준지도학습 기반 LDL-콜레스테롤 예측의 정확도 개선

양수빈[1], 김민태[1], 권수빈[1], 우나현[1], 김학재[2], 정태경[3], 이성주[1†]

[1]상명대학교 소프트웨어학과
[2] ㈜클래스액트
[3] 한림대학교 인공지능융합학부
{201921007[1], 201820985[1], 202020999[1], 202021020[1]}@sangmyung.kr
peacfeel[1†]@smu.ac.kr
krunvis@gmail.com
ttjeong@hallym.ac.kr

Improving the prediction accuracy for LDL-cholesterol based on semi-supervised learning

Su-Bhin Yang[1], Min-Tae Kim[1], Su-Bin Kwon[1], Na-Hyun Woo[1],
Hak-Jae Kim[2], Tai-Kyeong Jeong[3], Sung-Ju Lee[1†]
[1]Dept. of Software, Sang-myung University
[2]ClssAct Incorporated
[3]Dept. of Artificial Intelligence Convergence, Hal-lym University

요 약

이상지질혈증의 발병에 대한 조기 진단 및 관리하는 것은 중요한 문제이다. 이상지질혈증의 진단은 혈액계측 정보 중에서 네 가지 LDL, HDL, TG, 그리고 TC를 이용하여 진단하며, 이상지질혈증 관리를 위해서는 LDL을 추정하는 것이 중요하다. 본 논문에서는 나이, 성별, 그리고 BMI와 같은 신체계측 정보를 학습하여 LDL-콜레스테롤을 예측하기 위한 준지도학습(Semi-supervised learning) 기반 기계학습 방법을 제안한다. 제안 방법은 얕은 학습(Shallow Learning)기반의 MLP(Multi-Layer Perceptron)을 이용하고, 이상지질혈증 진단인자간의 상관관계를 고려하여 신체계측 정보로 예측된 HDL, TG, 그리고 TC을 이용하여 일반적인 기계학습을 이용한 예측방법의 정확도를 개선한다. 즉, 제안방법은 신체계측 정보를 이용하여 혈액계측 정보의 LDL, HDL, TG, 그리고 TC을 각각 예측하고, 신체계측에 혈액계측의 예측 정보를 추가하여 학습한 준지도학습 기반 얕은 네트워크를 설계한다. 실험결과, HDL, TG, 그리고 TC의 혈액예측 정보를 이용한 준지도학습 기반 LDL 예측 정확도는 71.4%로 신체계측 정보만을 이용한 예측 방법의 67.0% 보다 약 4.4% 개선할 수 있음을 확인한다.

1. 서론

사회가 발달함에 따라 만성질환인 비만과 이상지질혈증 유병률이 점차적으로 증가하고 있으며[1] 이상지질혈증 발병을 진단하는 네 가지 임상적 기준 중 LDL을 낮추는 것이 이상지질혈증 치료의 일차적인 목적으로 알려져 있다[2]. 또한, LDL은 심혈관 질환과의 연관성을 띄고 있어[3] LDL-콜레스테롤 수치와 이상지질혈증 발병에 대한 조기 진단 및 관리는 중요한 문제이다. 이상지질혈증의 유병 원인으로 수면의 질, 비만, 성별 등이 위험인자로 지목되었으며[4], 이상지질혈증의 조기 진단 및 관리 문제를 해결하기 위해서는 지목된 위험인자와 이상지질혈증 발병을 판단하는 네 가지의 임상적 수치를 활용하여 이상지질혈증 발병에 대한 예측 방법이 필요하다.

본 논문에서는 혈액 수집을 필요로 하는 특징을 제외하고 신체를 계측할 수 있는 나이, 성별 그리고 BMI 등의 신체계측 정보와 이상지질혈증 발병을 진단하는 혈액계측 정보 중 LDL을 제외한 HDL, TG, 그리고 TC의 세가지 특징들을 이용하여 LDL-콜레스테롤을 예측하는 모델을 설계한다. 또한, 신체계측 정보만을 이용하여 예측하는 모델보다 높은 성능을 제공하기 위해서 준지도학습을 이용하여 HDL, TG, 그리고 TC의 혈액예측 정보를 추가적으로

학습하는 모델을 설계한다. 또한, 본 논문에서 이용한 기계학습 모델은 특징의 개수가 12~15개로 많지 않기 때문에 얕은 학습(Shallow Learning)기반의 MLP(Multi-Layer Perceptron)을 이용한다.

