https://doi.org/10.7840/kics.2022.47.10.1637

거리 기반 레이블링 기법을 적용한 마이크로RNA-유전자 연관성 예측 딥러닝 모델 연구

김 재 인◆, 윤 승 원*, 황 인 우*, 이 규 철°

A Study on a Deep Learning Model for Predicting MicroRNA-Gene Association Using Distance-Based Labeling Methods

Jaein Kim◆, Seung-Won Yoon*, In-Woo Hwang*, Kyu-Chul Lee°

요 약

마이크로RNA는 유전자 발현을 조절하는 중요한 RNA이다. 마이크로RNA와 유전자 사이의 연관성을 찾는 것은 매우 작아서 실험이 어려운 마이크로RNA가 어떻게 작동하는지 이해하는 데에 도움이 되며 환자의 질병을 진단하는 데에도 효과적이다. 본 논문은 마이크로RNA-유전자 쌍의 거리 측정을 기반으로 하는 새로운 레이블링 방법과 함께 마이크로RNA와 유전자의 연관성을 예측하는 LSTM(Long-Short Term Memory) 기반 딥러닝 모델을 제시하며 그에 대한 5-fold 교차 검증 실험 결과를 제시한다. 본 연구에서 제시한 딥러닝 모델의 ROC 커브는 최대 0.95를 달성하였다.

키워드 : 마이크로RNA, 유전자, 마이크로RNA-유전자 연관성, 레이블링, 거리
Key Words : microRNA, gene, miRNA-gene associations, labeling, distance

ABSTRACT

MicroRNAs (miRNAs) are important RNAs that regulate gene expression. Finding associations between microRNAs and genes can help us understand how microRNAs that are so small and difficult to experiment, and are also effective in diagnosing a patient's disease. In this paper, we present a long-short term memory (LSTM) based deep learning model that predicts the association of microRNAs and genes together with a new labeling method based on distance measurement of miRNA-gene pairs. The verification test results are presented. The ROC curve of the deep learning model presented in this study achieved a maximum of 0.95.

Ⅰ. 서 론

마이크로RNA(miRNA)는 약 22개의 뉴클레오타이드로 구성된 작은 비암호화 RNA 분자이다. 마이크로RNA는 인체에서 유전자가 단백질로 발현되는 정도를 조절하는 역할을 수행한다. 즉, 질병을 일으킬 수 있는 단백질이 생성되는 것을 막을 수 있다. 마이크로RNA는 최근 기본적이고 중요한 여러 가지 생물학적

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021-1283-02).
◆ First Author : Chungnam National University, jaeinnn21@gmail.com, 학생회원
° Corresponding Author : Chungnam National University, kclee@cnu.ac.kr, 정회원
* Chungnam National University, yoonenoch11@gmail.com, 학생회원; allhpy35@gmail.com, 학생회원
논문번호 : 202205-075-C-RE, Received April 29, 2022; Revised July 4, 2022; Accepted August 16, 2022

현상들과 더불어 수많은 인간의 질병 원인과 밀접한 관련이 있는 것으로 밝혀짐에 따라 생물학 및 의학 분야에서 활발하게 연구되고 있다[1]. 마이크로RNA와 유전자의 연관성(association)을 예측하는 연구는 마이크로RNA의 역할과 그것의 매우 복잡한 메커니즘을 이해하는 데에 도움을 줄 수 있으며, 마이크로RNA 관련 질병이나 약물, 그리고 질병 관련 유전자를 예측하는 연구에 큰 도움을 줄 수 있다. 그러나 흔히 'wet lab experiment'라고 불리는 직접적인 실험으로 마이크로RNA와 유전자의 연관성을 찾는 것은 매우 오랜 시간이 소요되는 문제가 있다. 이로 인해 기계 학습, 딥러닝 모델과 같은 계산적(computational) 방법을 통해서 연관성을 예측할 필요성이 제기되었으며 최근 이와 같은 계산적 방법을 이용한 연관성 예측 연구가 다양하게 진행되고 있다[2].

