의사결정나무 기반 비만과 대사증후군 데이터 분류와 특징 중요도 분석

이종욱*, 김영호*, 백병현**, 황두성*
*단국대학교 SW소프트웨어학과
**(주)지아이비타
32143603@dankook.ac.kr, dudgh1002@naver.com,
byunghyun.baek@gi-vita.io, dshwang@dankook.ac.kr

Decision tree based obesity and metabolic syndrome data classification and feature importance analysis

Jongwook Lee*, Youngho Kim*, Byunghyun Baek**, Doosung Hwang*
*Dept. of Software Science, Dankook University
**GI VITA Inc.

요 약

비만은 다양한 합병증을 일으키는 위험요소로 현대인의 건강을 위협한다. 비만에 영향을 주는 요소들은 유전적 영향, 식습관, 신체활동 등이 연관된다. 비만 인구의 증가로 대사증후군의 발병률이 높아졌다. 대사증후군은 비만, 고지혈증과 고혈압 등의 여러 가지 성인병을 동반한다. 비만과 대사증후군 판별 요소 검출을 위한 개인의 신체 정보와 생활 정보 분석이 필요하다. 본 논문에서는 의사결정나무를 이용하여 비만과 대사증후군을 분류하고 분류 시 사용된 특징의 중요도 분석을 수행한다. 비만 분석 결과는 체중과 신장이 높은 특징 중요도를 나타냈으며 대사증후군은 HDL, 허리둘레, 혈압과 나이 등이 높은 특징 중요도를 보였다.

1. 서론

헬스케어 분야는 기계학습과 딥러닝을 이용하여 의료 기록과 환자 데이터에 대한 분석 연구가 활발히 진행되고 있다. 폐렴[1], 유방암, 당뇨병, 심장병[2] 등의 연구가 대표적이다. 보편적인 의료 서비스에 대한 연구와 더불어 식습관, 체중 변화, 수면시간과 신체활동 등 개인의 일상생활 데이터 분석을 통한 연구도 진행되고 있다[3].

비만은 중요한 연구 분야로 체내에 과도한 지방의 축적으로 발생하며 전 연령에서 발생이 증가하고 있다[4]. 비만은 제 2형 당뇨병, 관상동맥질환과 호흡기 합병증 등을 일으켜 현대인의 건강을 위협한다. 비만의 발생과 관련된 요소는 유전, 에너지 소비와 에너지 섭취 등이다. 비만은 복합적인 원인에 의해 발병하며 원인을 파악하는 것이 필요하다.

대사증후군은 고혈압, 고지혈증, 당뇨와 복부 비만 등의 위험한 성인병들이 동시다발적으로 나타나는 현상이며 인슐린 저항성이 주요한 발병 원인이다[1]. 인슐린 저항성은 혈당 조절 호르몬인 인슐린이 체내에서 정상적으로 작용하지 못하는 상태이다[2]. 대사증후군은 비만 발생률과 관련이 있으며 해마다 증가하는 추세이다[5, 6].

본 논문에서는 의사결정나무를 이용하여 대사증후군과 비만을 분류하고 특징 중요도를 분석한다. 2절에서는 기계학습을 통한 대사증후군과 비만 분류 선행 연구를 소개하고 3절에서는 의사결정나무를 설명한다. 4절에서는 실험 방법과 결과를 기술하고 마지막 절에서는 결론을 서술한다.

2. 관련 연구

Cervantes, R. C.외 1명은 UCI 공개 비만 데이터셋[3]에서 18-25세인 데이터를 선별하여 비만 여부를 의사결정나무와 SVM으로 분류했다. 의사결정나무는 97.2%, SVM은 64.0%의 정확도를 나타냈다. k-means를 통해 클래스를 재정의한 데이터셋에 대한 의사결정나무의 분류 성능은 98.5%의 정확도를 보였다[7].

