고지혈증 진단의 기계학습 정확도 개선을 위한 중요 임상데이터 특징 선택

이선민[1], 김민태[1], 양수빈[1], 김학재[2], 정태경[3], 이성주[1†]
상명대학교[1], ㈜클래스액트[2]
{201821005[1], 201820985[1], 201921007[1], peacfeel[1†] }@smu.ac.kr

{krunivs}@gmail.com,

{ttjeong}@hallym.ac.kr

Important Clinical Data Selection for Machine Learning Accuracy of Hyperlipidemia Digonosis

Seonmin Lee[1], Mintae Kim[1], Subin Yang[1], Hakjae Kim[2], Taikyeong Jeong[3], Sungju Lee[1†]
Department of Software, Sanmyung University[1]
CLASSACT Incorporated[2]
School of Artificial intelligence Convergence Hallym University[3]

요 약

기계학습 방법을 이용한 고지혈증의 정확한 진단을 위하여 임상데이터의 중요한 특징을 분석하는 것은 중요한 문제이다. 본 논문에서는 25가지의 고지혈증 임상데이터 특징 중에서 10가지의 중요한 특징만을 이용하여 고지혈증 진단 정확도를 개선하는 방법을 제안한다. 즉, 임상적으로 정의된 정상범위와의 관계성과 임상데이터 내의 특성을 분석한 특징 중요도를 분석해 우선순위가 높은 특징을 선택하여 기계학습의 정확도를 개선한다. 실험결과, 총 148명의 고지혈증 임상데이터를 SVM(Support Vector Machine), Decision Tree, 그리고 Random Forest를 통해 진단율을 각각 확인한다. 또한, 제안방법을 통하여 중요도 전체 25가지의 특징 중 10가지의 중요한 특징을 우선적으로 선택하여 정확도를 최대 13.3% 개선할 수 있음을 확인한다.

Ⅰ. 서 론

고지혈증은 혈액 중의 지질의 일종인 콜레스테롤이나 중성지방의 양이 정상수치보다 많은 상태를 뜻하며, 협심증, 심근경색, 뇌졸중 등 심혈관계 질환으로 위험한 합병증을 일으키기 때문에 정확한 진단 방법이 중요하다. 즉, 과다한 양의 지질이 혈액 내에 있을 경우, 지방 성분이 동맥벽에 침착되어 혈관이 좁아지게 되고, 그에 따라 심장과 뇌의 혈관질환의 발생 위험이 높아지게 된다[1]. 고지혈증 혈액검사 과정에서 주로 Cholesterol, LDL, TG, HDL 수치를 확인한다. 최근 이러한 수치뿐만 아니라, 추가적으로 고지혈증에 영향을 미치는 임상데이터 특징을 기계 학습적인 방법을 적용하는 연구가 보고되었다[2, 3].

본 논문에서는 25가지의 고지혈증 임상데이터 특징 중에서 중요한 임상데이터 특징 10가지만을 선택적으로 이용하여 고지혈증 진단 정확도를 개선하는 방법을 제안한다. 중요한 임상데이터 특징을 선택하기 위해서 잘 알려져 있는 임상적인 정상범위 기준과 환자와 비환자간의 관계성 분석 방법을 이용하여 중요한 특징을 선택한다. 즉, 기계학습에 이용하는 중요한 임상데이터는 임상적으로 정의된 기준으로 4가지의 특징을 우선적으로 선택하고, 임상적으로 정의되어있지 않은 나머지 21가지의 특징 중에서 임상데이터의 고지혈증 환자와 비환자의 특징을 비교하여 총 10가지의 특징을 선택한다.

제안방법은 148명의 고지혈증 여성 임상데이터(49~72세, 음성: 90명, 양성: 58명)에 대하여 SVM(Support Vector Machine)[4], Decision Tree[5], 그리고 Random Forest[6]의 세 가지 기계학습 방법을 이용하여 정확도를 각각 측정한다. 실험결과, 제안방법은 25가지의 특징 대신 10가지의 특징을 이용하여 기계학습을 통한 고지혈증 진단의 정확도를 최대 13.3% 개선할 수 있음을 확인한다.