어떤 바이오 이기종 데이터(heterogeneous data) 쌍의 요소들이 서로 관련이 있으면 그 데이터 쌍은 양성(positive), 그렇지 않으면 음성(negative)이라고 가정하자. 이 경우 일반적으로 마이크로RNA-유전자 등 바이오 연관성 데이터에는 마이크로RNA와 유전자가 서로 관련 있음이 'wet lab experiment'를 통해 검증된 양성 샘플은 존재하는 반면 음성 샘플은 존재하지 않는데, 이는 바이오 데이터 쌍은 서로 절대적으로 관련이 없다고 확실하게 정의될 수는 없기 때문이다. 검증되지 않은 쌍은 서로 연관성이 없음을 의미하는 것이 아니라 서로의 연관성이 아직 밝혀지지 않았을 뿐이다. 그런데 이진 분류를 위한 딥러닝 모델 성능이 좋기 위해서는 데이터셋의 양성 및 음성 샘플이 같은 비율로 존재해야 하므로 본 연구에서는 음성 샘플을 어떻게 구축할지가 중요하다. 본 논문에서는 더욱 정교한 데이터 구축을 위해 기존에 존재하고 있는 데이터셋에서 음성 데이터를 새롭게 레이블링하였다. 본 논문에서는 이를 위한 유클리드(Euclidean) 거리와 마할라노비스(Mahalanobis) 거리 기반 레이블링 기법과 기존 마이크로RNA-유전자 연관성 예측 연구 중 최고 성능의 모델을 능가하는 새로운 딥러닝 모델을 제시한다.

본 논문의 2장에서는 관련 연구 및 마이크로RNA와 유전자 연관성의 레이블링 방법과 연관성을 예측하는 딥러닝 모델의 설명을 제시하며 3장에서는 2장에서 설명한 모델에 대한 실험 결과를 제시한다. 마지막으로 이어지는 4장에서는 결론과 향후 연구 방향을 제시한다.

Ⅱ. 본 론

2.1 관련 연구

바이오 데이터 간의 연관성을 계산적인 방법을 통해 예측하는 연구는 최근 많아지고 있다. MLMDA[3]는 마이크로RNA와 질병 간의 연관성을 오토인코더와 랜덤 포레스트를 이용하여 예측하는 기계 학습 모델이다. 먼저 k-mer sparse matrix를 이용하여 마이크로RNA의 시퀀스 정보를 추출하고, 마이크로RNA의 기능적 유사성(functional similarity), 질병의 의미적 유사성(semantic similarity) 및 가우시안 정보 프로필 커널 유사성(GIP, Gaussian interaction profile kernel similarity) 정보와 결합한다. 이후 심층 오토인코더 뉴럴 네트워크를 통해 이들로부터 더 대표적인 특징을 추출한다. 마지막으로 랜덤 포레스트 분류기를 사용하여 잠재적 마이크로RNA-질병 연관성을 예측한다. Li[4]는 토폴로지 기반 유사성 측정을 통해 마이크로RNA-질병 연관성 예측 방법을 제시하였다. 해당 연구에서는 마이크로RNA와 질병의 연관성 네트워크 내에서 유사도를 계산하기 위해 딥러닝 기법인 DeepWalk를 활용한다. 또한 SMALF[5]는 스택 오토인코더를 이용하여 마이크로RNA-질병 연관성 매트릭스에서 마이크로RNA와 질병의 잠재적 특징(latent feature)을 학습한다. 그 후 마이크로RNA의 기능적 유사성, 마이크로RNA의 잠재적 특징, 질병의 의미적 유사성 그리고 질병의 잠재적 특징을 통합하여 특징 벡터를 얻는다. 마지막으로 XGBoost 알고리즘을 통해 알려지지 않은 마이크로RNA와 질병 사이의 연관성을 예측하는 모델이다. Tang[6]은 잠재적인 마이크로RNA와 질병의 관계를 예측하기 위하여 다중관점 딥러닝 프레임워크인 MMGCN을 제안하였다. 마이크로RNA와 질병의 다중 소스 유사성 관점을 융합하여 마이크로RNA-질병 예측 문제를 권장 작업으로 프레임화한다. MMGCN은 총 세 가지 모듈로 구성된 엔드 투 엔드(end-to-end) 모델이다. 각 모듈은 마이크로RNA와 질병의 멀티 소스 데이터를 인코딩하기 위한 다중 관점 GCN 인코더, GCN 인코더로 학습한 마이크로RNA 및 질병에 대한 서로 다른 채널 정보를 별도로 학습하는 멀티 채널 주의 메커니즘, 마지막으로 멀티 채널 주의 모듈의 특징을 재조합하여 마이크로RNA와 질병의 통합 임베딩을 얻는 CNN 결합기이다. CNN 결합기로 얻는 임베딩은 최종적으로 마이크로RNA와 질병 연관성 예측 작업의 입력으로 이용된다. GAEMDA[7]는 그래프 오토인코더 모델로, 엔드 투 엔드 방식으로 잠재적인 마이크로RNA와 질병의 관

