ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 52, No. 7, pp. 593-600, 2025. 7
https://doi.org/10.5626/JOK.2025.52.7.593

조기 예측을 위한 시계열 데이터 불균형 해소 기법

(Time Series Data Imbalance Resolution Techniques for Early Prediction)

안 응 선 † 권 태 형 † 김 도 국 ††
(Eungseon An) (Taehyoung Kwon) (Doguk Kim)

요 약 시계열 예측은 관측된 시계열 데이터를 분석하여 미래의 값을 예측하는 중요한 문제다. 그러나, 데이터가 불균형할 경우, 모델의 성능이 저하되고 예측 결과에 편향이 발생할 수 있다. 이를 해결하기 위해 최근 다양한 딥러닝 기법과 데이터 증강 방법들이 연구되고 있지만, 많은 연구들이 불균형 문제와 시계열 특성을 동시에 고려하지 못하여 근본적인 문제를 해결하지 못하고 있다. 본 연구에서는 시간적 패턴을 활용하여 샘플을 생성하는 조기 예측을 위한 방법을 제안한다. 제안된 기법은 긍정 및 부정 클래스를 효과적으로 구분할 수 있는 시점을 선정하여, 더 먼 시차에 대한 예측도 가능하게 한다. 본 연구에서 제안된 방법은 기존의 방법들보다 우수한 성능을 보였으며, 더 멀리 있는 시차에 대한 조기 예측의 가능성을 입증하였다.

키워드: 불균형 데이터, 시계열 데이터, 시간적 패턴, 조기 예측, 데이터 증강

Abstract Time series forecasting is a critical task that involves analyzing observed time series data to predict future values. However, when dealing with imbalanced data, model performance can degrade, leading to biased predictions. Although recent studies have explored various deep learning techniques and data augmentation methods, many fail to address challenges posed by data imbalance and the intrinsic characteristics of time series data simultaneously, leaving underlying issues unresolved. This study proposed a novel approach that could leverage temporal patterns to generate synthetic samples and extend the scope of early prediction. By identifying key moments that could effectively distinguish between positive and negative classes, our method enhanced the ability to predict further into the future. The method proposed in this study demonstrated superior performance to existing methods and proved the feasibility of early prediction for longer time lags.

Keywords: imbalanced data, time series data, temporal patterns, early prediction, data augmentation

· 본 연구는 인하대학교의 지원, 과학기술정보통신부 및 정보통신기획평가원의 생성AI선도인재양성사업 (IITP-2025-RS-2024-00360227)의 지원, 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(NO.RS-2022-00155915, 인공지능융합혁신인재양성(인하대학교))을 받아 수행된 연구임

† 비 회 원 : 인하대학교 전기컴퓨터공학과 학생
two01272@inha.ac.kr
dogsa333@gmail.com
†† 종신회원 : 인하대학교 전기컴퓨터공학과 교수(Inha Univ.)
dgkim@inha.ac.kr
(Corresponding author임)

논문접수 : 2024년 12월 5일
(Received 5 December 2024)
논문수정 : 2025년 5월 7일
(Revised 7 May 2025)
심사완료 : 2025년 5월 9일
(Accepted 9 May 2025)

정보과학회논문지 제52권 제7호(2025. 7)

1. 서 론

시계열 데이터는 시간의 흐름에 따라 관측된 데이터를 연속적으로 배열한 형태로, 금융, 의료, 제조업 등 다양한 산업에서 중요한 역할을 한다. 이러한 데이터는 과거 변화를 기반으로 미래의 트렌드를 예측하거나, 특정 사건의 발생 가능성을 평가하는 데 사용된다. 시계열 예측은 과거 데이터를 활용해 미래 값을 예측하는 작업으로, 그 정확성을 높이기 위한 다양한 연구가 지속적으로 이루어지고 있다. 그러나 시계열 데이터는 종종 불균형한 분포를 보이는데, 이는 예측 모델의 성능 저하를 유발하는 주요 요인 중 하나다. 불균형 데이터에서는 특정 클래스의 샘플 수가 매우 적거나 과도하게 많은 경우가 많아, 예측 결과가 왜곡될 위험이 있다. 예를 들어 제조업에서의 기계 고장 예측이나 의료 분야에서의 질병 조기 진단과 같은 중요한 응용에서는, 적은 데이터로 중요한 결정을 내려야 하는 만큼 소수 클래스의 예측이 특히 중요하다[1]. 이런 문제를 해결하기 위해 다양한 불균형 문제 해결 기법들이 제안되었다.

시계열 분류는 시계열 데이터를 주어진 카테고리에 맞게 분류하는 작업으로, 예측과 더불어 중요한 시계열 분석 방법 중 하나다. 예를 들어, 금융에서는 주식 시장의 변동성을 예측하고, 의료 분야에서는 환자의 생체 신호를 분석하여 질병의 조기 진단을 가능하게 하며, 제조업에서는 기계 고장의 조기 감지를 통해 유지보수 비용을 절감할 수 있다. 이러한 산업들에서는 예측의 정확성이 비즈니스 성과와 직결되며, 특히 시계열 데이터에서의 조기 예측은 사전 대응을 가능하게 하여 큰 이점을 제공한다. 불균형한 시계열 데이터의 분류 작업은 매우 까다로울 수 있으며, 잘못된 분류는 심각한 결과를 초래할 수 있다. 이러한 문제를 해결하기 위해, 기존의 데이터 증강 기법과 새로운 시계열 데이터 특화 기법들이 개발되고 있다.

