The Journal of the Convergence on Culture Technology (JCCT)
Vol. 9, No. 3, pp.737-744, May 31, 2023. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2023.9.3.737

JCCT 2023-5-87

기계학습 기반 근감소증 예측을 위한 데이터 전처리 기법

Data Preprocessing for Predicting Sarcopenia Based on Machine Learning

최윤*, 윤유림**

Yoon Choi*, Yourim Yoon**

요 약 근감소증은 노인들 사이에서 점점 더 흔하게 발생하고 있어, 최근 주목을 받고 있는 질병이다. 근감소증의 원인은 매우 다양하게 나타나지만, 노화, 식습관, 운동 부족등이 주요한 원인들 중 하나이다. 근감소증은 원인이 다양한 만큼 예방 및 치료에 전략을 개발하는 것이 중요하다. 하지만 요인이 다양한 만큼 사람이 근감소증을 정확하게 예측하기는 어렵다. 여기서 기계학습을 이용해 근감소증 예측의 정확도와 편의를 크게 높일 수 있다. 그러나 생활습관과 생체 데이터의 양은 방대한 만큼, 전처리 없이 데이터를 쓰기에는 시간복잡도와 정확성 측면에서 부적절할 수 있다. 본 논문에서는 근감소증과 그 원인에 대한 최신 문헌을 검토하고, 그에 맞게 기계학습 기만 근감소증 예측에 활용할 데이터를 전처리하는데 초점을 맞춘다.

주요어 : 근감소증, 노화, 기계학습, 데이터 전처리

Abstract Sarcopenia is an increasingly common disease among the elder that has recently received attention. Although the causes of sarcopenia are diverse, aging, dietary habits, lack of exercise are the one of the major factors. As the causes of sarcopenia are diverse, it is important to develop strategies for prevention and treatment. However, predicting sarcopnia accuartely is difficult due to the variety of factors involved. Here, machine learning can significantly improve the accuracy and convenience of predicting sarcopenia. However, since lifestyle habits and biological data are vast, using data without preprocessing may be inappropriate in terms of time complexity and accuracy. This paper reviews recent literature on sarcopnia and its causes, focusing on preprocessing the data to be used in sarcopnia prediction machine learning accrodingly.

Key words : Sarcopenia, Aging, Machine Learning, Data Preprocessing

Ⅰ. 서 론

근감소중은 골격근의 질량과 기능이 상실되는 것이 특징인 질환이다. 근감소증은 점진적인 골격근 질량 및 강도의 손실로 특징 된다[1]. 근감소증은 다양한 원인 요소를 가지고 있다. 나이, 유전적 요인, 자율신경계의 이상, 질병 및 외상 등이 근감소증을 일으킬 수 있다. 근감소증은 일상생활에서의 동작, 활동 등을 제한 할

*준회원, 가천대학교 컴퓨터공학과 학부연구생 (제1저자)
**정회원, 가천대학교 컴퓨터공학과 부교수 (교신저자)
접수일: 2023년 3월 30일, 수정완료일: 2023년 4월 14일
게재확정일: 2023년 5월 8일

Received: March 30, 2023 / Revised: April 14, 2023
Accepted: May 8, 2023
**Corresponding Author: yryoon@gachon.ac.kr
Dept. of Computer Engineering, Gachon Univ, Korea

수 있어, 환자의 삶의 질을 크게 저하시키는 요소이다. 특히 근감소증은 노령층에서 엄청난 삶의 질 저하를 일으킨다. 특히 근감소증을 조기 발견하고, 치료와 적극적인 관리가 필요하다.