계를 식별한다. GAEMDA는 그래프 신경망 기반 인코더를 적용하여 마이크로RNA 및 질병 노드의 저차원 임베딩을 생성하고 이기종 정보를 효과적으로 융합하기 위하여 노드의 이웃 정보를 집계하는 집계 기능 및 다층 퍼셉트론을 포함한다. 이후 이중 선형 디코더에 마이크로RNA와 질병 노드의 임베딩이 공급되어 마이크로RNA와 질병 노드 간의 잠재적 연결을 식별한다. 이처럼 바이오 데이터는 기계 학습이나 딥러닝 모델을 적용한 방법으로 연관성을 예측하는 연구들이 실제로 많이 행해지고 있으며 본 연구에서는 딥러닝 모델을 이용하여 마이크로RNA-유전자 연관성을 예측한다.

2.2 데이터 구축

본 논문에서는 SG-LSTM-FRAME[8]의 데이터셋에서 마이크로RNA와 유전자 각각의 시퀀스(sequence) 및 지형적(geometric) 특징을 통합하여 만들어진 128차원의 벡터값으로 표현된 마이크로RNA와 유전자 특징을 이용하여 연관성을 예측한다. 바이오 데이터를 벡터 형태로 표현하는 것은 연관성을 예측하는 데에 있어서 매우 중요하다. 벡터값으로 특징이 추출된다는 것은 곧 벡터 간의 거리를 측정할 수 있기 때문에 연관성 예측에 아주 좋은 지표가 된다. 또한 마이크로RNA와 유전자, 질병 등의 연관성을 딥러닝 모델을 통해 예측하기 위해서는 딥러닝 모델에 적합한 데이터를 구축하는 것이 중요하다. 일반적으로 딥러닝 모델에서의 입력은 1(양성)과 0(음성)으로 각각 레이블링된 샘플이 서로 동등한 비율로 존재하는 데이터셋을 필요로 한다. 그러나 명백히 관련 없음이 검증된 생물학적 데이터는 별도로 존재하지 않으므로 음성 샘플은 임의로 마이크로RNA와 유전자의 쌍을 만들어 구축하여야 한다. 이때 발생할 수 있는 문제점 중 하나는, 검증된 양성 샘플의 마이크로RNA-유전자 쌍의 거리만큼 임의로 생성된 특정 음성 샘플의 마이크로RNA-유전자 쌍의 거리가 가까울 수도 있다는 것이다. 별도의 기준을 가지고 음성 샘플을 만드는 것이 아니기 때문이다. 여기서 거리가 가깝다는 것은 마이크로RNA와 유전자 간의 연관성이 높다는 것을 의미한다. 이를 해결하고자 별도로 유클리드 거리와 마할라노비스 거리를 기반으로 한 레이블링을 진행함으로써 더욱 정교한 음성 샘플을 구축하고자 한다.

본 논문에서 이용하는 마이크로RNA와 유전자 연관성 데이터셋은 Xie[8]의 core 데이터셋이다. 해당 데이터셋의 각 샘플에는 시퀀스 특징 및 지형적 특징이 '그림 1'과 같이 128차원의 벡터값으로 통합되어 마이크로RNA와 유전자에 각각 존재한다. 데이터셋의 샘플들은 마이크로RNA와 유전자가 서로 연관이 있다고 검증된 15,540개의 레이블 1을 가지는 양성 샘플들과 연관성이 검증되지 않은 레이블 0으로 지정된 음성 샘플 15,540개로 총 31,080개의 샘플로 이루어져 있다.