본 논문의 구성은 다음과 같다. 2장에서는 중요도가 높은 상위 10가지의 특징을 선택하는 방법을 설명한다. 3장, 4장에서는 실험결과와 결론을 각각 설명한다.

Ⅱ. 본론

2.1. 주어진 임상데이터

본 논문에서는 고지혈증을 진단하기 위한 임상데이터를 총 25가지의 특징(Cholesterol, LDL, TG, HDL, Age, Neutrophil, AST, Waist, Lympho, Creatinine, Weight, Muscle, DBP, Fat, Height, Insulin, WBC, Glucose, HR, Fat percentage, SBP, WHR, BMI, ALT, CRP)을 이용하여 SVM, Decision Tree, 그리고 Random Forest를 통하여 기계학습의 정확도를 측정한다.

2.2. 임상적 정상범위

고지혈증은 표 1과 같이 총 Cholesterol, LDL, TG, 그리고 HDL의 4가지의 임상적 정상범위로 고지혈증 진단이 가능하다[1]. 따라서 임상적으로 정의된 4가지의 특징은 기계학습의 정확도를 개선하기 위한 중요한 특징으로 간주한다. 표 1은 고지혈증 진단을 위한 각 특징의 임상적 정상범

위를 보여준다.

표1. 고지혈증 진단을 위한 각 특징의 임상적 정상범위 [1]

검사명	정상범위
총 Cholesterol	200mg/dl 미만
LDL	100mg/dl 미만
TG	150mg/dl 미만
HDL	40mg/dl 이상

2.3. 고지혈증 환자와 비환자의 특징 분포 비교

임상적 정상범위가 정의되지 않은 나머지 21가지의 특징 중에서 중요한 특징을 선택하기 위해서 임상데이터 내의 환자와 비환자의 관계를 분석하여 중요한 특징을 선택한다. 즉, 환자의 특징 분포도와 비환자의 특징 분포도가 유사하지 않으면 고지혈증을 진단할 수 있는 특징 변화가 있다고 간주하여 정확도를 개선할 수 있는 중요 특징으로 선택한다. 정확한 수치로 환자와 비환자 간의 데이터 분포도 차이를 분석하기 위해서 환자와 비환자간의 특징 데이터를 MSE(Mean Squared Error) 방식을 통해 분석한다. 환자와 비환자간의 MSE 값이 클수록 두 데이터 의 분포도 차이가 커지고. 반대로 MSE 값이 작을수록 데이터 분포도 차이가 낮아진다.

표 2는 전체 25가지의 특징 중 표 1에서 정의된 4가지 특징(Cholesterol, LDL, TG, HDL)과(1번부터 4번) 임계값 3×10^{-3} 이상의 MSE 값을 갖는 중요도 상위 6개의 특징들을(5번부터 10번) 보여준다. Cholesterol, LDL, TG, HDL 이외의 5번부터 10번까지의 특징들은 임계값 3×10^{-3} 이상의 값을 갖기 때문에 정확도를 개선하기 위한 중요한 특징으로 간주한다. 나머지 15개의 특징들은 임계값 3×10^{-3} 미만의 값을 갖기 때문에, 고지혈증을 진단하는 중요한 특징으로 선택하지 않는다.

표 2. 임계값 (3×10^{-3}) 이상의 MSE 값을 지닌 특징 10개

번호	특징	MSE (10^{-3})	번호	특징	MSE (10^{-3})
1	Cholesterol	5.7	6	Neutrophil	5.4
2	LDL	4.6	7	AST	4.4
3	TG	1.2	8	Waist	4.1
4	HDL	1.1	9	Lympho	3.5
5	Age	14.5	10	Creatinine	3.0