노화에 따른 신체변화의 일반적인 것은 근골격의 약화와 지방량의 증가이다. 지방량의 증가는 대사성 질환의 유병률의 증가와 골격의 약화는 골다공증의 위험성과 골절 위험율의 증가로 나타난다. 하지만 근육량의 감소는 아직 의학적으로 큰 관심을 끌지 못하고 있다[2]. 노령화는 인간이 직면한 가장 큰 사회 문제 중 하나이다. 현재 우리 사회는 고령화 사회로 진입했고, 노인들의 비율 역시 높아지고 있다. 고령자가 전체 인구의 증가율 보다 8배 이상으로 빠르게 증가하게 있고, 의료 비용이 점점 비싸지고 있는 상황에서 사회적 정치적으로 노인들에 대한 복지문제는 중요한 위치를 차지하고 있을 수 밖에 없다[2]. 이에 따라 노화와 관련된 건강 문제 또한 더욱 중요한 이슈가 되고 있다. 노인들에게 근감소증이 치명적인 만큼, 근감소증은 각 나라의 평균 수명의 단축을 가져오고, 평균적인 국민 건강에 악영향을 끼친다. 따라서 이에 대한 보건의료의 관심이 늘어나고 있고, 근감소증에 대한 연구 또한 늘어나고 있다[3].

그 중에서는 기계학습을 근감소증 진단에 쓰려는 연구도 있었다. 근감소증에 있어서 진단에 가장 확실한 방법은 근육의 양과 질을 측정한 방법이다. 근육의 양과 질을 측정하기 위한 가장 정확한 방법은 컴퓨터 단층 촬영(CT)이나 자기 공명 영상 촬영을 통해서 근육량을 수동으로 측정하는 것이다[4]. CT를 이용해 촬영한 근육을 이용해서 기계학습을 학습시켜 근감소증을 예측한 연구가 있었다[5]. 하지만 CT를 이용한 촬영 방식에는 엄청난 시간과 에너지가 필요하다는 문제점이 있었다. 모든 환자에 대해서 CT 촬영를 실시하기에는 의료 자원의 문제와, 환자의 시간 등의 문제가 존재하기 때문이다.

또 다른 선행연구에서는 한국 질병관리청에서 조사한 국민건강영양조사의 2008년부터 2011년까지의 데이터를 사용했다[6]. 그리고 4개의 분류 모델을 사용해서 어떤 분류 모델이 가장 효과적인지를 연구하고, 어떤 요소가 sarcopenia에 더 영향을 미치는 요소인지를 밝혔다. 하지만, 선행 연구에서는 데이터 전처리 과정이 나와 있지 않았고, 데이터 전처리가 기계학습으로 sarcopenia를 진단하는데 있어서 어떤 영향을 주는지는 알 수 없었다.

기계학습에서 feature selection은 기계학습 모델을 위해 사용 되는 입력 데이터의 특징들 중 중요한 것들을 선택 하는 프로세싱 기법이다. 모델의 학습과 예측 위해 필요한 데이터의 특징을 줄이고, 불필요한 데이터를 제거하여 모델의 정확도를 향상 시키고, 더 빠른 학습을 가능하게 만든다[7]. 일반적으로 입력한 데이터의 특징이 많을 때, feature selection은 더 큰 효율을 보여준다. 입력한 데이터의 특징이 많으면 학습시간이 길어지고, 필요하지 않은 데이터로 인한 과대적합이 발생할 가능성이 커진다[8]. 이런 문제를 해결 하기 위해서 feature selection을 적용하면, 입력할 데이터의 특징을 줄이고, 불필요한 부분을 제거하여, 모델의 정확도와 효율성을 높일 수 있다. 생활 습관과 식습관에 관한 데이터의 양은 방대하기 때문에 데이터 전처리, 즉 feature selection 없이 기계학습에 적용한다면, 시간복잡도와 컴퓨터 자원 면에서 불필요한 소모를 초래할 수 있다. 이 논문에서는 데이터 전처리와 feature selection을 통해 기계학습을 통한 근감소증 예측의 시간효용성과 성능을 올리고자 한다.

II. 실험 방법

1. 데이터

실험에는 한국 질병관리청에서 진행한 국민건강영양조사 데이터와 한국고용정보원에서 진행한 고령화 패널 조사(KLoSA) 데이터를 활용해서 진행했다. 국민건강영양 조사의 데이터는 2008년에 진행된 조사의 데이터를 사용했다[9]. 고령화연구패널 조사의 데이터는 2020년에 진행된 8차 조사의 데이터를 활용했다[10].

2. 근감소증 진단 기준

국민건강영양조사의 데이터와 고령화연구패널 조사의 데이터의 내용이 다르기 때문에 서로 다른 진단 기준을 사용하였다.