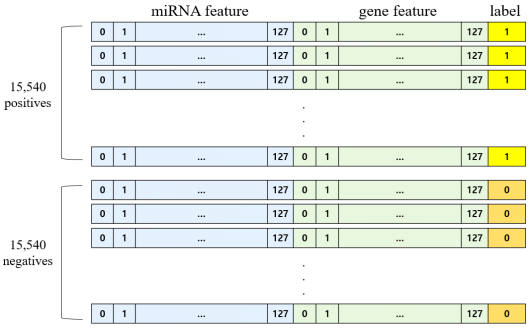

그림 1. 본 연구에서 사용한 데이터셋의 초기 구조도
Fig. 1. Initial structure of the dataset used in this study

주로 마이크로RNA와 유전자 또는 이외 생물학적 데이터들 간의 연관성을 예측하기 위해서는 유클리드 거리를 이용하는 경우가 많다. 본 연구에서는 음성 샘플 구축에 유클리드 거리뿐만 아닌 마할라노비스 거리까지 추가로 도입하였다. 마할라노비스 거리는 유클리드 거리에 공분산 계산을 추가한 형태로 이해할 수 있으며, 이로 인해 마할라노비스 거리는 데이터가 분포한 형태를 고려하여 거리를 측정할 수 있다. 이는 즉 데이터의 상관관계를 고려하여 거리를 측정할 수가 있다는 것이다. 수식 (1)은 마할라노비스 거리 공식으로, u와 v는 각각 마이크로RNA와 유전자의 배열이며 V는 마이크로RNA와 유전자 사이의 공분산 행렬로 이의 역행렬이 거리 계산에 사용된다.

$$\sqrt{(u-v)\,V^{-1}(u-v)^{T}} \tag{1}$$

음성 샘플의 레이블링 기준값(threshold)은 양성 샘플들의 거리값 평균으로 하였다. 먼저 바이오 데이터들 간의 거리를 측정할 때 주로 쓰이는 유클리드 거리만을 고려하는 방식으로, 레이블링 기준값보다 더 가까운 거리를 가지는 음성 샘플을 제거하였다. 이 작업은 양성으로 의심되는 음성 샘플을 제거하여 더욱 명백한 음성 샘플만을 남기기 위함이다. '그림 2'에서는 유클리드 거리로 측정한 레이블링 기준값보다 거리가 가까운 음성 샘플들의 거리값을 초록색 점으로 표시하였으며 x축이 음성 샘플들의 거리값을 나타낸다. 이때 레이블링 기준값은 빨간색 선으로 표시한 3.24이

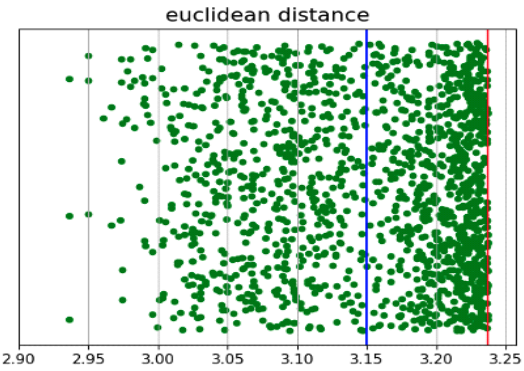

그림 2. 유클리드 거리로 측정한 기준값보다 더 짧은 값을 갖는 음성 샘플의 분포도
Fig. 2. Distribution of negative samples with values shorter than the threshold measured by Euclidean distance

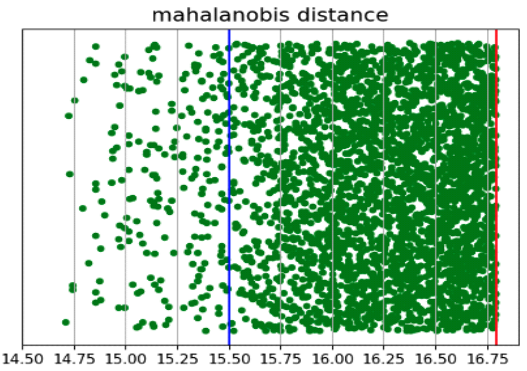

그림 3. 마할라노비스 거리로 측정한 기준값보다 더 짧은 값을 갖는 음성 샘플의 분포도
Fig. 3. Distribution of negative samples with values shorter than the reference value measured by Mahalanobis distance

다. 유클리드 거리 3.24보다 가까운 거리값을 가지는 음성 샘플은 총 1,465개로, 해당 1,465개의 음성 샘플을 제거한 뒤, 이진 분류 모델을 위한 데이터는 양성 샘플과 음성 샘플이 동일한 비율로 존재하여야 하므로 무작위로 1,465개의 양성 샘플을 제거해 주었다.