실험결과, HDL, TG, 그리고 TC의 혈액예측 정보를 이용한 준지도학습 기반 LDL 예측 정확도는 71.4%로 신체계측 정보만을 이용한 예측방법의 67.0% 보다 약 4.4% 개선할 수 있음을 확인한다.

본 논문의 구성은 다음과 같다. 2장에서는 연관인자의 특징, 임상적 정상범위 및 클래스 분류, 그리고 예측 결과를 추가 학습하는 준지도학습 방법을 설명한다. 3장에서는 실험 환경과 실험 결과를 설명하고, 4장에서는 결론을 설명한다.

2. 본론

2.1. 연관인자

본 논문에서는 LDL-콜레스테롤을 예측하기 위한 임상 데이터를 연세 의료원에서 제공받은 건강 관리 임상시험 데이터를 사용한다. 데이터 증가를 위해 226명 (20~68세의 남녀, 남성: 70명, 여성: 156명)의 8주간 1차, 2차 검사 결과 데이터를 분리하여 총 452개의 데이터로 진행한다. 혈액 수집을 요구하지 않는 신체계측을 통한 연관 인자는 총 12가지 AGE, SEX, BMI, PSQI, Muscle, Fat, SBP, DBP, HR, Waist, FatPercentage, 그리고 WHR를 이용한다.

2.2. 임상적 정상범위 및 클래스 분류

이상지질혈증은 표 1과 같이 Total Cholesterol, LDL, TG, 그리고 HDL의 4가지의 임상적 정상범위로 이상지질혈증 진단이 가능하다[5, 6]. 표 1은 이상지질혈증 진단을 위한 각 특징의 임상적 정상범위를 보여준다. 표 2는 표1을 기준으로 건강, 보통, 위험, 그리고 이상 수치의 특징을 나타내기 위해 임의적으로 정한 수치에 따른 클래스를 분류 기준을 보여준다.

<표 1> 이상지질혈증 임상적 정상범위 [5, 6]

검사명	정상범위
Total Cholesterol	200mg/dl 미만
LDL	100mg/dl 미만
TG	150mg/dl 미만
HDL	40mg/dl 초과

<표 2> 각 특징의 건강, 보통, 위험, 이상의 범위

	TC	LDL	TG	HDL
건강	~99	~49	~69	60~
보통	100~169	50~69	70~119	50~59
위험	170~199	70~99	120~149	49~41
이상	200~	100~	150~	~40

2.3. 준지도학습 기반 예측 정확도 개선 방법

이상지질혈증을 진단하는 각 특징은 서로에 대한 연관성이 있다[3]. 따라서 기존의 신체계측과 혈액계측 정보를 이용하여 정확도를 개선 할 수 있다. 준지도학습은 많은 양의 클래스가 없는 데이터와 적은 양의 클래스가 있는 데이터를 이용하여 좋은 분류기를 구축하여 정확도를 개선할 수 있다[7]. 따라서 본 연구에서는 신체계측과 혈액계측 정보를 이용한 준지도학습 기반 MLP를 사용해 LDL 예측 정확도를 개선한다. 즉, 학습이 완료된 준지도학습 방법의 레이블링 결과를 활용하여 학습 데이터의 특징의 종류를 확장한다. 따라서 추가적인 학습 데이터 생성을 통해 예측 성능을 높이는 방법인 준지도학습과 유사한 방법을 사용하여 정확도를 개선한다.

그림 1은 얕은 학습 네트워크의 MLP 모델을 이용하여 준지도학습 기반 LDL예측을 위한 학습 방법에 대한 구조를 보여준다. MLP의 학습을 위한 입력 (신체계측 및 혈액예측)과 출력 (혈액계측의 예측)에 따라 $M_{출력_입력}$으로 표현한다. 예를 들어, 신체계측(즉, X)를 이용하여 LDL, HDL, TG, 그리고 TC를 각각 예측하는 MLP는 각각 M_{LDL_X}, M_{HDL_X}, M_{TG_X}, 그리고 M_{TC_X}로 표현한다. 또한, LDL, HDL, TG, 그리고 TC의 예측은 LDL_{PRE}, HDL_{PRE}, TG_{PRE}, 그리고 TC_{PRE}로 각각 표현한다.