불균형 데이터는 머신러닝 모델의 학습 과정에서 다수 클래스에 편향된 예측을 초래하며[2], 특히 소수 클래스에 대한 예측 정확도가 크게 떨어질 수 있다. 이는 금융 사기 탐지에서의 소수 사기 거래 탐지, 의료 진단에서의 드문 질병 조기 진단, 그리고 제조업에서의 드문 기계 고장 감지와 같은 중요한 응용 분야에서 매우 치명적인 결과를 초래할 수 있다. 이러한 문제를 해결하기 위해 다양한 데이터 증강[3] 및 불균형 해소 기법들이 제안되어 왔다.

일반적으로 사용되는 기법 중 하나는 SMOTE (Synthetic Minority Over-sampling Technique)[4]로, 소수 클래스의 샘플 간의 선형 보간을 통해 새로운 샘플을 생성함으로써 불균형을 해결하려고 한다. 하지만 SMOTE는 시계열 데이터의 시간적 구조를 고려하지 않아, 복잡한 패턴을 효과적으로 반영하지 못하는 한계를 가지고 있다. 또한 ADASYN(Adaptive Synthetic Sampling Approach)[5] 같은 방법도 존재하지만, 이 역시 각 샘플의 복잡성을 충분히 반영하지 못하며, 과적합의 문제를 야기할 수 있다.

이에 반해, 본 연구에서 제안한 기법은 기존의 이러한 한계들을 보완하는 데 중점을 두고 있다. 특히, T-SMOTE (Temporal Synthetic Minority Over-sampling Technique)[6]를 개선하여, 시계열 데이터의 특성을 충분히 반영할 수 있는 새로운 방법을 제시한다. 기존의 T-SMOTE는 소수 클래스 샘플들 간의 선형 보간을 통해 새로운 샘플을 생성하는 방식으로, 시계열 데이터의 시간적 패턴을 충분히 반영하지 못한다는 한계가 있다. 이에 따라 본 연구에서는 Dynamic Time Warping(DTW)[7] 기법을 도입하여, 시계열 데이터의 비선형적인 시간적 패턴을 더 효과적으로 반영할 수 있는 새로운 샘플 생성 방법을 제안한다. 또한, 기존 T-SMOTE에서 사용된 시간적 이웃 정의가 고정된 임계값에 의해 제한됨으로써 발생하는 조기 예측의 한계를 극복하기 위해, 새로운 임계값 설정 방법을 도입했다. 이 방법은 시계열 데이터의 시간적 패턴을 보다 정확하게 반영하며, 더 먼 시차를 고려하여 조기 예측의 범위를 확장하는 데 기여한다. 구체적으로, 높은 예측값을 가진 샘플들뿐만 아니라, 적절히 설정된 낮은 임계값 이하의 샘플들도 고려하여 조기 예측의 신뢰도를 높였다. 이러한 개선된 방법론은 다양한 시계열 데이터셋에서 기존의 T-SMOTE 및 다른 최신 기법들과 비교했을 때 우수한 성능을 보였다. 특히, 본 연구의 기법은 더 넓은 시차에 걸쳐 높은 예측 성능을 유지하며, 조기 예측의 가능성을 크게 확장시켰다. 이를 통해, 시계열 데이터에서의 불균형 문제를 효과적으로 해결하였다.

본 연구는 기존 기법들이 가진 한계를 넘어서는 새로운 접근 방식을 제안함으로써, 이를 통해 조기 예측의 범위를 넓히고, 시계열 데이터의 복잡한 패턴을 더 잘 반영함으로써, 다양한 응용 분야에서의 실질적인 성과를 기대할 수 있다.

2. 관련 연구

2.1 데이터 불균형 문제 해결 기법

데이터 불균형 문제는 머신러닝 모델의 성능 저하를 초래하는 주요 요인 중 하나이다. 특히 소수 클래스에 대한 예측 정확도가 떨어질 수 있다는 문제점을 가진다. 이러한 문제를 해결하기 위해 다양한 전통적인 기법들이 제안되어 왔다. 우선, Repeat 기법은 소수 클래스의 샘플을 단순히 반복하여 다수 클래스와의 불균형을 줄이는 가장 기본적인 방법이다. 이 방법은 구현이 쉽고

빠르지만, 샘플의 다양성이 부족하여 과적합의 위험이 크다. 또한, SMOTE는 소수 클래스 샘플들 간의 선형 보간을 통해 새로운 샘플을 생성하여 불균형을 해결하는 기법이다. SMOTE는 샘플의 다양성을 높여 과적합을 방지할 수 있지만, 생성된 샘플이 기존 데이터의 경계를 제대로 반영하지 못할 수 있는 단점이 있다.