위에서 사용한 방식은 유클리드 거리만을 레이블링에 이용한다. 더욱 정교한 레이블링을 위해 추가로 거리 기준을 도입할 필요가 있다. 이로 인해 데이터의 분포를 고려하여 거리를 측정하는 마할라노비스 거리를 추가로 레이블링에 이용한다. 기준값보다 거리가 짧은 음성 샘플은 유클리드 거리 기준에서 해당하는 것을 먼저 삭제한 뒤 마할라노비스 거리 기준에 해당하는 것을 추가로 삭제하고 기준값보다 거리가 먼 음성 샘플은 유지하는 방법으로 레이블링 작업을 진행하였다. 이는 즉 양성으로 의심되는 샘플을 제거하여 명백히 음성으로 판단되는 샘플만 유지하도록 하는 방법이다. 이러한 방식에서 본래 유클리드 거리에서의 기준값은 약 3.24, 마할라노비스 거리에서의 기준값은 약 16.79이다. 이때 더욱 명백하게 양성에 가까운 연관성을 지니는 샘플만을 제거하기 위해 데이터 분포를 고려하여 기준값을 조정하였다. 두 가지 거리값을 기준으로 할 때 음성 샘플 분포가 적은 지점인 유클리드 거리에서의 3.15와 마할라노비스 거리에서의 15.5를 새로운 기준값으로 하였다. '그림 2'과 '그림 3'의 두 그래프에서 파란색 세로선은 새롭게 조정한 레이블링 기준값이다.

본 연구에서 조정한 유클리드 거리 기준 3.15, 마할라노비스 거리 기준 15.5의 새 기준값을 기준으로 삭제할 음성 샘플은 유클리드, 마할라노비스 거리 기준에서 각각 358개, 200개로 중복을 제외한 총 365개의 음성 샘플을 제거하였다. 샘플을 제거하면 양성, 음성 샘플의 비율이 달라지므로 이전의 방식과 마찬가지로 무작위로 365개의 양성 샘플 역시 제거하여 데이터를 최종적으로 구축하였다.

2.3 딥러닝 모델 설계

'그림 4'는 직전 문단에서 설명한 레이블링 방법을 통해 딥러닝 모델을 사용하여 최종적으로 연관성을 예측하는 과정을 나타낸 것이다. (a)는 유클리드 거리를 통해 계산된 358개의 삭제되어야 할 음성 샘플들

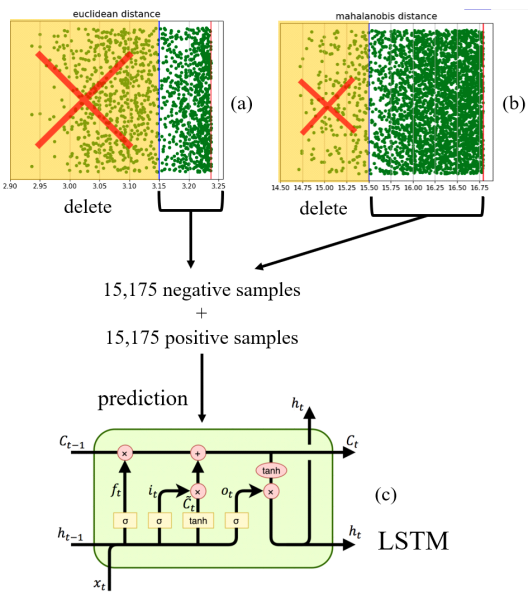

그림 4. 새로운 레이블링 방법을 적용한 딥러닝 모델의 개념도
Fig. 4. Overview of a deep learning model applying a new labeling method

이며 (b)는 마할라노비스 거리를 통해 계산된 200개의 삭제되어야 할 음성 샘플들이다. 이들의 합집합인 365개의 음성 샘플과 365개의 양성 샘플들을 삭제하여 구축한 새로운 데이터를 딥러닝 모델을 통해 예측한다. 본 연구에서는 입력 데이터가 시계열 특성을 가지므로 RNN 계열의 딥러닝 모델인 LSTM(Long-Short Term Memory) 모델을 구축하여 연관성을 예측하였다. 전통적인 RNN은 비교적 짧은 시퀀스에 대해서만 효과를 보이며, RNN의 time step이 길어질수록 앞의 정보가 뒤로 잘 전달되지 못하는 장기 의존성 문제가 발생한다. LSTM은 이러한 문제를 해결할 수 있는 모델로, 은닉층의 메모리 셀에 입력 게이트, 망각 게이트, 출력 게이트를 추가하여 불필요한 기억을 지우는 과정을 거친다. 따라서 LSTM은 RNN에 비해서 긴 시퀀스의 입력을 처리하는 데 탁월한 성능을 보인다.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