국민건강영양조사에서는 팔과 다리의 제지방량과 키, 그리고 성별을 통해서 근감소증을 판별했다. 먼저 Skeletal muscle mass index(SMI)를 측정하는 공식을 사용했다. SMI는 사지 골격근량을 키로 나눈 지표이다. Asia Working Group for sarcopenia에서 남자의 경우 SMI$<$7.0kg/m^2 일 경우, 여자의 경우 SMI$<$5.4kg/m^2

일 경우 근감소증으로 진단했다[11].

고령화연구패널 조사에서는 악력 데이터와 성별을 사용하여 근감소증을 판별했다. 여자의 경우 악력 16kg 미만, 남자의 경우 악력 28 kg 미만일 경우 근감소증으로 진단했다[12].

3. 실험 환경 및 프로그램

근감소증 진단을 위한 기계학습 모델은 데이터마이닝 분석 패키지인 WEKA를 사용했다. 기계학습 실험은 3.2GHz AMD Ryxzen 7 8500H 프로세서와 16GB의 RAM을 가진 컴퓨터에서 수행되었다.

4. 실험 설정

먼저 실험에 쓰인 데이터들에 사용된 진단기준에 따라서 근감소증인지 아닌지를 나타내는 레이블을 추가하였다. 그 이후에, 기계학습 실험의 객관성을 위해서 근감소증 진단에 사용된 특징들을 삭제했다. 그 이후, 데이터 처리를 거친 후에 실험을 개시했다. 실험에서 쓰인 분류기로는 로지스틱 회귀를 사용했다. 로지스틱 회귀는 기계학습 기반 분류 방법 중에서 이진 분류에 속하는 알고리즘으로, 어떤 문제에 대해서 0과 1로 답을 할 수 있는 분류기다. 즉, 근감소증인지 아닌지 판별할 때, 적절한 답을 내놓을 수 있는 분류기라고 할 수 있다. Batch size는 전체 트레이닝 데이터들을 여러 작은 그룹으로 나누었을 때, 그 나눠진 그룹의 크기를 의미한다. k 개의 클래스가 있다고 가정하고 m개의 특징을 가진 n 개의 데이터가 있다고 하면, 파라미터 행렬 B는 m * (k-1) 행렬이 된다. 여기서 마지막 클래스를 제외한 어떤 클래스 j에 대한 가능성은 다음과 같다.

$$P_j(X_i) = \exp(X_i B_j) / (\sum_{j=1}^{k-1} \exp(X_i * B_j) + 1)$$

그리고 마지막 클래스는 다음 수식과 같은 가능성을 갖는다.

$$1 - (\sum_{j=1}^{k-1} P_j(X_i)) = 1 / (\sum_{j=1}^{k-1} \exp(X_i * B_j) + 1)$$

따라서 가능도 함수의 로그는 다음과 같다.

$$L = -\sum_{i=1}^{n} (\sum_{j=1}^{k-1} Y_{ij} * \ln(P_j(X_i)) + (1 - (\sum_{j=1}^{k-1} Y_{ij})) * \ln(1 - \sum_{j=1}^{k-1} P_j(X_i))) + ridge * (B^2)$$

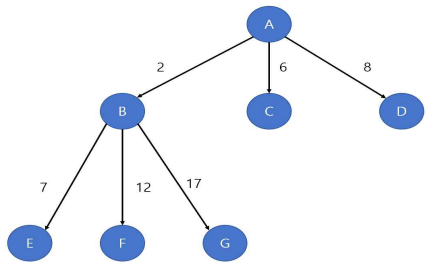
본 논문에서 ridge의 값은 1.0E-8 로 설정했다[13].

그리고 데이터 전처리를 통해서 각각 원본과 다른 데이터 두 개를 만들어서 실험했다. 각각의 특징 별로, 그 특징이 가질 수 있는 nominal 데이터의 종류를 기준으로 데이터를 삭제했다. 예를 들어, 어떤 특징이 [0, 1]를 종류로 가지고 있는 경우, 특징의 종류를 두 개 가지고 있는 것으로 생각한다. 국민건강영양조사와 고령화연구패널조사 모두 서로 다른 3개의 버전을 만들어서 실험했다. 서로 다른 3개의 버전은 원본(ver.1), nominal 데이터의 종류가 3개 이하인 것을 삭제한 버전(ver.2), nominal 데이터의 종류가 5개 이하인 것을 삭제한 버전(ver.3) 으로 나눴다. 즉 어떤 특징의 값의 가지수가 3개 이하면 그 특징은 모든 레이블에서 삭제한다. 이것이 ver.2 가 된다. 마찬 가지로, 특징이 가지는 값의 가지수가 5개 이하인 것을 삭제한 것이 ver.3 이다.