이를 개선한 B-SMOTE(Borderline-SMOTE)[8]는 SMOTE의 진화된 버전으로, 기존 SMOTE가 모든 소수 클래스 샘플을 대상으로 샘플을 생성하는 반면, B-SMOTE는 소수 클래스의 경계 근처 샘플에 중점을 두어 새로운 샘플을 생성함으로써, 데이터의 경계를 더 잘 반영하는 샘플을 생성할 수 있다. 그러나 이 기법 역시 데이터의 복잡한 시간적 패턴을 반영하는 데 한계가 있다. 이러한 한계를 극복하기 위해 MBS(Modified Borderline-SMOTE)[9], INOS(Improved Neighborhood Oversampling)[10], MBO(Minority Based Oversampling)[11] 등의 기법이 제안되었다. MBS는 경계에서 멀리 떨어진 안전한 영역에서도 샘플을 생성하여 과적합 문제를 완화하고, INOS는 데이터 공간의 밀도 분포를 고려하여 밀집된 지역에서 더 많은 샘플을 생성한다. MBO는 소수 클래스 샘플이 데이터 공간 전반에 고르게 분포되도록 새로운 샘플을 생성한다. 이러한 진보된 방법론들은 전통적인 기법들이 가지는 한계를 일부 극복하였으나, 시계열 데이터와 같은 고유한 시간적 특성을 가진 데이터에 적용될 때는 여전히 그 한계를 완전히 극복하지 못하는 경우가 많다. 이 때문에, 시계열 데이터의 특성을 반영한 특화된 기법들이 요구된다.

2.2 시계열 데이터의 불균형 문제와 기존 접근 방법

시계열 데이터는 시간에 따라 순차적으로 발생하는 데이터로, 각 데이터 포인트 간의 시간적 상관관계가 매우 중요한 특징이다[12]. 이러한 시간적 특성은 데이터의 패턴과 경향을 이해하는 데 중요한 역할을 하며, 특히 예측 모델의 성능에 큰 영향을 미친다. 그러나 불균형한 시계열 데이터에서는 소수 클래스의 데이터가 부족하여 모델이 해당 클래스를 제대로 학습하지 못하는 문제가 발생한다. 이를 해결하기 위해 여러 데이터 불균형 처리 기법들이 제안되었지만, 대부분의 기법들은 비시계열 데이터를 대상으로 개발되었기 때문에 시계열 데이터의 고유한 시간적 특성을 충분히 반영하지 못한다는 한계가 있다. 이러한 한계를 극복하기 위해 T-SMOTE 방법이 제안되었다. T-SMOTE는 시계열 데이터의 시간적 순서를 고려하여 시간적 이웃 간의 선형 보간을 통해 새로운 합성 샘플을 생성하는 기법이다. 이를 통해 시계열 데이터의 시간적 특성을 반영하면서 불균형 문제를 완화할 수 있다. 또한, Spy Method를 활용하여 이상치를 제거하고, 모델의 예측 성능을 향상시키는 전략을 사용한다. 그러나 기존의 T-SMOTE 방법은 시계열 데이터의 비선형적인 패턴을 충분히 반영하지 못하며, Spy Method에서 설정되는 임계값 h로 인해 조기 예측 범위가 제한되는 문제가 있다. 따라서 시계열 데이터의 복잡한 패턴을 효과적으로 반영하고, 조기 예측의 범위를 확장할 수 있는 개선된 기법이 필요하다.

Algorithm 1 Spy method for Obtaining l values

Input:

Negative samples N, Positive samples P, Spy sample percentage s

Output:

Thresholds h_{high}, h_{low}

Set of indices L corresponding to specific prediction intervals

1: P_{spy} = Randomly select s%of N

2: $P' = P \cup P_{spy}$ & $N' = N - P_{spy}$

3: Re-balance $P'$to match class ratios by repeating positive samples

4: Train an LSTM model using P' and N'

5: Obtain prediction scores $\{s_i\}$ from the trained model

6: Set thresholds:

$$h = \begin{cases} h_{high} = \max_i s_i \\ h_{low} = Percentile_p(s_i), where\ p \in [95, 99] \end{cases}$$

7: Identify index sets:

$$L = \begin{cases} L_{high} = i | s_i > high \\ L_{low} = i | h_{low} < s_i \le h_{high} \end{cases}$$

그림 1 제안 기법에서의 Spy Method 임계값 설정 방식
Fig. 1 Threshold Setting Method of the Spy Method in the Proposed Technique

2.3 기존 접근 방법의 한계와 Spy Method의 문제점

Spy method[13]는 주로 positive-unlabeled learning에서 사용되는 방법으로, 여기서는 부정적인 데이터를 일부 긍정적인 것으로 재분류한 후에 학습을 진행하여 최종적으로 부정 샘플을 효과적으로 분류할 수 있도록 도와주는 기법이다. 제안 기법에서의 Spy method 임계값 설정 방식은 위 (그림 1)에 나타나있다. 구체적으로, 원래 부정 샘플의 일부(약 5~15%)를 spy 샘플로 설정하여 긍정 샘플에 혼합하고, 이 새로운 데이터를 이용해 학습을 진행한다. 이는 데이터 불균형 정도에 따라 유동적으로 적용하기 위함이다. 불균형 정도가 극심해 양성 샘플이 극히 적은 경우에는 Spy 비율을 낮게(약 5%) 설정해 과도한 오라벨링을 방지하고, 반대로 불균형이 상대적으로 심하지 않은 경우에는 최대 15%까지 높여도 양성 클래스 판단에 무리가 없도록 했다. Spy method는 이 과정에서 부정 샘플을 긍정으로 라벨링하고, 예측 점수가 가장 높은 임계값 h를 기준으로 긍정 클래스와 부정 클래스를 구분한다. 이 h값은 학습된 모델이 얼마나 정확하게 부정 샘플을 감지할 수 있는지를 평가하는 데 사용된다. 예를 들어, 긍정 클래스로 라벨링 된 spy 샘플이 임계값 h 이상일 경우, 긍정 클래스일 가능성이 매우 높다. 이 방식은 부정 샘플을 필터링하는 데 유용하