수식 (2)는 망각 게이트에 대한 수식으로, h_{t-1}과 x_t를 받아서 0에서 1 사이의 값을 C_{t-1}에 보낸다. 이때 수식에서 결과값이 1이면 보존, 0이면 버린다. 수식 (3)과 (4)는 입력 게이트에 관한 수식으로, 수식 (2)에서 어떤 것을 기록하고 버릴지 정한 반면 여기서는 기록하는 것들 중 어떤 값을 업데이트하는지 정하게 된다. 수식 (5)에서는 과거의 셀 상태를 업데이트한다. 이 때 곱셈을 이용하여 망각할 것은 0으로 만들고 기록해야 할 것은 덧셈을 통해 기록한다. 수식 (6)과 (7)은 출력 게이트에 대한 수식으로 여기서는 무엇을 출력으로 내보낼지 결정하게 된다. 셀 상태에서 전체를 내보내는 것이 아니라 특정 부분만을 출력으로 내보내도록 한다.

본 연구에서 설계한 모델의 LSTM 레이어는 3으로 지정하였고, 손실 함수는 Cross-Entropy, 최적화 함수는 Adam 최적화 함수를 사용하였다. 딥러닝 모델의

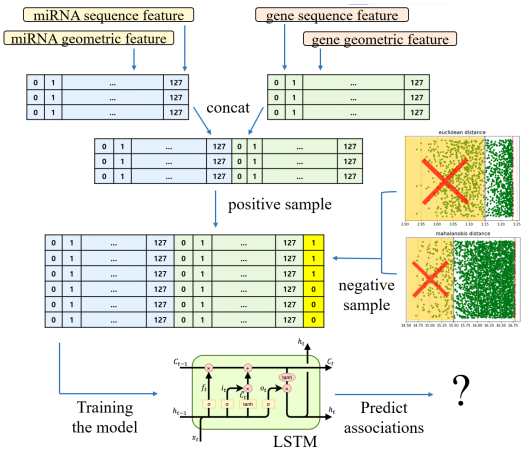

그림 5. 본 연구 모델의 전체 흐름도
Fig. 5. The overall workflow of model of this study

배치(batch) 사이즈는 64, 학습률(learning rate)은 0.001로 하였다. 모든 실험은 Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz, 32GB RAM, and GeForce GTX 1080 Ti GPU 환경에서 진행하였다.

본 연구의 특징 추출 및 딥러닝 모델 예측까지의 흐름도는 그림 5와 같다.

2.4 실험 데이터

실험은 총 네 가지 실험 데이터들을 본 연구에서 설계한 모델에 적용하여 진행하였다. 첫 번째 실험 데이터에서는 SG-LSTM-FRAME[8]에서 사용한 데이터의 양성 샘플의 마이크로RNA와 유전자의 리스트를 가지고 무작위로 짝을 지어 15,540개의 음성 샘플을 생성하였다. 해당 샘플과 기존의 양성 샘플을 합친 31,080개의 샘플로 이루어진 데이터가 첫 번째 실험 데이터이다. 두 번째 실험 데이터는 유클리드 거리 기준값인 3.24보다 가까운 음성 샘플 쌍을 제거한 후 양성 샘플 또한 제거된 음성 샘플 개수만큼 무작위로 추가로 제거하여 구축한 데이터이다. 세 번째 실험 데이터는 2.2절의 데이터 구축에서 조정되기 이전의 기준값인 유클리드 거리 기준값 3.24와 마할라노비스 거리 기준값 16.79를 기준으로 하여 음성 샘플을 삭제하고, 이에 맞게 양성 샘플 또한 삭제한 총 29,564개의 데이터이다. 마지막으로 네 번째 실험 데이터는 새로운 두 개의 기준값을 기준으로 가까운 거리값을 가지는 음성 샘플을 제거하고 마찬가지로 양성 샘플도 무작위로 제거하여 구축된 데이터이다.

2.5 실험 결과

평가 지표로는 AUC-ROC 커브를 이용하였다.

AUC-ROC 커브는 다양한 임계값에서 모델의 분류 성능을 측정한 그래프이다. AUC는 0과 1 사이의 값을 가지며 AUC 값이 높다는 것은 클래스를 구별하는 모델 성능이 뛰어나다는 것을 의미한다. 실제 임상에서 AUC-ROC 커브는 일반인 및 환자 클래스를 구분하는 모델의 성능 평가 지표로 흔히 사용된다. 또한 모든 실험에서는 5-fold 교차 검증을 적용하였다.