2.4. 실험 환경

본 연구의 데이터는 연세 의료원에서 제공하는 고지혈증 임상시험 데이터를 사용하였다. 총 148개 (49~72세의 여성, 음성: 90명, 양성: 58명)의 고지혈증 임상데이터를 중심으로 진행하였으며, 총 25가지의 특징 샘플 (Cholesterol, LDL, TG, HDL, Age, Neutrophil, AST, Waist, Lympho, Creatinine, Weight, Muscle, DBP, Fat, Height, Insulin, WBC, Glucose, HR, Fat percentage, SBP, WHR, BMI, ALT, CRP)을 이용하였다. 기계학습의 학습데이터는 118개의 훈련 집합 및 30개의 테스트 집합으로 구성하였다. SVM, Decision Tree, 그리고 Random Forest를 통하여 주어진 고지혈증 임상데이터를 기반으로 진단 정확도를 측정하였다.

2.5. 실험 결과 분석

본 논문에서는 잘 알려져 있는 임상적인 정상범위 기준과 환자와 비환자간의 관계성을 분석하여 우선순위가 높은 특징선택의 결과를 기반으로 25가지의 임상데이터 특징 중에서 10가지의 임상데이터(Cholesterol, LDL, TG, HDL, Age, Neutrophil, AST, Waist, Lympho, Creatinine) 특징을 선택하였다. 표 3은 기존의 25가지의 특징과 각 임계값 기준에 따른 특징 개수에 대한 SVM, Decision Tree, 그리고 Random Forest의 정확도 비교를 보여준다. 25가지 모든 특징을 사용한 경우 정확도는 각각 53.3%, 63.3%, 그리고 59.8%로 측정되었다. 반면에 임계값이 3×10^{-3} 이상인 10가지 특징을 사용한 경우, 표 3와 같이 정확도가 각각 66.6%, 73.3%, 그리고 70.2%로 측정되었고, 13.3%, 10%, 그리고 10.4% 개선되었음을 확인하였다.

표 3. 기계학습의 고지혈증 진단 정확도 비교

임계값 (10^{-3})	사용 특징개수	SVM (%)	Decision Tree (%)	Random Forest (%)
0	25	53.3	63.3	59.8
1	23	56.6	66.6	68.5
2	17	63.3	66.6	68.5
3	10	66.6	73.3	70.2

Ⅲ. 결론

본 논문에서는 고지혈증 임상데이터의 중요한 특징을 선택하여 기계학습을 이용한 고지혈증 진단의 정확도를 개선하는 방법을 제안하였다. 25가지의 임상데이터 특징 중, 임계값 3×10^{-3}을 넘는 10가지의 특징을 선택한 결과, 약 10~13.3% 진단 정확도를 개선하였다. SVM, Decision Tree, 그리고 Random Forest의 세 가지 방법 중, SVM 방법이 13.3%로 정확도가 가장 많이 개선되었고, Decision Tree 방법이 73.3%로 가장 높은 정확도를 제공하였다.

ACKNOWLEDGMENT

본 연구는 2020년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임. [S2935743]

참 고 문 헌

[1] I. Jeong, 고지혈증/비만. 대한외과학회 학술대회 초록집, 2018

[2] S. Lee, and T. Shin, “Development and application of prediction model of hyperlipidemia using SVM and meta-learning algorithm,” Korea Intelligent Information Systems Society, 2018

[3] Y. Liu, and et. al., “Deep Learning-Based Method of Diagnosing Hyperlipidemia and Providing Diagnostic Markers Automatically,” Diabetes Metab Syndr Obes, 2020

[4] S. Park, K. Kim, J Lee, and S Lee, “Red Tide Prediction using Neural Network and SVM,” Proceeding of The Institute of Electronics and Information Engineers, 2011

[5] S. Lee, and H. Jin, “Analysis and Prediction of Diabetic Patients using Decision Tree,” Proceeding of The Institute of Electronics and Information Engineers, 2013

[6] J. Moon, S. Jung, H. Kim, and E. Hwang, “Daily Occurrence Prediction of Regional Infectious Diseases Using Random Forest,” Proceeding of The Korean Institute of Information Scientists and Engineers, 2019