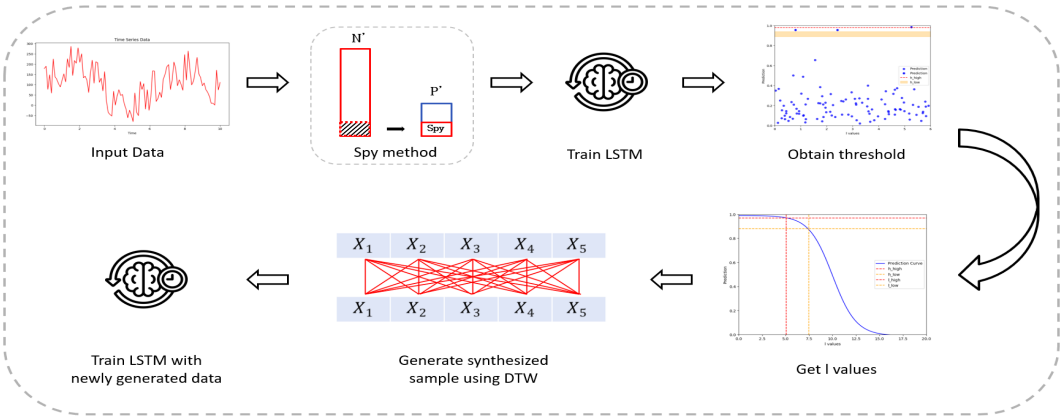

그림 2 제안된 T-SMOTE+의 전반적인 흐름
Fig. 2 Overall Workflow of the Proposed T-smote+ Method

며, 예측 성능을 높이는 데 기여할 수 있다. 기존 Spy method의 한계는 임계값 h가 지나치게 높게 설정될 경우, 결과적으로 얻어지는 l값이 작거나 없을 수도 있다. 이는 조기 예측의 범위를 제한하게 된다. 예를 들어, 시계열 데이터에서 spy 샘플을 이용해 학습한 결과 h값이 너무 높게 설정되면, 시차 l을 이용한 추가 샘플링이 어려워지고, 조기 예측의 의미가 퇴색될 수 있다. 이러한 문제점을 해결하기 위해 본 연구에서는 Spy Method의 임계값 설정 방식을 개선하고, 시계열 데이터의 특성을 반영한 새로운 샘플 생성 방법을 제안한다. 제안된 T-SMOTE+의 전반적인 흐름은 (그림 2)에 나타나 있다.

3. 제안 방법

3.1 제안 알고리즘 개요

- 조기 예측의 범위가 협소했던 한계를 극복하고자, 본 연구에서는 더 넓은 시차와 더 많은 시계열 정보를 활용하는 새로운 접근 방식을 제안한다. h_{low}와 h_{high}의 임계값 설정을 통해, 고위험 샘플과 저위험 샘플을 구분하며, 이를 바탕으로 조기 예측의 성능을 향상시켰다.
- SPY method에서 발생할 수 있는 잠재적인 이상치 문제를 개선하기 위해, 본 연구에서는 이웃 샘플들의 유사성을 더욱 정밀하게 평가하고, 이상치로 인한 예측 오류를 최소화하였다. 이를 통해 기존 기법에서 놓칠 수 있었던 샘플 간의 미세한 차이를 반영하고, 전체 모델의 예측 정확도 및 안정성을 강화하였다. 특히, 예측 성능의 변동성을 줄여 강인한 성능을 확보하였다.
- DTW기법을 적용하여 비선형적인 시계열 패턴까지 반영한 샘플링 방법을 제안한다. 이를 통해 시간에 따른 비선형적 변화와 복잡한 패턴을 더 잘 반영할 수 있으며, 특히 장기적인 시차를 고려한 예측의 정확성을 향상시켰다.

3.2 Spy method의 개선

기존의 Spy Method는 가장 높은 예측값을 기준으로 임계점 h를 설정하기 때문에, 이 값이 높게 고정됨에 따라 더 먼 시차의 데이터를 충분히 반영하지 못하는 문제가 발생한다. 단일한 h 값을 설정하면 특정 데이터셋에서 샘플이 너무 적게 선정되거나, 조기 예측에 필요한 더 멀리 있는 시차의 정보를 담지 못할 위험이 있다. 이를 해결하기 위해, 우리는 임계값 h를 단일 값이 아닌 두 개의 값 h_{high}와 h_{low}로 설정하는 방식을 제안한다. 아래에 있는 (그림 3)은 시차 l 값에 따른 예측 점수의 변화를 나타내며, 임계값 설정의 필요성을 시각적으로 보여준다.

구체적으로, h_{high}는 가장 높은 예측값으로 설정하고, h_{low}는 예측값 분포의 95%에서 99% 사이의 percentile 값으로 설정한다. 이를 통해 데이터셋의 분포에 따라 유

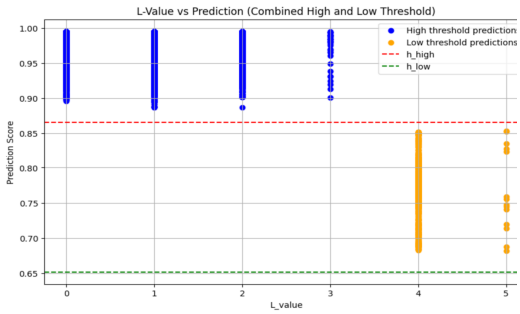

그림 3 값에 따른 예측 점수 그래프
Fig. 3 Prediction Score Graph by l Value

연하게 h_{low}값을 조절할 수 있다. 이러한 설정은 긍정 클래스의 예측값을 기반으로 샘플을 세 그룹으로 나누는데 활용된다.