준지도학습의 정확도를 개선하기 하기 네 개의 M_{LDL_X}, M_{HDL_X}, M_{TG_X}, 그리고 M_{TC_X}에서 정확도가 높은 순서대로 나열한 뒤, 신체계측 정보와 이전 모델이 예측한 값을 학습하도록 준지도학습 기반 네트워크를 설계한다. 즉, X를 이용하여 LDL_{PRE}을 예측하고, 정확도가 높은 순서인 HDL_{PRE}는 이전의 LDL_{PRE}을 이용하여 $M_{HDL_X+LDL_{PRE}}$를 이용하여 학습한다. TG_{PRE}는 $X+LDL_{PRE}+HDL_{PRE}$를 이용하여 $M_{TG_X+LDL_{PRE}+HDL_{PRE}}$를 통하여 생성하고, TC_{PRE}는 $X+LDL_{PRE}+HDL_{PRE}+TG_{PRE}$를 이용하여 $M_{TC_X+LDL_{PRE}+HDL_{PRE}+TG_{PRE}}$를 통하여 TC_{PRE}를 생성한다. 마지막으로 $M_{LDL\text{-}X+HDL_{PRE}+TG_{PRE}+TC_{PRE}}$은 $X+HDL_{PRE}+TG_{PRE}+TC_{PRE}$를 학습하여 LDL_{PRE_LAST}를 예측한다.

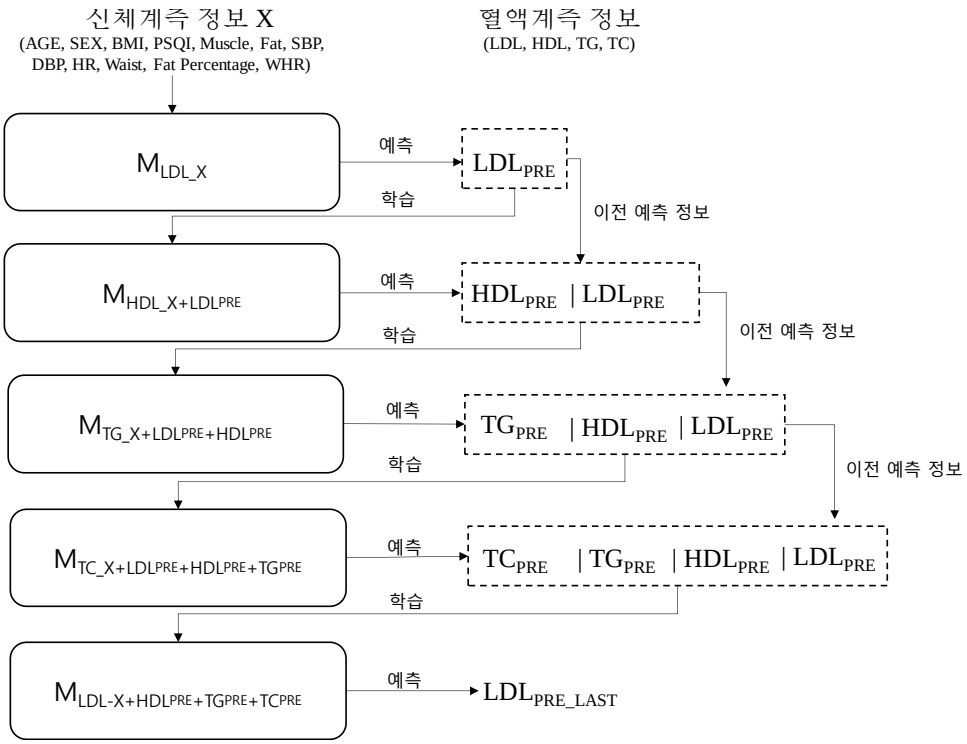

그림 1. MLP 와 준지도학습 기반 신체계측과 혈액예측을 이용한 LDL 예측 방법

표 3은 MLP 기반 LDL 예측을 위하여 필요한 각각의 학습 데이터 요소를 보여준다.