또한 데이터 별로 실험은 두 번씩 진행했다. 첫 번째는 별 다른 수정을 하지 않고 실험했고 , 두 번째는 feature selection을 적용하고 실험했다. 특징평가 방식은 상관관계 기반 feature selection을 사용했고, 탐색 방법은 최상 우선탐색을 사용했다. 상관 관계 기반 feautre selection은 각 특징의 개별 예측 능력과 각 특징 간의 중복 정도를 함께 고려해 부분 집합의 가치를 평가하는 방법이다. 목표 특징과 상관관계는 높지만, 상관 관계는 낮은 특징의 하위 집합이 선호된다. 최상 우선 선택 방식은 확장 중인 노드들 중에서 목표 노드까지 남은 거리가 가장 짧은 노드를 확장 하여 탐색하는 방법이다.

그림 1의 예에서, 수치가 낮을수록 최적이라고 가정하면 최상 우선 탐색은 A 노드에서 B를 택할 것이고, B 노드에서 E를 택할 것이다. 실험에서는 이와 같은 방식으로 feature selection을 진행했고, 그림 1과 같이 각 노드 간의 수치를 계산하는데 상관관계 기반 feautre selection을 사용하여 적용했다.

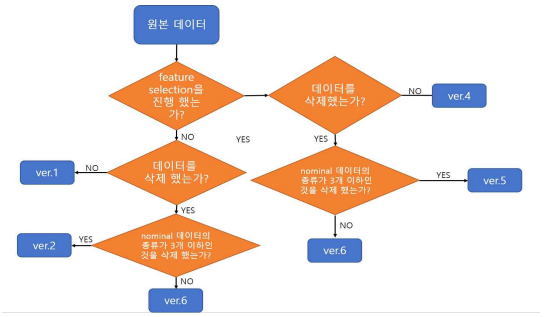
그림 1. 최상 우선 탐색의 예
Figure 1. The Example of the Best First Search Method



원본 데이터에 feature selection을 적용한 것을 ver.4로, nominal 데이터의 종류가 3개 이하인 것을 삭제한

데이터에 feature selection을 적용한 버전(ver.5), nominal 데이터의 종류가 5개 이하인 것을 삭제한 데이터에 feature selection을 적용한 버전(ver.6)에 대해 실험 결과를 정리하였다. 전체 데이터 전처리에 대한 흐름도를 그림 2에 정리하였다.

그림 2. 데이터 전처리 흐름도
Figure 2. The Flowchart of the Data Preprocessing



III. 실험 결과

1. 국민건강영양조사

표 2은 국민건강영양조사의 데이터로 실험한 결과의 특징 수와 시간과 기계학습 모델의 예측 정확도를 표로 정리한 것이다. feature selection을 하지 않고 실험을 진행했을 때는 특징을 어느 정도 삭제할수록 정확도와 시간 모두 좋아지는 결과를 보여줬다. 특히 nominal 데이터의 종류가 5개 이하인 것을 삭제한 ver.3의 경우에는 어떤 데이터도 삭제하지 않았을 때와 비교해서 거의 5배의 시간 효율성을 보여주고, 정확도의 경우에도 약 5% 상승한 것을 보여줬다.