High 그룹: $s_i > h_{high}$, 긍정일 가능성이 매우 높다.

Low 그룹 : $h_{low} < s_i < h_{high}$, 긍정일 가능성이 높지만 상대적으로 낮다.

Outlier 그룹 : $s_i < h_{low}$, 긍정일 가능성이 낮은 이상치로 간주된다.

각 그룹에 해당하는 가중치는 다음과 같이 결정한다[14].

High 그룹 : $w_i = \max(0, s_i - h_{high})$

Low 그룹 : $w_i = \max(0, s_i - h_{low})$

High 그룹의 샘플들은 전체 생성되는 샘플의 90%에서 95%를 차지하도록 한다. 이는 이 그룹의 샘플들이 긍정일 가능성이 매우 높기 때문에, 모델의 학습에 더 큰 비중을 두어야 하기 때문이다. Low 그룹의 샘플들은 전체 생성되는 샘플의 5%에서 10%만을 차지하도록 한다. 이는 조기 예측의 범위를 넓히면서도 모델의 안정성을 유지하기 위한 전략이다. 특히, l값의 범위가 크게 늘어나지 않은 경우에는 후보 샘플이 적으므로 최대 10%까지 허용하고, 반면에 l 범위가 크게 확장된 경우에는 후보 샘플이 충분하기에 5% 정도로 제한한다. 이는 시차가 크게 확장된 데이터셋일수록 긍정 샘플을 좀 더 유연하게 늘리고, 반대로 시차 확장이 미미하거나 불균형이 극심해 양성 정보가 희소한 경우에는 오분류 위험을 줄이기 위해 Low 그룹 비율을 최소화하려는 의도이다. 아울러 Outlier 그룹을 배제하는 방식은, 잘못된 예측으로 인한 손실 비용이 큰 산업 데이터 등에서 특히 유리하다. 부정확한 긍정 샘플을 모델에 섞으면 치명적 오판이 발생하기 쉽기 때문에, 일정 임계값 아래의 이상치는 과감히 제거함으로써 안전성을 높일 수 있다. 본 연구에서는 세 개의 대표 시계열 데이터로 실험했으며, 비용 민감도가 높은 실제 산업 현장에서도 이러한 Outlier 적용이 효과적일 것으로 기대한다.

3.3 샘플 생성 방법

기존의 T-SMOTE 방법은 시간적 이웃 간의 선형 보간을 통해 샘플을 생성하지만, 이는 비선형적인 시계열 패턴을 충분히 반영하지 못한다는 한계가 있다. 이를 해결하기 위해, 본 연구에서는 DTW기법을 도입하여 비선형적인 시계열 패턴을 반영한 새로운 샘플 생성 방법을 제안한다. DTW는 두 시계열 간의 유사성을 계산할 때 시간 축을 비선형적으로 변형하여 가장 유사한 패턴을 찾는다. 이를 통해 시계열 데이터의 시간적 변형이나 속도 차이를 보정할 수 있어, 패턴의 유사성을 보다 정확하게 반영할 수 있다. (그림 4)는 기존 방식과 DTW방식을 비교하여, 각 기법이 시계열 패턴을 어떻게 반영하는지를 시각적으로 보여준다. DTW의 이론적 계산 복잡도는 시계열 길이를 L이라 할 때 $O(L^2)$로 알려져 있다. 그러나 본 연구의 접근에서는 모든 시차(l) 구간에 대해 DTW를 전면적으로 적용하지 않고, 일정 임계값(Outlier 배제)을 만족하는 소수의 구간에만 비선형 합성을 수행한다. 따라서 실제 대규모 데이터셋에서도 L이 극도로 커지지 않는 한 계산 부담이 과도하지 않으며, 다양한 시계열 환경에서 본 방법을 적용하는 데 큰 문제는 없다.

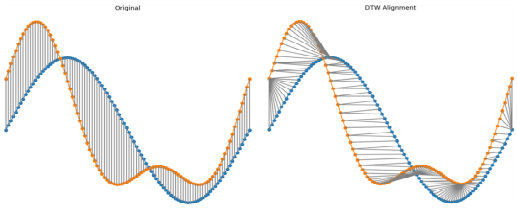

그림 4 기존 기법과 DTW를 적용한 기법 비교

Fig. 4 Comparison of Original Method & Dynamic Time Wraping method

본 연구에서는 DTW를 활용하여 양성 샘플 간의 비선형적인 시간 정렬을 수행하고, 이를 기반으로 새로운 합성 샘플을 생성한다.

$$X_{new} = \alpha * X_{il}^{aligned} + (1-\alpha) * X_{il+1}^{aligned} \quad (1)$$

샘플 X_{il}과 시간적 이웃 샘플 X_{il+1} 사이에 DTW 알고리즘을 적용하여 최적의 정렬 경로를 계산하고 두 샘플의 비선형적인 시간적 관계를 정렬하고, 정렬된 시계열을 기반으로 임의의 계수 α를 사용하여 합성샘플 X_{new}를 얻는다. 여기서 α는 $Beta(s_{il}, s_{il+1})$에서 추출한 값이며, s_{il}과s_{il+1}는 각각 샘플 X_{il}과 X_{il+1}의 예측 점수이다. 합성 샘플의 예측점수 s_{new}는 다음과 같다.