<표 3> 다양한 특징

	MLP	학습을 위한 특징
시나리오1	$M_{LDL\text{-}X}$	신체계측 정보 X (AGE, SEX, BMI, PSQI, Muscle, Fat, SBP, DBP, HR, Waist, Fat Percentage, WHR)
시나리오2	$M_{HDL\text{-}X+LDLPRE}$	$X + LDL_{PRE}$
	$M_{TG\text{-}X+LDLPRE+HDLPRE}$	$X + LDL_{PRE} + HDL_{PRE}$
	$M_{TC\text{-}X+LDLPRE+HDLPRE+TGPRE}$	$X + LDL_{PRE} + HDL_{PRE} + TG_{PRE}$
	$M_{LDL\text{-}X+HDLPRE+TGPRE+TCPRE}$	$X + HDL_{PRE} + TG_{PRE} + TC_{PRE}$
시나리오3	$M_{LDL\text{-}X+HDL+TG+TC}$	X + HDL + TG + TC

3. 실험 환경 및 실험 결과

3.1. 실험 환경

본 연구의 데이터는 226명의 8주간 측정한 1차, 2차 검사 결과 데이터를 분리하여 총 452개 (20~68세 남녀, 남성: 70명, 여성: 156명)의 임상 데이터로 진행하였다. MLP의 학습데이터로 361개의 훈련 데이터 및 91개의 테스트 집합으로 구성하였다.

본 논문에서는 LDL 예측을 위하여 MLP를 사용하여 모델을 구성하고, 예측 정확도를 측정하였다. 임상 데이터에서 추출한 표 3의 특징들을 Standard Scaler을 사용하여 데이터 정규화를 진행한 후, 데이터를 80% 비율인 361개의 훈련집합과 20% 비율인 91개의 테스트 집합으로 나눈 뒤 레이블 데이터를 원-핫 인코딩(One-hot Encoding)하여 학습을 진행하였고, 테스트 집합을 사용하여 정확도를 측정하였다. 이상지질혈증 진단 인자를 예측할 MLP 모델은 n개의 특징으로 구성된 훈련집합을 입력층에 넣어 128-64-32개의 노드를 가진 세 개의 은닉층을 거친 뒤, 네 개의 출력층 노드를 통해 예측 결과를 분류하였다. 또한, 과적합 방지를 위하여 세 개의 은닉층 사이에 두번의 Dropout (비율은 0.2)를 적용하고, 하이퍼파라미터 조정을 위한 배치 크기(batch_size)를 8, 학습 횟수(epoch)는 150으로 각각 설정하였다.

3.2 실험 결과

본 논문에서는 준지도학습 기반 제안방법의 정확도 개선 정도 비교를 위한 세 가지 시나리오를 구성하였다. 첫 번째 시나리오는 신체계측 정보 X를 사용하여 LDL을 예측하는 방법이다. 또한, 두 번째 시나리오는 준지도학습 방법을 이용하여 HDL_{PRE}, TG_{PRE}, 그리고 TC_{PRE}를 순차적으로 예측하여, 예측

결과를 추가로 학습하여 다음 예측을 위한 학습 데이터로 사용하여 LDL을 예측하는 방법이다. 마지막으로 세 번째 시나리오는, X와 혈액계측 정보 HDL, TG, 그리고 TC를 학습하여 LDL을 예측하는 방법으로 두 번째 시나리오와의 정확도 비교를 위해 구성하였다.

표 4는 LDL을 예측하기 위한 MLP 모델 기반 세 가지 시나리오의 예측 정확도를 보여준다. 신체계측 정보 X를 학습한 뒤 LDL, HDL, TG, 그리고 TC로 예측한 M_{LDL_X}, M_{HDL_X}, M_{TG_X}, M_{TC_X}의 예측 정확도는 각각 67.0%, 56.0%, 53.9%, 그리고 49.5% 였다. 따라서 순차적으로 LDL, HDL, TG, 그리고 TC의 예측 순서대로 준지도학습을 적용하여 각각의 혈액계측 정보를 예측하였다. 또한 시나리오 1의 신체계측 정보 X를 이용한 LDL 예측 정확도는 67.0% 이었다. 시나리오 2는 신체계측 정보 X에 따라 $M_{HDL\text{-}X+LDL_{PRE}}$, $M_{TG\text{-}X+LDL_{PRE}+HDL_{PRE}}$, $M_{TC\text{-}X+LDL_{PRE}+HDL_{PRE}+TG_{PRE}}$, $M_{LDL\text{-}X+HDL_{PRE}+TG_{PRE}+TC_{PRE}}$순으로 각각의 혈액 계측 정보를 예측하였으며 LDL은 71.4%로 측정되었다. 마지막으로 혈액 예측 정보가 아닌 계측 정보를 이용하였을 때의 정확도는 79.1%로 측정되었다.

