베이지안 네트워크와 특이값 분해 알고리즘을 이용한 운동 추천 시스템

신아영, 임유진
숙명여자대학교 IT공학과
ayooung7@sookmyung.ac.kr, yujin91@sookmyung.ac.kr

An exercise recommendation system using bayesian network and singular value decomposition algorithm

A-Young Shin, Yujin Lim
Dept. of IT engineering, Sookmyung Women's University

요 약

본 논문에서는 코로나-19로 인해 홈 트레이닝 시장이 성장하고 있는 상황 속에서 효율적인 운동을 위해 사용자의 식습관, 신체조건, 선호도 등을 바탕으로 적합한 운동을 추천해주는 시스템을 제안한다. 먼저 K-최근접 이웃 알고리즘을 활용해 비만의 정도에 따라 사용자를 분류하고, 운동 데이터를 소모 칼로리에 따라 클러스터링 한다. 다음으로 비만의 정도와 운동 레벨에 따라 정해진 추천 점수를 통해 사전 선호도 확률을 계산하고, 베이지안 네트워크를 통해 사후 확률을 구한다. 이를 바탕으로 특이값 분해 알고리즘(SVD)를 활용하여 사용자 맞춤형 운동을 추천한다. 제안 시스템의 성능을 검증하기 위해 비교 실험을 진행하여 회귀 문제 평가 척도인 RMSE 값 측면에서 성능을 분석하였다.

1. 서론

최근 코로나-19의 영향으로 건강관리에 관심을 가지는 사람들이 늘어났다. 하지만, 코로나-19의 장기화로 실내 체육시설 이용이 어려워진 상황 속에서 홈 트레이닝 시장이 빠르게 성장하고 있다. 이에 사람들은 자신의 신체조건 등을 고려한 효율적인 운동 추천을 필요로 하고 있다. 따라서 개인의 선호도와 조건들을 고려한 개인화 추천 시스템의 중요성이 대두되고 있다.

본 논문은 홈 트레이닝을 하는 사용자들의 성향 정보와 선호도를 고려한 개인 맞춤화 운동추천 시스템을 제안한다. 비만 데이터 셋을 활용하여 사용자의 비만 정보, 신체정보 등을 토대로 사용자를 분류하고, 운동 데이터 셋 또한 운동의 소모 칼로리를 기준으로 데이터를 클러스터링 한다. 해당 그룹의 선호도를 바탕으로 베이지안 네트워크를 활용한 가중치 선호도 정보와 특이값 분해 알고리즘(SVD)을 사용하여 추천 시스템을 구현한다. 해당 시스템의 추천이 제대로 제공되는지 확인하기 위해 평균 제곱근 오차인 RMSE 값을 기반으로 다른 추천 방법과 비교 실험을 통해 성능을 검증한다.

2. 관련 연구

2.1 k-최근접 이웃 알고리즘 (K-Nearest Neighbor)

k-최근접 이웃 알고리즘은 가장 가까운 이웃을 찾아 새로운 사용자에 대한 예측 및 분류 작업을 하는데 사용되는 지도학습 기법 중 하나이다. 새로운 사용자에 대해 전체 사용자 데이터로부터 가장 가까운 k개의 이웃을 선택하여 근접 정도에 따른 가중치 평균으로 분류 또는 예측 값을 계산한다[1]. 본 논문에서는 k-최근접 이웃 알고리즘을 사용하여 사용자의 비만의 정도를 결정하고 분류하였다.

2.2 베이지안 네트워크 (Bayesian network)

베이지안 네트워크란 사전에 일어난 일을 바탕으로 사후의 확률을 추론하는 방법[2]으로, 본 논문에서는 사용자가 선택한 운동의 이력을 분석하여 다음에 무엇을 선택할지 예측하는 데 사용하였다.

2.3 특이값 분해 알고리즘 (SVD, Singular Value Decomposition)

특이값 분해 알고리즘은 하나의 행렬을 여러 개의 행렬의 곱으로 분해하는 방법이다. 모든 사용자와 운동에 대한 $m \times n$ 크기의 행렬 M을 특이값 분

해하면 $M = U \sum V^T$ 와 같이 세 가지 행렬의 곱으로 나타낼 수 있다. 이때, $U_{m \times m}$은 사용자 행렬을 나타내고, $\sum_{m \times n}$은 특이값을 대각항으로 가지는 대각 행렬, ${V^T}_{n \times n}$는 운동 행렬을 나타낸다[3]. 이러한 SVD는 고차원의 행렬을 저차원의 행렬로 축소시킴으로써 분석의 정확성과 계산 속도를 향상시켜 실시간 추천을 할 수 있다[4]. 본 논문에서는 특이값 분해 알고리즘을 사용하여 사용자에게 적합한 운동을 실시간으로 추천해주는데 사용하였다.