첫 번째 실험 데이터로 진행한 실험에서는 검증 결과 최고 AUC 0.70 정도의 성능을 나타내었다. 이는 즉, 양성 클래스와 음성 클래스를 구별할 확률이 70% 임을 의미하는데, 이와 같이 샘플을 별도의 레이블링 기준을 정하지 않거나 별도의 조치 없이 무작위로 선정하는 것은 마이크로RNA와 유전자의 연관성 예측에는 적합한 방식이 아니라고 할 수 있다. 따라서 연관성 예측을 위한 데이터는 앞의 '본론'에서 설명되었던 레이블링 기법을 적용하여 정교하게 구축되어야 한다.

다음으로 두 번째 실험 데이터의 학습 및 검증 결과는 '그림 6'과 같이 최고 AUC 0.9450의 성능을 나타내었다. 무작위로 음성 샘플을 선정하는 방식에 비해서 눈에 띄게 성능이 상승함을 알 수 있다. 또한 기존 마이크로RNA-유전자 연관성 예측에서 최고 성능을 나타내었던 ROC 커브 0.94의 성능을 낸 Xie[8]의 연구와 비교하였을 때 소폭의 성능 상승이 있는 것을 알 수 있다.

조정되기 이전의 기준값을 적용한 세 번째 실험 데이터를 이용한 모델의 5-fold 교차 검증의 결과는 최고 AUC 0.9481을 달성하였다. 또한 새로 조정된 기준값과 새로운 레이블링 방법을 적용한 네 번째 실험 데이터를 이용한 모델의 5-fold 교차 검증으로 학습 및 테스트한 결과는 최고 AUC 0.9522를 달성하였다. 이는 기존 최고 성능을 나타내었던 논문의 0.94의 AUC에 대해서 1%의 개선점이 존재한다. 또한 세 번째 실험과 네 번째 실험을 비교하였을 때 기준 거리값을 조금 더 짧게, 즉 더욱 정교하게 적용하였을 때의 성능 수치가 조금 더 좋았음을 확인할 수 있다.

'표 1'에서 볼 수 있듯, 별도의 레이블링 작업 없이 무작위로 선정한 음성 샘플의 첫 번째 데이터는 분류 성능이 뛰어나지 않다. 유클리드 거리를 이용하여 기초적인 레이블링을 진행한 두 번째 실험 데이터의 경우 첫 번째 데이터에 비해서 눈에 띄는 성능 개선이 있었다. 그러나 여기에서 추가로 마할라노비스 거리를 적용하여 더욱 정교한 레이블링을 진행한 세 번째 데이터의 경우 기존 가장 최고의 성능을 보인 논문에 비해서 1% 이상의 성능 개선이 나타났다. 따라서 본 논문의 레이블링 기법을 적용하면 딥러닝 분류 모델의 성능 개선에 도움을 줄 수 있다.

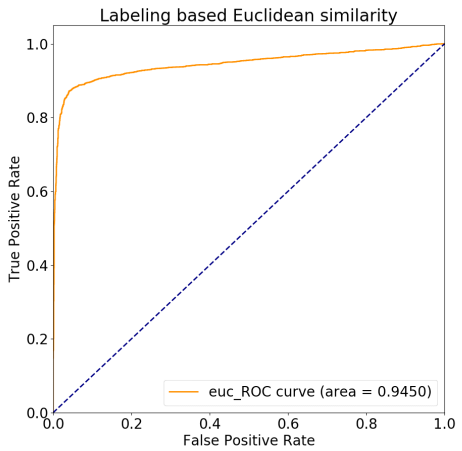

그림 6. 유클리드 거리 기준값보다 가까운 값을 삭제한 데이터로 학습한 결과
Fig. 6. A result of learning with data with values closer than the Euclidean distance reference value deleted

표 1. 세 가지 실험 데이터를 이용한 5-fold 교차 검증 결과 AUC 값
Table 1. fold cross-validation result AUC value using three experimental data

data 1	data 2	data 3	data 4
0.70	0.9450	0.9481	**0.9522**

Ⅲ. 결 론

본 논문에서는 마이크로RNA와 유전자의 연관성을 예측하는 딥러닝 모델의 성능을 향상시킬 수 있는 레이블링 방법과 함께 LSTM을 이용한 딥러닝 모델의 개발 및 실험 결과를 제시하였다. 본 논문과 같은 마이크로RNA와 기타 바이오 데이터의 연관성을 예측하는 딥러닝 모델의 연구는 최근 의학, 생물학, 약학 등의 분야에서 질환의 진단이나 신약 개발을 위해 많이 이루어지고 있다. 본 연구에서 제시한 레이블링 방법은 마이크로RNA와 유전자 사이의 연관성뿐만 아니라 수많은 질병과 관련 있는 유전자 등을 예측할 때도 쓰일 수 있을 것이다.