1) www.ydp.go.kr/health/index.do
2) www.amc.seoul.kr/asan/healthinfo/disease/diseaseSubmain.do
3) Estimation of obesity levels based on eating habits and physical condition dataset

Cui, Tong외 4명은 Cervantes, R. C.외 1명이 사용한 데이터셋을 이용했다. 체중과 신장을 제외한 비만에 영향을 주는 특징의 통계적 분석과 기계학습을 사용하여 비만을 분류했다. 특징의 상관관계와 비만 클래스별 특징 값의 분포를 분석했다. 비만 분류 실험은 LDA와 PCA를 통해 생성된 특징을 학습에 사용했다. 트리 기반 앙상블 모델은 85.9%의 정확도를 나타냈고 k-NN, SVM, 의사결정나무 등의 모델들을 집계한 정확도는 86.3%를 보였다[8].

Karimi-Alavijeh외 3명은 Isfahan Cohort Study의 대사증후군 데이터셋을 이용했다. 데이터셋은 클래스 불균형이 존재하여 SMOTE 방식으로 오버샘플링했다. 의사결정나무와 SVM으로 대사증후군 분류 실험을 진행했다. 의사결정나무는 75%의 SVM은 76%의 정확도로 대사증후군을 분류했다[9].

3. 의사결정나무

3.1. 분류 알고리즘

의사결정나무는 입력 값으로 수치형(numeric)데이터와 범주형(categorical)데이터가 가능하다. 모델은 이진 트리구조이며 학습 과정에서 부모 노드로부터 분기하여 자식 노드를 재귀적으로 생성한다. 분기는 부모 노드가 갖는 데이터 집합의 불순도(impurity)가 자식 노드에서 감소하도록 진행된다. 데이터 집합의 불순도는 엔트로피(entropy)를 통해 계산하며 식1로 정의된다. D는 데이터 집합이며 k는 집합이 가지는 클래스이다.

$$E(D) = -\sum_{k} p_k \log_2(p_k), k = 1,2,\cdots,k \quad (1)$$

데이터 집합이 가지는 특징 중에서 어떤 하나의 특징 값을 이용하여 데이터를 분류할 때 감소하는 엔트로피 값인 정보 이득량(information gain)은 식2로 정의된다. A는 데이터 집합이 가지는 특징이고 v는 특징 A의 값이며 D_v는 데이터 집합 D내 v값을 갖는 부분 집합이다.

$$IG(D,A) = E(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} E(D_v) \quad (2)$$

주어진 데이터를 이용하여 형성된 트리는 과적합(overfitting)문제를 가지고 있다. 의사결정나무는 가지치기(pruning)를 사용하여 해결하며 가지치기의 종류는 사전 가지치기와 사후 가지치기가 있다.

사전 가지치기는 트리의 깊이나 트리의 노드가 가지는 최소 데이터 수를 조절하여 과적합을 감소시킨다. 사후 가지치기는 다양한 알고리즘이 존재하지만 비용 복잡도 가지치기(cost complexity pruning)를 사용했다. 비용 복잡도 가지치기는 부모 노드와 자식 노드의 오분류율이 동일한 노드를 제거하여 루트 노드만 존재할 때까지 반복한다. 여러 개의 가지치기된 트리들이 생성되며 테스트 데이터에 대해 가장 좋은 성능을 보이는 서브 트리를 선택하여 과적합을 감소시킨다.

3.2. 특징 중요도

트리의 형성이 완료되면 트리를 구성할 때 사용된 특징에 대한 중요도를 추출할 수 있다. 특징 중요도는 불순도를 낮추는데 기여한 정도를 나타내어 데이터를 분류할 때 중요하게 작용한 특징을 찾아낼 수 있다. 감소한 불순도를 계산할 수 있는 식2를 이용하여 식3이 정의된다. 식3의 $FI(f_i)$는 데이터셋의 i번째 특징에 대한 중요도를 나타내고 $D(j)$와 $A(j)$는 노드 j가 가지는 데이터 집합과 특징이다.

