절에서는 학습 데이터의 라벨이 일부 오류를 포함하는 상황을 가정하고, (i) 증강 단계에서 생성되는 샘플이 실제 라벨 구조와 얼마나 정합적인지(실험 1), (ii) 라벨 오류가 최종 분류 성능에 어떤 영향을 미치는지(실험 2)를 정량적으로 분석한다. 라벨 노이즈는 실제 라벨의 p만큼 비율의 라벨을 뒤집어 관측 라벨을 구성하는 방식으로 주입하였다. 이후 증강 및 학습 파이프라인은 관측 라벨을 사용하여 수행하되, 성능 평가는 테스트 데이터의 실제 정답을 기준으로 F1 점수를 산출하였다. 실험은 10개의 랜덤 시드 반복으로 진행하였으며, 클래스 비율은 normalized entropy 구간(0.85-1.00, 0.70-0.85,

Table 8. Evaluation of augmented data quality under label noise for each method

Normalized Entropy	p = 0.15		p = 0.30	
	SMOTE	TriMa	SMOTE	TriMa
0.55 ~ 0.70	0.6210 ± 0.1880	**0.7837 ± 0.1402**	0.3701 ± 0.1460	**0.7065 ± 0.1730**
0.70 ~ 0.85	0.6585 ± 0.1835	**0.8339 ± 0.1245**	0.4414 ± 0.1804	**0.7211 ± 0.1681**
0.85 ~ 1.00	0.8222 ± 0.1547	**0.8700 ± 0.0709**	0.6209 ± 0.2128	**0.7529 ± 0.0869**

0.55-0.70)으로 구분하였다. 각 normalized entropy에 따른 결과는 해당 구간에 속하는 3개의 서로 다른 클래스 비율 시나리오 결과를 평균으로 제시하였으며, 본 절에서 사용하는 normalized entropy 정의와 시나리오 구성은 앞선 4.2.2 불균형 실험 설정과 동일하다.

실험 1에서는 라벨 노이즈가 존재할 때 증강 산출물이 실제 라벨 구조를 얼마나 보존하는지를 품질 지표로 측정하였다. 다만 TriMa와 SMOTE는 증강 산출물의 형태가 다르므로 동일한 방식으로 정답 정합성을 정의하기 어렵다. TriMa는 비라벨 풀에서 선택된 샘플에 의사 라벨을 부여하므로, 실험 환경에서 해당 샘플의 실제 라벨을 알고 있을 때 의사 라벨의 정합성을 직접 측정할 수 있다. 이에 TriMa의 품질 지표는 선택된 샘플에 대해 실제 라벨과 일치하는 비율로 정의하였다. 이는 노이즈가 포함된 관측 라벨을 기반으로 학습하더라도 TriMa가 실제 라벨을 얼마나 정확히 복원하여 증강하는가를 직접 반영한다. 반면 SMOTE는 관측 라벨을 이용해 feature space에서 보간하여 synthetic 샘플을 생성하며, 생성된 샘플은 원천적으로 실제 라벨이 정의되지 않는다. 따라서 SMOTE의 경우 실험 1에서 TriMa와 동일한 의미의 정답 일치율을 직접 계산할 수 없으므로, 본 연구에서는 간접 품질 지표로 oracle-consistency를 사용하였다. 먼저 노이즈 주입 전의 깨끗한 라벨로 학습한 oracle 분류기를 별도로 구축하여 SMOTE가 생성한 synthetic 집합에 대해 SMOTE가 부여한 라벨과 oracle 예측이 일치하는 비율을 측정하였다.

<Table 8>에 따르면, p=0.15에서 SMOTE의 oracle-consistency는 0.55-0.70 구간에서 0.6210로 낮고 변동 폭도 큰 반면, TriMa의 의사 라벨 정합성은 0.7837로 더 높게 유지된다. 0.70-0.85 및 0.85-1.00에서도 TriMa 정합성은 각각 0.8339, 0.8700으로 SMOTE 대비 전반적으로 높거나 더 안정적인 경향을 보인다. 노이즈가 더 큰 p=0.3에서는 차이가 더욱 뚜렷해지며, 특히 0.55-0.70에서 SMOTE는 0.3701까지 급격히 감소하는 반면 TriMa는 0.7065로 상대적으로 높은 값을 유지한다. 또한 0.85-1.00에서도 TriMa(0.7529)가 SMOTE(0.6209)보다 높고 표준편차 역시 더 작게 나타난다. 이러한 결과는 노이즈가 포함된 관측 라벨이 증강 과정에 반영될 때, 단순 보간 기반 합성은 오류가 포함된 라벨 구조를 학습 공간에 확장할 위험이 커질 수 있는 반면, TriMa는 앙상블 기반 예측과 margin 조건을 통해 상대적으로 신뢰도가 높은 샘플을 선택함으로써 오류 전파를 완화할 수 있음을 시사한다.