표 2. 국민건강영양조사의 실험 결과
Table 2. The Result of Korea National Health and Nutrition Examination Surveys

국민건강영양조사			
버전	특징 수	시간	예측 정확도
ver.1	2619	659.94s	88.353%
ver.2	1731	510.1s	89.9436%
ver.3	788	123.85s	93.3299%
ver.4	36	1.2s	96.9728%
ver.5	36	1.3s	96.9728%
ver.6	47	2.84s	96.8702%

feature selection을 추가로 진행했을 때는 시간 효율성과 정확도 모두에서 주목할만한 성능의 상승을 이끌어 냈다. 동일한 데이터를 이용한, ver.1 과 ver.4를 비교했을 때 8% 이상의 정확도의 개선이 이루어졌다.

feature selection을 진행 하고 실험을 진행했을 때 특이한 점이 있었다. 일단 ver.4와 ver.5 의 경우, WEKA 프로그램이 완전히 동일한 특징을 선택했고, 그 결과 동일한 예측 정확도를 보여줬고, 시간 효율성 면에서도 거의 동일했다. 하지만 ver.6 에서는 더 많은 특징을 선택했으며, 정확도도 아주 미세하게 떨어지는 모습을 보여줬다. 특징 데이터를 더 많이 삭제하면서 근감소증 예측에서 중요한 위치를 차지하고 있는 특징을 삭제하면서 감소한 것으로 추측할 수 있다.

표 3는 feature selection으로 선택된 특징들을 표로 나열한 것이다. ver.4 와 ver.5는 완전히 동일하게 선택되었기 때문에 하나로 표기했다. 여기서 ver.4, ver.5와 ver.6 모두에서 등장하는 특징들은 근감소증 예측에 다른 특징들 보다 더 중요한 영향을 준다고 추측할 수 있다. 예를 들어, 나이,체중, 사지의 무게 등의 경우에는 모든 feature selection된 특징들에서 등장했다. 반대로 ver.6 에서만 등장하는 특징들은 상대적으로 근감소증 진단에 있어서 덜 중요한 특징들이라고 추측 가능하다. 하지만 선택된 특징들 모두 기계 학습을 통한 근감소증 진단에 영향을 준다는 것은 확실하다.

표 3.. 국민건강영양 조사에서 feature selection으로 선택된 특징
Table 3. The Features seleted in KNHNES

선택된 특징	
ver.4, ver.5	ver.6
나이	나이
가구조사 가중치	가구조사 가중치
만성 폐쇄성 폐질환 증상 2_기침기간	가구총소득 :소득액
신부전 외래경험	가구총소득 소득코드
당뇨병 의사진단여부	주생계부양자 3순위자
녹내장 치료	뇌졸증 와병일수
주관적 체형인식	류마티스성 관절염 와병일수
1년간 체중 변화 여부	결핵주치료기관
체중	주관적 체형인식
체질량지수	1년간 체중 증가량
비만유병여부	2차 이완기 혈압
공복혈당	체중
GOT	허리둘레
혈중크레아티닌	체질량지수
1초간 노력성 호기량	공복혈당
흉부영상 판독결과	GOT
기타 발현 여부	유로빌리노겐

1일 비타민 A 섭취량	1초간 노력성 호기량
제4번요추 골밀도	1초간 노력성 호기량/노력성 폐활량
대퇴골 전체 T-score(아시아기준)	흉부영상 판독결과
왼팔무게	우식경험 영구치수
오른팔무게	1일 비타민 A 섭취량
몸통 지방량	제 4번 요추 골밀도
오른다리 무게	대퇴골 전체 T-score(아시아기준)
총제지방량(머리제외)	총 골밀도
총무게	왼팔무게
총 제지방량	오른팔 무게
안명부 동통이나 압박감 여부	몸통 지방량
이과시진_안면신경마비	몸통 무게
시력검사_자동굴절/핀홀_교정시력_좌안	왼다리 무게
녹내장_안압_우안	오른다리 무게
손상2 발생시 활동	총제지방량(머리제외)
외래 3 이용이유 질병코드	총무게(머리제외)
외래 7 이용이유 질병코드	총지방량
	총제지방량
	외이도 협착유무_우이
	후두내시경
	어지럼증 검사_조건_시간
	시력검사_자동굴절/핀홀_교정시력_좌안
	녹내장_안압_우안
	녹내장_안압_좌안
	손상2 발생시 활동
	외래2 본인부담금
	외래 3 이용이유 질병코드
	외래 7 이용이유 질병코드
	외래 8 이용이유 질병코드