본 연구에서는 유클리드 거리와 마할라노비스 거리를 이용한 정교한 데이터 구축을 위한 레이블링을 제시하였으나, 연구에서 이용한 데이터는 양성, 음성 포함 약 3만 개 정도로 그 양이 적다는 한계가 있다. 이를 해결하기 위해 향후 본 연구팀만의 데이터를 새롭

게 구축하여 많은 양의 데이터를 확보하여 추가적인 실험을 진행할 계획이다.

References

[1] S.-J. Kim and H.-J. Cho, “A review on the correlation between the pathology of alzheimer’s disease and microRNA,” *Biomed. Sci. Lett.*, vol. 27, no. 4, pp. 208-215, 2021. (https://doi.org/10.15616/BSL.2021.27.4.208)

[2] Z.-H. You, et al., “PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction,” *PLoS Computational Biology*, vol. 13, no. 3, 2017. e1005455. (https://doi.org/10.1371/journal.pcbi.1005455)

[3] K. Zheng, et al., “MLMDA: A machine learning approach to predict and validate MicroRNA-disease associations by integrating of heterogenous information sources,” *J. Translational Med.*, vol. 17, no. 1, pp. 1-14, 2019. (https://doi.org/10.1186/s12967-019-2009-x)

[4] G. Li, et al., “Predicting microrna-disease associations using network topological similarity based on deepwalk,” *IEEE Access*, vol. 5, pp. 24032-24039, 2017. (https://doi.org/10.1109/ACCESS.2017.2766758)

[5] D. Liu, et al., “SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost,” *BMC Bioinformatics*, vol. 22, no. 1, pp. 1-18, 2021. (https://doi.org/10.1186/s12859-021-04135-2)

[6] X. Tang, et al., “Multi-view multichannel attention graph convolutional network for miRNA-disease association prediction,” *Briefings in Bioinformatics*, vol. 22, no. 6, 2021. (https://doi.org/10.1093/bib/bbab174)

[7] Z. Li, et al., “A graph auto-encoder model for miRNA-disease associations prediction,” *Briefings in Bioinformatics*, vol. 22, no. 4, 2021. (https://doi.org/10.1093/bib/bbaa240)

[8] W. Xie, et al., “SG-LSTM-FRAME: A computational frame using sequence and geometrical information via LSTM to predict miRNA-gene associations,” *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 2032-2042, 2021. (https://doi.org/10.1093/bib/bbaa022)

김 재 인 (Jaein Kim)

2021년 2월 : 충남대학교 컴퓨터공학과 졸업

2021년 3월~현재 : 충남대학교 바이오AI융합학과 석사과정

<관심분야> 인공지능, 딥러닝, 머신러닝, 생물정보학

[ORCID:0000-0002-2902-6146]

윤 승 원 (Seung-Won Yoon)

2018년 2월 : 충남대학교 컴퓨터공학과 졸업

2018년 3월~현재 : 충남대학교 컴퓨터공학과 석박사통합과정

<관심분야> 인공지능, 딥러닝, 머신러닝, 생물정보학

황 인 우 (In-Woo Hwang)

2021년 2월 : 목원대학교 정보통신융합공학부 졸업

2021년 3월~현재 : 충남대학교 바이오AI융합학과 석사과정

<관심분야> 딥러닝, 바이오AI, 생물정보학, feature extraction

이 규 철 (Kyu-Chul Lee)

1984년 2월 : 서울대학교 컴퓨터 공학과 졸업
1986년 2월 : 서울대학교 컴퓨터 공학과 석사
1990년 2월 : 서울대학교 컴퓨터 공학과 박사
1989년 3월~현재 : 충남대학교 컴퓨터공학과 교수

<관심분야> 데이터베이스, 인공지능, 빅데이터