실험 2에서는 동일한 노이즈 주입 조건에서 최종 분류 성능의 변화를 비교하였다. 비교 방법은 증강 없이 학습하는 PLAIN, 관측 라벨 기반으로 클래스 균형화를 수행하는 SMOTE, 그리고 TriMa로 구성하였다. <Table 9>에서 p=0.15를 보면, 강한 편향(0.55-0.70) 조건에서 F1 점수는 PLAIN은 0.7152, SMOTE는 0.7478, TriMa는 0.7617로 나타나 증강 기법이 성능 향상에 기여하며 그중 TriMa가 가장 높은 평균 성능을 보인다. 중간 편향(0.70-0.85)에서도 TriMa는 0.8115로 SMOTE(0.7888) 대비 개선이 확인된다.

노이즈가 더 큰 p=0.3에서는 세 방법 간 차이가 명확해진다. 모든 normalized entropy에서 TriMa가 SMOTE를 상회하며, 0.55-0.70에서 TriMa는 0.6425로 SMOTE(0.6135) 대비 높고, 0.70-0.85에서도 TriMa(0.6884)가 SMOTE(0.6717)를 상회한다. 특히 0.85-1.00 구간에서 TriMa는 0.7310으로 SMOTE(0.7001) 대비 개선 폭이 크게 나타난다. 이는 증강 품질 차이가 최종 분류 성능으로 연결될 수 있음을 뒷받침하며, 라벨 노이즈가 커질수록 관측 라벨 기반 합성은 오류가 포함된 구조를 확장할 위험이 증가하는 반면, TriMa는 선택적 의사 라벨을 통해 상대적으로 신뢰 가능한 샘플을 중심으로 학습 데이터를 확장하여 성능 저하를 방어하는 강건한 프레임워크임을 입증한다.

Table 9. F1 Scores of each method under label noise conditions

Normalized Entropy	p = 0.15			p = 0.30		
	PLAIN	SMOTE	TriMa	PLAIN	SMOTE	TriMa
0.55 ~ 0.70	0.7152 ± 0.1040	0.7478 ± 0.1075	**0.7617** ± 0.1141	0.6015 ± 0.1381	0.6135 ± 0.1333	**0.6425** ± 0.1595
0.70 ~ 0.85	0.7889 ± 0.0862	0.7888 ± 0.0866	**0.8115** ± 0.0894	0.6714 ± 0.1181	0.6717 ± 0.1194	**0.6884** ± 0.1063
0.85 ~ 1.00	0.8116 ± 0.0629	0.8237 ± 0.0638	0.8195 ± 0.0622	0.6796 ± 0.0678	0.7001 ± 0.0718	**0.7310** ± 0.0732

5. 결 론

본 연구는 라벨 데이터가 부족하고 클래스 불균형이 심한 BSS 수요 예측 환경에서, 정확도와 강건성을 동시에 확보하기 위한 새로운 데이터 증강 기법인 TriMa를 제안하였다. 일평균 교환량 수준에 따라 저(low), 중(medium), 고(high)의 세 개 범주로 구성된 분류 문제에 대해, TriMa는 *Static, Dynamic, Entropy*의 세 가지 상호 보완적인 self-training 기준을 병렬로 적용하였다. 또한 검증 데이터에서 산출된 F1 점수 기반의 softmax 가중치를 이용해 모델별 기여도를 적응적으로 반영한 결합 확률을 계산하였다. 이를 통해 margin 기반 필터링으로 신뢰도가 높은 샘플만을 최종 의사 라벨로 채택하고, SMOTE를 결합하여 소수 클래스 데이터를 보강함으로써, 초기 단계에서 발생하기 쉬운 의사 라벨 노이즈와 클래스 불균형 문제를 동시에 해결하였다.

실험을 통해 TriMa가 실제 BaaS 운영 데이터뿐 아니라 인위적인 데이터 부족 및 편중이 적용된 UCI 공개 데이터에서도 일관된 성능 우위를 보임을 확인하였다. Seed 데이터에서는 심각한 불균형 구간에서 평균 F1을 향상시키고 반복 실험 간 표준편차를 크게 낮추었으며, Satellite 데이터에서는 라벨 데이터양이 줄어들수록 오히려 TriMa의 강건성이 더욱 두드러졌다. 실제 BaaS 수요 데이터 실험에서도 TriMa는 SMOTE와 비교하여 유의미한 성능 개선 효과와 안정적인 AULC를 달성하였다. 이로써 TriMa가 제한된 데이터 환경에서 예측 정확도와 반복 실험 간 결과 일관성을 동시에 높일 수 있음을 실험적으로 증명하였다.