2. 고령화 연구 패널 조사

표 4은 고령화연구패널조사의 데이터로 실험한 결과의 특징 수와 시간과 기계학습 모델의 예측 정확도를 표로 정리한 것이다. Feature selection 하지 않고 실험을 진행했을 때는 특징을 어느 정도 삭제할수록 정확도와 시간 모두 좋아지는 듯 했으나, ver.3 정도로 과하게 삭제해서 근감소증 예측에 중요한 데이터를 삭제하게 될 경우 오히려 예측 정확도가 근소하게 떨어지는 모습을 보여줬다. Feature selection을 추가로 진행한 경우에는, 오히려 예측 정확도가 떨어지는 모습을 보여줬다. 이는 데이터를 삭제하면서 근감소증 예측에 효과적인 데이터가 삭제되어서 정확도가 떨어지는 것으로 추측된다. 물론 feature selection 자체의 효과는 존재했다. ver.1 과 ver.4를 비교해봤을 때, 예측 정확도

표 4. 고령화 연구 패널 조사의 결과
Table 4. The Result of KLoSA

고령화 연구 패널 조사			
버전	특징 수	시간	예측 정확도
ver.1	2272	294.79s	74.114%
ver.2	1099	150.37s	76.2034%
ver.3	911	161.98s	73.8952%
ver.4	15	0.11s	80.8937%
ver.5	14	0.16s	80.7396%
ver.6	22	0.24s	78.9676%

는 3% 이상의 개선을 보여줬고, 시간 효용도도 비교할 수 없을 정도로 개선이 이루어졌다. 이는 ver.2 와 ver.5를 비교했을 때도 마찬가지였다.

고령화 연구 패널 조사의 ver.5 과 ver.6을 비교해봤을 때, 특징 수는 늘어났지만 정확도가는 약 6.5% 정도 떨어진 것을 볼 수 있다. 이를 미루어 보아, 근감소증 예측의 정확도에 큰 영향을 미치는 특징이 삭제될 경우, 기계학습 모델은 정확도를 올리기 위한 특징을 더 탐색하지만, 오히려 시간 효율성과 정확도 모두 더 안 좋은 결과를 내놓는 것을 알 수 있다. 이는 국민건강영양조사의 ver.5와 ver.6의 데이터를 비교해봐도 알 수 있다. ver.5에 비해서 ver.6에서는 11개의 특징을 더 가졌지만, 정확도는 매우 근소하게 나마 떨어진 것을 확인 할 수 있었다. 고령화 연구패널 조사에서 한가지 더 특기할 것은 ver.3 과 ver.6을 비교하면 알 수 있다. 국민건강영양조사의 ver.3 과 ver.6을 비교하면 정확도가 개선된 것을 알 수 있지만, 반대로 고령화연구패널조사에서의 ver.3과 ver.6을 비교해보면, feature selection을 적용했을 때, 오히려 정확도가 소폭 떨어진 것을 확인 할 수 있었다. 특징을 과하게 삭제한 것이 feature selection에까지 영향을 미친 것이다.

표 5는 고령화연구패널 조사에서 feature selection으로 선택된 특징들을 나타낸 것이다. 여기서 정확도가 가장 높은 ver.5의 특징들이 근감소증과 관계가 가장 크다고 볼 수 있다. 여기서도 국민건강영양조사의 경우와 마찬가지로 특징들이 더 많이 선택되었다고 정확도가 높아지는 것은 아니라고 알 수 있다. 정확도에 영향을 미치는 것은 얼마나 많은 특징들이 선택되었나 보다는 어떤 특징이 선택되었나 임을 알 수 있다.

표 5. 고령화연구패널조사에서 feature selection으로 선택된 특징
Table 5. The Features seleted in KLoSA