$$s_{new} = \alpha * s_{il} (1-\alpha) * s_{il+1} \quad (2)$$

각 샘플 X_{il}에 대해 생성할 합성 샘플의 수 m_{il}은 다음과 같이 결정한다.

$$m_{il} = \frac{r * n * s_{il}}{\sum_{i=1}^{n} \sum_{l=1}^{L} s_{il}} \quad (3)$$

r : imbalance ratio

n : the number of true positive sample

$\sum_{i=1}^{n} \sum_{l=1}^{L} s_{il}$: sum of prediction scores of all positive samples

표 1 다양한 불균형 데이터 처리 기법별 성능
Table 1 Results of Different Imbalanced Data-Handling Methods

Dataset	AUC							
	Repeat	SMOTE	B-SMOTE	MBS	INOS	MBO	T-SMOTE	**T-SMOTE+**
Wafer	0.9629	0.9782	0.9760	0.9793	0.9890	0.9888	**0.9992**	0.9987
TwoPats	0.8731	0.8833	0.8877	0.8967	0.9077	0.9096	0.9130	**0.9998**
Sleaf	0.9295	0.9584	0.9720	0.9729	0.9767	0.9823	**0.9987**	0.9981
Dataset	F1							
	Repeat	SMOTE	B-SMOTE	MBS	INOS	MBO	T-SMOTE	**T-SMOTE+**
Wafer	0.9349	0.9456	0.9497	0.9553	0.9603	0.9641	**0.9756**	**0.9756**
TwoPats	0.6451	0.6521	0.6531	0.6603	0.6716	0.6705	0.6826	**0.9989**
Sleaf	0.8624	0.9000	0.9107	0.9207	0.9320	0.9288	0.9484	**0.9574**
Dataset	AUPRC							
	Repeat	SMOTE	B-SMOTE	MBS	INOS	MBO	T-SMOTE	**T-SMOTE+**
Wafer	0.9318	0.9461	0.9507	0.9548	0.9669	0.9680	**0.9885**	0.9873
TwoPats	0.6796	0.7016	0.7048	0.7136	0.7310	0.7264	0.7487	**0.9996**
Sleaf	0.9440	0.9535	0.9532	0.9632	0.9703	0.9712	**0.9886**	0.9842

4. 실 험

4.1 데이터 셋

본 연구에서는 제안된 방법의 성능을 검증하기 위해 UCR(Time Series Classification Archive)에 공개된 세 가지 시계열 데이터셋을 사용하였다. 해당 데이터셋들은 시계열 데이터 분석 분야에서 널리 활용되며, 다양한 특성과 난이도를 지니고 있어 알고리즘의 일반화 성능을 평가하는 데 적합하다.

Wafer: 반도체 제조 공정에서 웨이퍼의 이상 상태를 감지하기 위한 데이터로, 정상 상태와 이상 상태의 두 가지 클래스로 구성된다. 총 2,174개의 시계열 샘플이 있으며, 각 샘플은 152개의 Timestamp를 가진다. 클래스 불균형이 존재하여 정상 샘플이 다수를 차지한다.

TwoPatterns: 인공적으로 생성된 데이터셋으로, 두 가지 패턴을 식별하는 이진 분류 문제를 다룬다. 총 1,000개의 샘플로 구성되어 있으며, 각 샘플은 128개의 Timestamp을 가진다. 클래스 간의 비율이 균등하지 않아 불균형 문제를 평가하는 데 활용된다.

Sleaf: 스웨덴산 나뭇잎의 외곽선을 기반으로 한 시계열 데이터로, 15종의 서로 다른 나무 종을 분류하는 다중 분류 문제이다. 총 1,125개의 샘플이 있으며, 각 샘플은 128개의 Timestamp로 이루어져 있다. 본 연구에서는 불균형 문제를 강조하기 위해 특정 클래스를 대상으로 이진 분류로 변환하여 사용하였다.

4.2 실험 결과

4.2.1 실험 환경

본 연구는 Python 3.9.19와 PyTorch 2.4를 사용하여 실험을 수행하였으며, GPU로는 NVIDIA GeForce RTX 3090을 활용하였다. 모델은 LSTM을 기반으로 설계되었으며[15], 불균형 데이터를 처리하기 위해 양성 클래스에 가중치를 부여한 이진 크로스엔트로피 손실 함수를 사용하였다. 옵티마이저로는 Adam을 사용하였고, 조기 종료 기법을 적용하여 과적합을 방지하였다. 시계열 데이터의 조기 예측을 위해, 데이터셋의 시계열 길이를 고려하여 최대 l값을 20으로 설정하였다. 이는 기존의 T-SMOTE 논문에서 사용된 설정을 참고한 것으로, 조기 예측의 범위를 효과적으로 설정하기 위함이다 이에 따라 각 데이터셋의 타임스탬프에서 20을 뺀 값을 모델의 입력 시퀀스 길이로 사용하였다. 즉, 각 샘플의 시계열 데이터 중 초기부터 마지막 20개의 Timestamp 이전까지의 데이터로 모델이 예측을 수행하도록 설계되었다. 이를 통해 조기 예측의 범위를 확장하여 더 이른 시점에서의 예측이 가능하도록 하였다. 학습 과정으로 생성된 새로운 긍정 샘플들은 기존에 있던 부정 샘플과 균형을 맞추어 학습 데이터셋을 구성하였다. 이는 클래스 불균형 문제를 완화하고 모델의 예측 성능을 향상시키기 위함이다. 이렇게 구성된 균형 잡힌 데이터셋을 사용하여 LSTM 모델을 재학습하였다.