더 나아가 라벨 노이즈가 존재하는 조건에서도 TriMa는 SMOTE 대비 증강 과정에서의 라벨 신뢰도를 더 높게 유지하였고, 노이즈 수준이 증가할수록 F1 성능 저하를 더 효과적으로 완화함을 실험을 통해 확인하였다. 이는 TriMa가 앙상블 기반 결합 확률과 margin 필터링을 통해 저신뢰 의사 라벨의 채택을 억제함으로써, 증강 단계에서 발생하는 오류 전파를 구조적으로 방지할 수 있음을 의미한다.

그러나 TriMa는 세 개의 self-training 모델을 병렬 실행하고, 반복적인 margin 필터링 및 확률 가중 결합 단계를 거치므로, 데이터 규모가 커질수록 연산량이 증가하는 계산 복잡도의 한계를 지닌다. 특히 self-training의 반복적 특성상 전체 학습 시간은 전체 데이터 크기와 반복 횟수에 의존적이다. 따라서 대규모 실시간 환경에 적용하기 위해서는 self-training 단계의 경량화, GPU 가속 및 분산 학습 기법 도입 등을 통해 연산 효율성을 개선할 필요가 있다. 또한, 본 연구에서는 지역 간 상호작용이나 시공간적 의존성을 충분히 반영하지 못했다는 한계가 있으며, 급격한 수요 패턴 변화에 실시간으로 적응하는 기능도 보완이 필요하다.

향후 연구에서는 TriMa 프레임워크를 경량화하여 대규모 서비스 환경에서도 실시간 수요 예측이 가능하도록 개선할 예정이다. 구체적으로, 의사 라벨링 과정을 효율화하거나 그래프 신경망을 도입하여 지역 간 공간적, 시공간적 의존성을 학습하는 방향으로 모델을 고도화하고자 한다. 이를 통해 전기 이륜차 배터리 교환소 인프라의 효율적 운영 및 확장 기획에 실질적인 기여를 제공할 수 있을 것으로 기대된다.

참고문헌

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020), Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning, *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN 2020)*, IEEE, Glasgow, United Kingdom, 1-8.

Cerqueira, V., Moniz, N., Inácio, R., and Soares, C. (2024), Time Series Data Augmentation as an Imbalanced Learning Problem, in Santos, M. F., Machado, J., Novais, P., Cortez, P., and Moreira, P. M. (eds.), Progress in Artificial Intelligence, *EPIA 2024, Lecture Notes in Computer Science*, Vol. 14968, Springer, Cham, Switzerland, 335-346.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P., and Lukasik, S. (2010), Seeds [Dataset], UCI Machine Learning Repository, University of California, Irvine, CA, USA.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), SMOTE: Synthetic Minority Over-Sampling Technique, *Journal of Artificial Intelligence Research*, **16**, 321-357.

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. (2023), SoftMatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations*.

Choi, M., Ko, G. J., Boongoen, T., Iam-On, N., Zu, S., and Cheong, T. (2025), Driving-Pattern-Based Ensemble Clustering for SOC Prediction in Battery-Swapping Electric Two-Wheeled Vehicles, *IEEE Access*, **13**, 192006-192022.

Feng, Y. and Lu, X. (2022), Deployment and Operation of Battery Swapping Stations for Electric Two-Wheelers Based on Machine Learning, *Journal of Advanced Transportation*, Article 8351412.

Guo, L. Z. and Li, Y. F. (2022), Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding, In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022), Proceedings of Machine Learning Research*, Vol. 162, PMLR, 8082-8094.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005), Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang (Eds.), *International Conference on Intelligent Computing*, 878-887.

He, H. and Hong, Y. (2025), TrustMatch: Mitigating pseudo-label bias in semi-supervised learning with trust-aware refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 594-603.

Hu, X., Zhang, Z., Fan, Z., Yang, J., Yang, J., Li, S., and He, X. (2024). GCN-Transformer-based spatio-temporal load forecasting for EV battery swapping stations under differential couplings, *Electronics*, **13**(17), 3401.

Hua, M., Pereira, F. C., Jiang, Y., Chen, X., and Chen, J. (2025), Transfer learning for cross-modal demand prediction of bike-share and public transit, *Journal of Intelligent Transportation Systems*, **29**(6), 640-653.