선택된 특징		
ver.4	ver.5	ver.6
응답자의 출생년월	응답자의 출생 년월	응답자의 출생 년월
참여하고 있는모임(단체) 없음	인터뷰 날짜	인터뷰 날짜
인터뷰 날짜	C1 도움정도	첫번째 자녀의 연령
3번째 기타가족으로부터 수령 여부 현물지원	민간의료보험 가입여부	지난 기본조사 당시와 비교한 주관적 건강상태
C1 도움정도	C1 도움정도	청력
일상생활의 주변 도움 필요정도)교통수단을 이용하여 외출하기	민간의료보험 가입여부	틀니 치아수
민간의료보험 가입여부	악력측정 수락 여부	지난 1주일간의 느낌과 행동_우울감
1차 건강검진시 문제가 발견되어 무료 2차검진을 받은 경험여부	IADL 지수화	장소지남력_시/구/동/번지
지난 기본조사 이후 기타 병/의원 외래진료비로 지불한 총액	인지기능 구분	악력측정의 안정성 확인(손 수술 및 외상 경험등)
악력측정 수락 여부	금융자산 이자	악력 측정시 응답자의 노력정도
악력 측정시 응답자의 노력정도	월평균 예상 수급액	IADL 지수화
교육훈련 프로그램 경험 여부	주관적 기대감_나는 자식 세대가 우리 세대보다 더 나은 경제적/사회적 환경에서 살 수 있을 것 같다	ADL 지수화
월평균 예상 수급액	지난 1년간의 자원봉사활동 월평균 참여시간	인지기능 점수
주관적 기대감_나는 자식 세대가 우리 세대보다 더 나은 경제적/사회적 환경에서 살 수 있을 것 같다		도움정도
지난 1년간의 자원봉사활동 월평균 참여시간		지난해 월평균 소비_보건의료비
		월평균 예상 수급액
		현금 1천만원이 생길경우 항목별 사용금액_빚을 갚는다
		삶의 만족도_자신의 건강상태
		월평균 용돈
		지난 1년간의 여행,관광,나들이 경험 횟수
		지난 1년간의 자원봉사활동 월평균 참여시간

실험 결과, 누락된 값이 있는 데이터를 삭제하는 경우, 어느 정도 정확도와 효율성 면에서 향상이 있었으나, 데이터를 과하게 삭제했을 때는 오히려 정확도가 떨어지는 현상이 발생했다.

이는 데이터를 삭제 할 때, 너무 많은 특징을 삭제하면서, 근감소증 판단에 있어서 중요한 정보까지 삭제해버리기 때문에 발생한 것으로 보인다. 또한 feature selection을 진행한 경우에는 확실한 성능 향상을 기대할 수 있었다. 시간 효율성 면에서 feature selection을 진행하고 실험을 한 경우에는 매우 뛰어난 성능의 향상이 있었다. 다만 feature selection의 경우에도 데이터를 과하게 삭제한 경우에는 효율성이 떨어지는 경우가 있었다. 따라서 데이터를 일괄 삭제하는 것보다는, 적절한 방법을 통해서 근감소증 판단에 있어서 어떤 데이터가 중요한지 알아낸 이후에 데이터를 삭제하는 것이 좋을 것이다.

Ⅳ. 결론

본 연구에서는 근감소증 예측 기계학습 모델을 위한 데이터 전처리 기법에 대해 다루었다. 국민건강영양조사와 고령화 연구패널조사를 이용해서 데이터를 수집하고, 기준에 따라서 일정 부분의 데이터를 삭제하는 전처리와 feature selection을 통해서 모델의 정확도를 높이

고, 자원의 효율성을 높이고자 했다.

국민건강영양조사와 고령화연구패널조사의 데이터가 가지고 있는 데이터의 종류가 달랐기 때문에, 각각 다른 근감소증 진단 기준을 적용했다. 국민건강영양조사는 SMI 수치를 통해서 판별했고, 고령화 연구패널 조사는 악력 데이터를 통해서 판별했다.

기준을 통해서 데이터를 일괄 삭제하는 전처리는 두 데이터 모두 효과가 있었지만, 고령화연구패널 조사의 경우, 과하게 삭제하는 경우에는 근감소증 판별에 있어서 중요한 데이터를 삭제하는 경우가 생겨서 오히려 정확도가 떨어지는 경우가 있었다.

feature selection을 진행했을때는 두 데이터 모두 시간 효용도에서는 상당한 개선을 보여줬다. 예측 정확도에서도 개선이 되는 것을 보여줬지만, 국민건강영양조사에서 데이터가 과하게 삭제되었을 경우에는 오히려 정확도가 소폭 떨어지는 경우가 있었다.

따라서 데이터를 일괄 삭제하는 것보다는, 적절한 기준을 세우고, 그에 맞춰서 데이터를 삭제하는 것이 기계학습을 통한 근감소증 진단에 효율적일 것이다.