$$FI(f_i) = \frac{\sum_{j} IG(D(j), A(j))}{\sum_{k} IG(D(k), A(k))} \quad (3)$$

4. 실험

4.1. 데이터

비만 분류 실험에 사용된 UCI 공개 비만 데이터셋은 남미 지역의 사람들로부터 설문을 통해 수집했다. 데이터의 77%는 SMOTE방식으로 오버샘플링했으며 16개의 특징과 2,111개의 데이터로 구성된다. 특징 집합은 기초정보 4가지, 식습관 6가지와 신체활동 4가지, 가족력과 흡연여부이다. 클래스는 체질량지수(BMI)를 세계보건기구(WHO)가 정의한 비만의 정도와 종류이며 총 7개이다[10].

대사증후군 분류 실험은 질병관리청 한국인유전체사업[4]의 대사증후군 데이터셋을 사용한다. 전체 데이터 수는 31,070개이며 나이, 성별, 혈액검사결과 등을 포함하는 9개 특징과 대사증후군 발병 여부를 나타내는 2개 클래스가 있다.

4.2. 실험 방법

학습 데이터와 테스트 데이터 비율은 전체 데이터의 80%와 20%로 설정했다. 비만 정도를 큰 단위로 재구성하여 분류 성능과 특징 중요도의 차이를 비교하기 위해 3가지 비만 분류 실험을 구성했다. 첫 번째 실험은 비만 여부를 파악하기 위한 이진 분류 문제 T_1을 정의했다. 분류 문제 T_2는 정상, 과체

4) https://kdca.go.kr/contents.es?mid=a40504020100

중과 비만을 구분하여 3개 클래스 분류 문제 T_2를 정의했다. 마지막 실험은 7개 클래스 분류 문제 T_3를 구성했다(표 1).

표 1. 비만 분류 문제별 데이터셋 구성

Problem			No. of data
T_1	T_2	T_3	
Normal	Normal	Underweight	351
		Normal	324
	Overweight	Level I	297
		Level II	290
Obesity	Obesity	Type I	290
		Type II	287
		Type III	272
Total			2,111

대사증후군 데이터셋은 클래스 불균형이 존재하여 클래스 비율 변화에 따른 분류 성능과 특징 중요도 비교를 위한 3가지 분류 문제를 구성했다. 학습 데이터와 테스트 데이터 비율은 80%와 20%로 설정했다. 정상 데이터는 대사증후군인 데이터 수의 정수 배만큼 랜덤하게 선택하여 추출했다(표 2).

표 2. 대사증후군 분류 문제별 데이터셋 구성

Problem / State	T_4	T_5	T_6
Normal	3,640	7,280	10,920
Metabolic syndrome	3,640	3,640	3,640
Total	7,280	10,920	14,560

4.3. 분류 성능

분류 문제 T_1는 99.2%의 정확도를 보여주고 T_2는 95.4%, T_3는 93.2%의 정확도를 나타냈다. 클래스 수가 증가하면서 유사한 데이터들 간의 오분류율이 증가하여 정확도가 최대 6%까지 감소했다. 사후 가지치기 적용 결과는 분류 성능이 최대 2.5%까지 높아졌다(표 3).

표 3. 분류 문제 T_1, T_2, T_3 테스트 결과

Problem / Metrics	Before			After		
	T_1	T_2	T_3	T_1	T_2	T_3
Accuracy	99.2	95.4	93.2	99.4	97.9	99.5
F1-score	99.2	95.1	92.6	99.4	97.5	93.8
Precision	99.2	95.0	92.7	99.4	97.7	94.0
Recall	99.2	95.2	92.6	99.4	97.3	93.6

분류 문제 T_4는 99.0%의 정확도를 보였다. 클래스 불균형 나타나는 T_5와 T_6는 클래스 비율이 동일한 T_4와 정확도가 동일하거나 증가했다(표4). 가지치기 후 최대 0.1%의 분류 성능이 증가했다.