<표 4> 특징에 대한 시나리오 별 LDL 예측 정확도

특징에 따른 정확도(%)		
시나리오1	신체계측	67.0
시나리오2	신체계측 + 혈액예측 (제안방법)	71.4
시나리오3	신체계측 + 혈액계측	79.1

시나리오별 정확도를 비교했을 때, 시나리오 3은 시나리오 1보다 12.1% 높은 정확도를 제공하였다. 즉, LDL, HDL, TC, 그리고 TG는 서로 연관성이 있고, 따라서 혈액계측 정보를 신체계측 정보와 함께 학습데이터로 활용하는 것이 신체계측 정보만 이용하는 것 보다 높은 정확도를 제공하였다. 즉, 혈액계측 정보를 얻기 힘든 상황에서 신체계측 정보 뿐만 아니라 혈액예측 정보를 함께 이용하여 LDL을 예측하는 제안방법의 시나리오 2는 시나리오 1 보다 4.4% 높은 성능을 제공하였음을 확인하였다. 따라서 제안방법은 준지도학습 기반의 이전 혈액예측 데이터를 학습하여 LDL의 예측 성능을 개선할 수 있음을 확인하였다.

4. 결론

본 논문에서는 나이, 성별, 그리고 BMI와 같은 신체계측 정보를 학습하여 LDL-콜레스테롤을 예측하는 모델의 정확도 개선을 위해 준지도학습 방법을 적용한 기계학습 기반 LDL-콜레스테롤을 예측하는 방법을 제안하였다. 제안 방법은 얕은 학습기반의 MLP을 이용하였고, 이상지질혈증 진단인자간의 상관관계를 고려하여 신체계측 정보로 예측된 HDL, TG, 그리고 TC를 준지도학습을 이용하여 정확도를 개선하였다. 실험결과, 신체계측 정보를 학습하여 LDL을 예측한 시나리오 1, 준지도학습 방법을 적용한 시나리오 2, 그리고 신체계측과 혈액계측 정보를 학습하여 LDL을 예측하는 시나리오 3의 정확도는 각각 67.0%, 71.4%, 그리고 79.1%로 측정되었다. 특히, 신체계측과 혈액계측 정보를 학습한 시나리오 3의 방법은 79.1%의 정확도를 제공하였다. 따라서 각 이상지질혈증을 위한 혈액계측 정보는 서로간의 연관성이 있음을 확인하였고, 본 논문에서는 정확도 개선을 위한 방법으로 혈액계측 정보를 예측하는 준지도학습의 아이디어를 적용하였다. 실험결과 시나리오 1과 비교하여 제안방법의 시나리오 2는 혈액계측 정보를 예측하는 준지도학습을 적용하여 4.4%의 정확도를 개선하였다.

ACKNOWLEDGEMENT

본 연구는 2022년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임[S2935743], 본 연구는 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20002781)

참고문헌

[1] Ah-Reum Han, The association between sleep and metabolic syndrome, Department of Medicine the Graduate School, Yonsei University, 2007.

[2] Sang-Hak Lee, Diagnosis and treatment of dyslipidemia, The Korean Journal of Medicine, Vol.74, No.4, 2008.

[3] Committee of Clinical Practice Guideline of the Korean Society of Lipid and Athersclerosis, Korean Guidelines for the Management of Dyslipidemia 4th ed, 22,2018.

[4] Su-bhin Yang, Min-tae Kim, Su-bin Kwon, Hak-jae Kim, Tai-kyeong Jeong, Sung-ju Lee, HDL Cholesterol Prediction based on Shallow Learning using Dyslipidemia-related factors, Proceedings of Symposium of the Korean Institute of Information Scientists and Engineers, 1490-1492, 2021.

[5] Seon-min Lee, Min-tae Kim, Su-bin Yang, Hak-jae Kim, Tai-kyeong Jeong, Sung-ju Lee, Important Clinical Data Selection for Machine Learning Accuracy of Hyperlipidemia Digonosis, Proceedings of Symposium of the Korean Institute of communications and Information Sciences, 1426-1427, 2021.

[6] I. Jeong, 고지혈증/비만. 대한외과학회 학술대회 초록집, 161-162, 2018.

[7] A-Leum Kim, Sung-Bae Cho, A Method for Generating Data based on Semi-supervised Learning for Predicting Defaults in Social Lending, Proceedings of Symposium of the Korean Institute of Information Scientists and Engineers, 696-698, 2017.