근감소증을 판별하는 확실한 진단 기준만 존재한다면, 데이터의 종류에 크게 상관없이, 간단한 데이터 전처리와 feature selection을 통해서, 근감소증 예측 기계학습 모델에 성능을 올리는데 효과적임을 알 수 있었다. 그리고 feature selection을 진행해서, 어떤 특징들이 근감소증 진단에 관련이 있는지 알아 낼 수 있었다.

하지만, 선택된 특징들과 근감소증 간의 선후관계까지는 기계학습으로 알아낼 수는 없었다. 미래연구에서는 의료 분야와 컴퓨터 분야의 긴밀한 협업을 통해서, 근감소증에 관련된 특징들이 무엇인지를 밝히고, 더 나아가 그 특징들이 근감소증과 정확히 어떤 관계를 가지는지까지 알아낼 수 있을 것이다.

References

[1] Santilli V, Bernetti A, Mangone M, Paoloni M. Clinical definition of sarcopenia. Clin Cases Miner Bone Metab. 2014 Sep;11(3):177-80. PMID: 25568649; PMCID: PMC4269139.

[2] Hong S, Choi WH. Clinical and physiopathological mechanism of sarcopenia. Korean J Med. 2012;83:444 - 54. doi: 10.3904/kjm.2012.83.4.444.

[3] Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, Jang HC, Kang L, Kim M, Kim S, Kojima T, Kuzuya M, Lee JSW, Lee SY, Lee WJ, Lee Y, Liang CK, Lim JY, Lim WS, Peng LN, Sugimoto K, Tanaka T, Won CW, Yamada M, Zhang T, Akishita M, Arai H. Asian Working Group for Sarcopenia: 2019 Consensus Update on Sarcopenia Diagnosis and Treatment. J Am Med Dir Assoc. 2020 Mar;21(3):300-307.e2. doi: 10.1016/j.jamda.2019.12.012. Epub 2020 Feb 4

[4] Albano, D.; Messina, C.; Vitale, J.; Sconfienza, L.M. Imaging of sarcopenia: Old evidence and new insights. Eur. Radiol. 2020, 30, 2199 - 2208, DOI : https://doi.org/10.1007/s00330-019-06573-2

[5] Kim YJ. Machine Learning Models for Sarcopenia Identification Based on Radiomic Features of Muscles in Computed Tomography. International Journal of Environmental Research and Public Health. 2021; 18(16):8710. https://doi.org/10.3390/ijerph18168710

[6] Kang YJ, Yoo JI, Ha YC. Sarcopenia feature selection and risk prediction using machine learning: A cross-sectional study. Medicine (Baltimore). 2019 Oct;98(43):e17699. doi:10.1097/MD.0000000000017699.

[7] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94 (November 2018), 45 pages. DOI : https://doi.org/10.1145/3136625

[8] Xue Ying 2019 J. Phys.: Conf. Ser. 1168,022022, DOI : 10.1088/1742-6596/1168/2/022022

[9] https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do

[10]https://survey.keis.or.kr/klosa/klosa04.jsp

[11]Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, Jang HC, Kang L, Kim M, Kim S, Kojima T, Kuzuya M, Lee JSW, Lee SY, Lee WJ, Lee Y, Liang CK, Lim JY, Lim WS, Peng LN, Sugimoto K, Tanaka T, Won CW, Yamada M, Zhang T, Akishita M, Arai H. Asian Working Group for Sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc 2020;21:300-307. e2 , DOI : DOI:https://doi.org/10.1016/j.jamda.2019.12.012

[12]WON, Chang Won. Diagnosis of sarcopenia in primary health care. J. Korean Med. Assoc, 2020, 63: 633-641. DOI : http://dx.doi.org/10.5124/jkma.2020.63.10.633

[13]S. Cessie, J. C. Houwelingen, Ridge Estimators in Logistic Regression, Journal of the Royal Statistical Society Series C: Applied Statistics,

Volume 41, Issue 1, March 1992, Pages 191 - 201,
https://doi.org/10.2307/2347628

※ 이 논문은 정부(과학기술정보통신부, 교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No.2022R1F1A1066017,
NRF-2022S1A5C2A07090938)