표 4. 분류 문제 T_4, T_5, T_6 테스트 결과

Problem / Metrics	Before			After		
	T_4	T_5	T_6	T_4	T_5	T_6
Accuracy	99.0	99.0	99.4	99.0	99.1	99.5
F1-score	99.0	99.0	99.4	99.0	99.1	99.5
Precision	99.0	99.0	99.4	99.0	99.1	99.5
Recall	99.0	99.0	99.4	99.0	99.1	99.5

4.4. 특징 중요도

T_1에서 신장은 32.7%, 체중 20.8%, 전자기기 사용 정도 12.2% 등이 높은 중요도를 나타냈다(그림 1). T_2의 특징 중요도는 신장 35.4%, 나이 22.4%, 체중 18.0% 등으로 나타났으며 T_1에 비해 나이의 중요도가 높아지고 체중의 중요도는 감소했다(그림 2). T_3는 T_2의 특징 중요도 항목과 동일하지만 항목의 값들은 변경되었다(그림 3). 분류 문제 T_4, T_5와 T_6는 HDL, 혈압(DBP, SBP), 허리둘레(WAIST)가 높은 특징 중요도를 보였다(그림 4, 5, 6). 성별, 혈중중성지방량(TG)과 나이는 낮은 중요도를 보였으며 체질량지수는 가장 낮은 중요도를 보였다.