4.2.2 성능 평가

성능 평가는 F1 스코어, AUC(Area Under the ROC Curve), AUPRC(Area Under the Precision-Recall Curve)를 기준으로 수행되었다. (표 1)은 각 데이터셋에 대해 기존의 여러 불균형 데이터 처리 기법인 Repeat, SMOTE, B-SMOTE, MBS, INOS, MBO 그리고 T-SMOTE와 본 연구에서 제안한 방법의 성능을 비교하였다.

4.2.3 비교 및 분석

제안된 방법은 조기 예측의 범위를 확장함으로써 실제 응용에서의 활용도를 높였다. 조기 예측 범위를 넓혔기 때문에 일부 메트릭의 값이 다소 하락하는 것은 예상되는 결과로, 이는 모델이 더 이른 시점의 데이터로부

터 예측을 수행해야 하기 때문이다. 그럼에도 불구하고 제안된 방법은 안정적이고 미미한 성능 저하만을 보였으며, 이는 모델의 강인성과 일반화 능력을 입증한다. 특히 TwoPatterns 데이터셋에서는, 기존 T-SMOTE 대비 F1이 약 31%p 향상되었고, AUC와 AUPRC 또한 각각 약 8%p와 25%p가량 증가하여 비선형 패턴을 더욱 효과적으로 학습했음을 보여주었다. 반면 Wafer와 SwedishLeaf 데이터셋에서는 일부 메트릭 값이 약 0.01 내외로 소폭 낮아졌으나, 전체적인 예측 성능은 여전히 우수하였고 표준편차 역시 안정적으로 유지되었다. 이는 기존 방법이 조기 예측 범위가 협소하여 실용성이 제한적이었던 부분을 보완한 것으로, 제안된 방법의 중요한 장점 중 하나이다.

본 연구에서 제안한 방법은 기존 기법과 비교하여 복잡한 시계열 데이터의 불균형 문제를 효과적으로 해결할 수 있음을 보였다. 특히, DTW를 활용한 샘플링 기법은 비선형적인 시간적 변화를 반영하여 모델의 예측 능력을 향상시켰다. 또한, Spy Method의 개선을 통해 이상치로 인한 성능 저하를 최소화하였으며, 조기 예측의 범위를 확장하여 실제 적용 가능성을 높였다. 한편, 일부 데이터셋에서 성능 향상이 미미하거나 기존 기법과 유사한 결과를 보인 것은 데이터셋의 특성이나 모델의 한계로 인한 것일 수 있다.

5. 결 론

본 논문에서는 시계열 데이터의 불균형 문제를 해결하고 조기 예측의 성능을 향상시키기 위한 새로운 방법을 제안하였다. 제안된 방법은 DTW를 활용하여 비선형적인 시계열 패턴까지 반영함으로써, 복잡한 시간적 변화를 더욱 정확하게 포착할 수 있었다. 이를 통해 소수 클래스의 데이터 증강 효과를 극대화하여 장기적인 시차를 고려한 예측의 정확성을 높였다. 또한, 기존 Spy Method에서 발생할 수 있는 이상치 문제를 개선하기 위해 이웃 샘플들의 유사성을 정밀하게 평가하였으며, 이를 통해 예측 오류를 최소화하고 모델의 안정성을 강화하였다. 특히, 예측 성능의 변동성을 줄여 다양한 데이터셋에서도 일관된 성능을 보였다. 실험 결과, Wafer와 Sleaf 데이터셋에서는 기존 방법과 비슷한 수준의 성능을 유지하면서도 성능의 안정성을 확보할 수 있었다. 특히 TwoPatterns 데이터셋에서는 기존 방법에 비해 탁월한 성능 향상을 보였으며, 이는 제안된 방법이 복잡한 패턴을 가진 데이터셋에서도 우수한 성능을 발휘함을 나타낸다. 이러한 결과는 제안된 방법이 조기 예측의 범위를 확장하면서도 안정적인 메트릭을 제공하여 실제 응용에서의 적용 가능성을 높인다는 것을 시사한다.