Kenaka, S. P., Cakravastia, A., Ma'ruf, A., and Cahyono, R. T. (2025), Enhancing Intermittent Spare Part Demand Forecasting: A Novel

Ensemble Approach with Focal Loss and SMOTE, *Logistics*, **9**(1), 25.
Lee, D.-H. (2013), Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, in *Proceedings of the Workshop on Challenges in Representation Learning, ICML 2013*, Atlanta, GA, USA, 1-6.
Lee, M. Y., Sung, K. W., and Han, S. W. (2021), Transfer Learning with Seasonal Adjustment for Automotive Spare Part Long-Term Demand Forecasting, *Journal of the Korean Institute of Industrial Engineers*, **47**(3), 302-314.
Leeuw, J. D., Bukhsh, Z.A., and Zhang, Y. (2023), Parcel loss prediction in last-mile delivery: deep and non-deep approaches with insights from Explainable AI, ArXiv, abs/2310.16602.
Lei, D., Qi, Y., Liu, S., Geng, D., Zhang, J., Hu, H., and Shen, Z.-J. M. (2025), Pooling and Boosting for Demand Prediction in Retail: A Transfer Learning Approach, *Manufacturing & Service Operations Management*, **27**(6), 1779-1794.
Li, Z., Zheng, Y., Chen, C., and Jbabdi, S. (2024), Learning Label Refinement and Threshold Adjustment for Imbalanced Semi-Supervised Learning. ArXiv, abs/2407.05370.
Ministry of Land, Infrastructure and Transport (2023), Registered Motorcycles by Metropolitan City and Province [Dataset], Total Registered Motor Vehicles - Motorcycle Registration Status by Region, MOLIT Statistics Service, Sejong, Korea.
Moniz, N., Branco, P., and Torgo, L. (2017), Resampling Strategies for Imbalanced Time Series Forecasting, *International Journal of Data Science and Analytics*, **3**(3), 161-181.
National Data Office (2024), Population Census: Households by Age of Household Head and Household Size (Ordinary Households) [Dataset], KOSIS - Korean Statistical Information Service, Daejeon, Korea.
Prati, R. C., Batista, G. E. A. P. A., and Monard, M. C. (2004). Class imbalances versus class overlapping: An analysis of a learning system behavior. In *Mexican International Conference on Artificial Intelligence*, 312-321.
Son, J., An, W., Im, J., and Cho, Y. (2022), A Case Study on the Demand Forecasting of the Navy Repair Parts Using Machine Learning, *Journal of the Korean Institute of Industrial Engineers*, **48**(3), 320-326.
Srinivasan, A. (1993), Statlog (Landsat Satellite) [Dataset], UCI Machine Learning Repository, University of California, Irvine, CA, USA.
Vuttipittayamongkol, P. and Elyan, E. (2021), On the class overlap problem in imbalanced data classification, *Knowledge-Based Systems*, **212**, 106631.
Wang, L., Mykityshyn, A.L., Johnson, C., and Marple, B. D. (2020), Deep Learning for Flight Demand Forecasting.
Wang, S., Chen, A., Wang, P., and Zhuge, C. (2023), Short-term electric vehicle battery swapping demand prediction: Deep learning methods, *Transportation Research Part D: Transport and Environment*, **119**, 103746.
Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., Schiele, B., & Xie, X. (2022). FreeMatch: Self-adaptive thresholding for semi-supervised learning.
Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020), Self-Training with Noisy Student Improves ImageNet Classification, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), IEEE, Seattle, WA, USA, 10687-10698.
Yang, W., Zhang, R., Chen, J., Wang, L., and Kim, J. (2023), Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Volume 1: Long Papers, Association for Computational Linguistics, Toronto, Canada, 16369-16382.
Yang, Y., Khorshidi, H. A., and Aickelin, U. (2024), A Review on Over-Sampling Techniques in Classification of Multi-Class Imbalanced Datasets: Insights for Medical Problems, Frontiers in Digital Health, 6, 1430245.
Zhang, Y., Cheng, Y., Huang, X., Wen, F., Feng, R., Li, Y., and Guo, Y. (2021). Simple and Robust Loss Design for Multi-Label Learning with Missing Labels. ArXiv, abs/2112.07368.

저자소개

유선호: 고려대학교 산업경영공학과 학사과정에 재학 중이다. 연구분야는 기계학습, Structured Prediction, 그래프 이론이다.

주승돈: 한국항공대학교에서 항공우주공학 학사, Texas A&M University에서 항공우주공학 석사, Georgia Institute of Technology에서 경영학 석사학위를 취득하였다. 현재 주식회사 젠트로피의 CEO로 재직 중이며, 전기 이륜차와 배터리 교환 시스템을 서비스하고 있다.

고광종: 한국공학대학교에서 IT 경영학과와 컴퓨터공학 학사학위를 취득하였다. 현재 고려대학교 산업경영공학과 석·박사통합과정 재학중이며, 연구분야는 매타휴리스틱과 기계학습을 이용한 제조시스템 최적화이다.

정태수: 고려대학교에서 산업공학 학사, 한국과학기술원에서 산업공학 석사, Georgia Institute of Technology에서 산업시스템공학 박사 학위를 취득하였다. 현재 고려대학교 산업경영공학과의 정교수로 재직 중이며, 연구분야는 AI와 최적화(OR)를 이용한 산업계 의사결정 및 최적화이다.