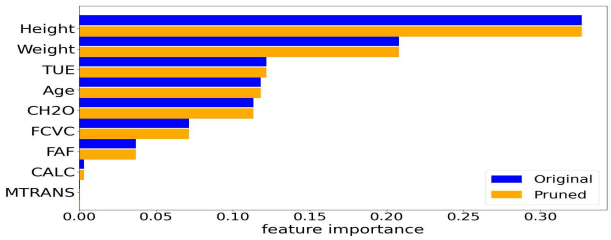

그림 1. 분류 문제 T_1 특징 중요도

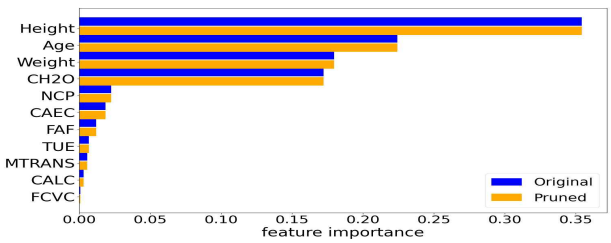

그림 2. 분류 문제 T_2 특징중요도

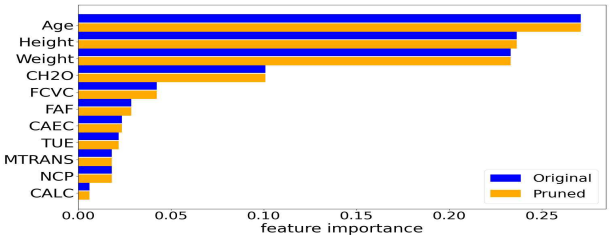

그림 3. 분류 문제 T_3 특징 중요도

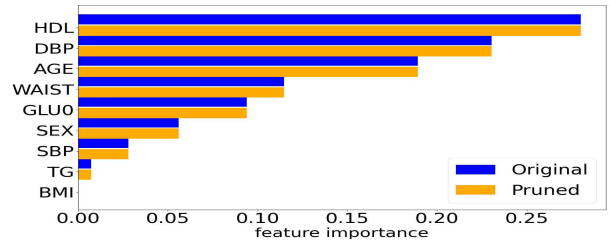

그림 4. 분류 문제 T_4 특징 중요도

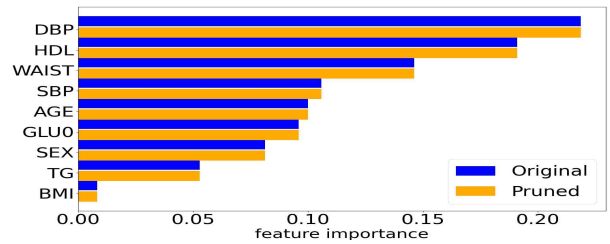

그림 5. 분류 문제 T_5 특징 중요도

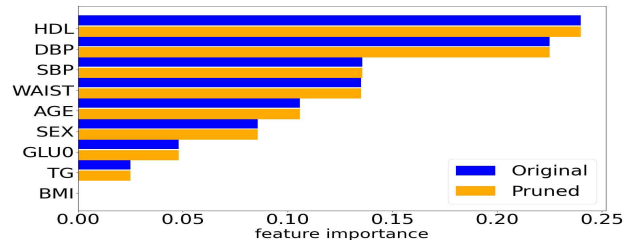

그림 6. 분류 문제 T_6 특징 중요도

5. 결 론

의사결정나무를 이용한 3가지 비만 분류 실험은 이진 분류 99.2%, 3개 클래스 분류 95.4%, 7개 클래스 분류 93.2%의 정확도를 보였다. 가지치기를 적용 후 과적합이 감소하였고 분류 성능은 증가했다. 성별을 제외한 기초 정보는 높은 중요도를 보였으며 식습관과 신체활동은 상대적으로 낮은 중요도를 보였다. 대사증후군 분류 실험은 클래스 불균형과 관계없이 약 99.0%의 정확도를 나타냈다. 대사증후군 분류 시 HDL, 혈압, 허리둘레가 높은 중요도를 보였으며 나이와 혈중 중성지방량은 비교적 낮은 중요도를 보였다.

실험에 사용한 데이터셋은 식습관과 신체활동에 대해 빈도로 표현하여 개인이 섭취한 음식의 종류와 양, 신체 활동량 정보가 부족하다. 개인의 식습관과 신체 활동량에 대해 자세한 정보를 갖는 데이터 셋을 분석할 수 있다면 비만과 대사증후군에 영향을 주는 요인을 정밀하게 분석할 수 있을 것이다. 비만과 대사증후군을 유발하는 요소를 찾기 위해 개인의 칼로리 섭취와 소비에 대한 데이터를 수집하고 요소별 특징 중요도를 이용하여 분석한다면 보다 정밀한 결과를 도출할 수 있을 것이다.

참 고 문 헌

[1] Cooper, Gregory F., et al. "An evaluation of machine learning methods for predicting pneumonia mortality." Artificial intelligence in medicine 9.2 (1997): 107-138.

[2] Kohli, Pahulpreet Singh, and Shriya Arora. "Application of machine learning in disease prediction." 2018 4th International conference on computing communication and automation (ICCCA). IEEE, 2018.

[3] Wei, Zhao, et al "Wireless sensor networks for in-home healthcare: Issues, trend and prospect." Proceedings of 2011 International Conference on Computer Science and Network Technology. Vol. 2. IEEE, 2011.

[4] Kopelman, Peter G. "Obesity as a medical problem." Nature 404.6778 (2000): 635-643.

[5] Eckel, Robert H., et al. "The metabolic syndrome." The lancet 365.9468 (2005): 1415-1428.

[6] Lee, Seung Eun, et al. "Trends in the prevalence of metabolic syndrome and its components in South Korea: Findings from the Korean National Health Insurance Service Database (2009 - 2013)." PloS one 13.3 (2018): e0194490.

[7] Cervantes, et al. "Estimation of obesity levels based on computational intelligence." Informatics in Medicine Unlocked 21 (2020): 100472.

[8] Cui, Tong, et al. "Estimation of Obesity Levels Based on Decision Trees." 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM). IEEE, 2021.

[9] Karimi-Alavijeh, et al. "Predicting metabolic syndrome using decision tree and support vector machine methods." ARYA atherosclerosis 12.3 (2016): 146.

[10] Palechor, et al. "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico." Data in brief 25 (2019): 104344.