향후 연구에서는 추가적인 데이터셋과 다양한 실험을 통해 제안된 방법의 일반성을 더욱 검증할 예정이다. 특히 Spy Method가 단순 배제 방식 위주로 동작하는 현 체계를 한층 정교화하여, 일부 Outlier 가능성이 있는 샘플도 비용 민감도나 시점별 특성에 따라 유연하게 재활용하는 방안을 모색할 계획이다. 이를 통해 긍정 클래스로 잘못 분류되었을 수도 있는 샘플들에 대한 활용도를 높이고, 더욱 안정적인 예측이 가능하도록 개선하고자 한다. 본 연구에서 활용한 T-SMOTE 계열 증강 방식은, 시점별 정보를 단계적으로 전달, 기억하는 LSTM과 결합했을 때 탁월한 시너지 효과를 보였다. LSTM은 게이트 구조를 통해 중요한 과거 정보를 누적, 제어할 수 있어, 근접 시차(1값이 작은 구간)에서 생성된 샘플을 안정적으로 학습하며 조기 예측 성능을 끌어올린다. 한편, 최근 각광받는 Transformer나 TCN은 병렬처리 능력과 전역 패턴 학습에 강점을 갖추고 있어, 장기 시점이나 대규모 시계열 데이터에서 유리할 것으로 기대된다. 따라서 본 논문에서는 우선 LSTM 기반 접근에 집중하였으나, 향후 연구에서는 Transformer, TCN 등 최신 모델과 T-SMOTE를 융합하여 근접 시차 구간은 LSTM의 세밀한 국소 패턴 학습에 맡기고, 더 긴 범위나 전역 의존성이 중요한 구간에는 Transformer의 어텐션 메커니즘을 활용함으로써 시계열 전 구간에 대한 예측 성능을 한층 높일 수 있을 것으로 보인다. 이러한 확장을 통해 본 논문에서 제안한 증강 기법의 산업적 적용 범위를 더욱 넓히고, 다양한 시계열 환경에서의 강건성과 재현성을 강화할 수 있을 것으로 기대된다. 본 연구의 결과는 시계열 데이터 분석 분야에서 불균형 문제와 조기 예측의 한계를 극복하는 데 기여하며, 다양한 응용 분야에서 더욱 신뢰도 높은 예측을 제공하는 토대가 될 것이다.

References

[1] Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). “Deep learning for time series classification: A review,” *Data Mining and Knowledge Discovery*, Vol. 33, No. 4, pp. 917–963.

[2] S. Yi, H. Choi, S. Kwon, J. Han, and K. Im, "Rethinking class imbalance in deep learning: A Bayesian approach," *IEEE Access,* Vol. 9, pp. 5538–5549, 2021.

[3] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time Series Data Augmentation for Deep Learning: A Survey," arXiv preprint, Vol. 2002, No. 12478v4, Mar. 2022.

[4] J. Jiang, R. Song, H. Fang, and F. Tse,"An improved SMOTE imbalanced data classification method based on duple differential evolution algorithm," *Applied Intelligence,* Vol. 51, No. 4, pp.

2429 - 2440, 2021.
[5] T. S. Shie, J. C. Tsai, and T. Y. Kuo, "Improved ADASYN for enhancing classification performance of imbalanced datasets," *IEEE Access,* Vol. 8, pp. 72695 - 72707, 2020.
[6] P. Zhao, C. Luo, B. Qiao, L. Wang, S. Rajmohan, Q. Lin, and D. Zhang, "T-SMOTE: Temporal-Oriented Synthetic Minority Oversampling Technique for Imbalanced Time Series Classification," *Proc. of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22),* pp. 2406-2412, 2022.
[7] K. Roy and M. Sharma, "An improved dynamic time warping-based approach for time series classification," *Journal of Intelligent & Fuzzy Systems,* Vol. 40, No. 5, pp. 9203 - 9214, 2021.
[8] B. Liu, J. Guo, and H. Song, "Temporal Borderline-SMOTE: A robust oversampling strategy for imbalanced time series classification," *IEEE Access*, Vol. 9, pp. 173210 - 173220, 2021.
[9] C.-L. Liu, and P.-Y. Hsieh, "Model-Based Synthetic Sampling for Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering,* Vol. 32, No. 8, pp. 1543-1555, Aug. 2020.
[10] X. Chen, J. Liu, and Y. Zhang, "An enhanced integrated oversampling framework for imbalanced time series classification," *IEEE Access,* Vol. 8, pp. 96300 - 96312, 2020.
[11] Z. Gong, and H. Chen, "Model-Based Oversampling for Imbalanced Sequence Classification," *Proc. of the 2016 ACM Conference on Information and Knowledge Management (CIKM 2016)*, pp. 1009-1018, Oct. 2016.
[12] A. Bagnall, J. Lines, J. Hills, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, Vol. 31, No. 3, pp. 606 - 660, 2017.
[13] X. Li and B. Liu, "PU learning in the wild: A unified framework and a comparative study," *Knowledge-Based Systems,* Vol. 239, 107875, 2022.
[14] G. Pang, C. Shen, and L. Cao, "Deep Learning for Anomaly Detection: A Review," *IEEE Transactions on Knowledge and Data Engineering, early access,* pp. 1 - 20, 2021.
[15] J. Tang, C. Xie, and X. Li, "Addressing imbalanced time series classification via data augmentation and LSTM neural networks," *Neurocomputing,* Vol. 414, pp. 159 - 170, 2020.

안 응 선
2024년 인하대학교 통계학과 졸업(학사) 2024년~현재 인하대학교 전기컴퓨터공학과 석사과정. 관심분야는 시계열 데이터, 데이터 증강, 머신러닝

권 태 형
2024년 인하대학교 컴퓨터공학과 졸업(학사). 2024년~현재 인하대학교 전기컴퓨터공학과 석사과정. 관심분야는 시계열 데이터, 데이터 증강, 딥러닝

김 도 국
2012년 한국과학기술원 컴퓨터공학과 졸업(학사). 2014년 한국과학기술원 정보보호대학원 졸업 (석사). 2018년 한국과학기술원 전산학부 졸업 (박사). 2018년~2020년 하나금융융합기술원 책임. 2020년~2021년 카카오엔터프라이즈 책임. 2021년~현재 인하대학교 인공지능공학과 조교수. 관심분야는 시계열 예측, 자연어 처리, 금융 인공지능