3. 베이지안 네트워크와 특이값 분해 알고리즘을 이용한 운동 추천 시스템

추천 시스템을 위한 데이터 셋은 비만 데이터 셋으로 'Obesity based on eating habits & physical cond.'[5]를 사용하였고 운동 데이터 셋으로는 'Calories burned during exercise and activities'[6]을 사용하였다.

Obesity dataset은 컬럼비아, 페루, 멕시코 인들의 식습관, 신체조건, 비만 정도 등의 데이터를 제공한다. 비만 정도를 나타내는 컬럼인 NObeyesdad는 저체중, 정상 체중, 과체중 1-2 레벨, 비만 1-3 타입으로 7개의 값을 가진다. 시스템의 결과 성능을 향상시키기 위해 이를 저체중, 정상, 과체중, 비만 4가지의 값을 가지도록 재분류하였다. 또한, 저체중을 1, 정상을 2, 과체중을 3, 비만을 4로 각각 숫자 형식으로 변환해 적용하였다[1].

그리고 비만의 정도를 나타내는 컬럼과 다른 컬럼들 간의 상관관계를 확인하기 위해 선형 회귀 실험을 진행하였다. 실험 결과는 <표 1>과 같다.

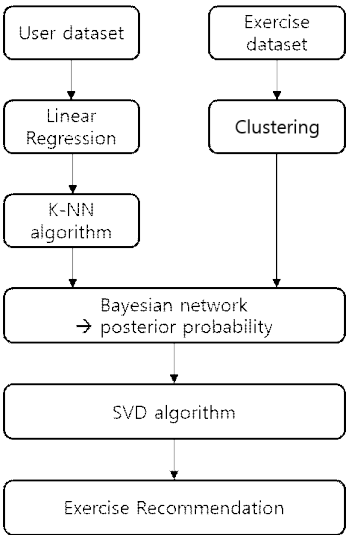
<표 1> 선형 회귀 실험 결과

feature	coefficient
Gender	0.044156
Age	0.413213
Height	0.132500
Weight	0.880412
family_history_with_overweight	0.504324
FAVC	0.275896
FCVC	0.124506
NCP	-0.120563
CAEC	0.374978
SMOKE	0.009515
CH2O	0.148771
SCC	-0.184015
FAF	-0.157407
TUE	-0.090338
CALC	-0.117531
MTRANS	-0.057896
Nobeyesdad	1.000000

선형 회귀 실험 결과에 따라 상관 계수가 큰 값을 갖는 Weight, family_history, Age, CAEC(식사 사이 음식 섭취), FAVC(고열량 식품 섭취 빈도) 총 다섯 개의 컬럼으로 데이터 셋을 재구성하였다.

재구성한 데이터셋이 비만의 정도를 잘 결정하는지 확인해보기 위해 k-최근접 이웃 실험을 진행하였다. k 값이 4 일 때 약 91.53%의 정확도를 얻을 수 있었고, 재구성한 데이터셋이 비만의 정도에 따라 사용자를 잘 분류할 수 있음을 확인할 수 있었다.

운동 데이터 셋에는 운동의 레벨을 나타내는 컬럼을 추가하고, kg당 소모 칼로리를 기준으로 1-4의 레벨로 클러스터링 하였다. 운동의 소모 칼로리를 기반으로 클러스터링하기 위해 k-means 기법을 적용한다.



(그림 1) 제안하는 시스템의 구조.

제안하는 시스템의 전반적의 구조는 그림 1과 같다. 재구성한 데이터 셋에 베이지안 네트워크를 통해 사용자의 선택으로 변화된 사후 데이터를 반영하여 운동에 대한 선호도 확률을 계산한다. 이 선호도 확률을 이용해 User-Item matrix를 구성하고 특이값 분해 알고리즘을 적용하여 운동을 추천한다.

먼저, 재구성 된 사용자 데이터셋과 운동 데이터 셋을 이용하여 각 사용자에 대해 모든 운동의 1차 추천 점수를 계산한다. 1차 추천 점수는 사용자의 비만 정도를 나타내는 값과 운동 레벨의 차의 절댓값을 추천 점수의 최대 값인 4에서 빼 구한 값으로 한다[1]. 그리고 베이지안 네트워크를 통해 사후 선호도 확률을 계산한다.

예를 들어, 사용자가 운동 추천 점수 1-4에 따른 운동 종류 $\{R_1, R_2, R_3, R_4\}$에 대한 선택 횟수가 {1, 1, 3, 5}로 사용자의 총 선택 횟수는 10회 일 때, 1차 추천 점수로 제공된 추천을 통해 사용자가 추천 점수가 4점인 R_4 운동을 5회 추가로 선택한다고 하면, R_4 운동에 대한 사전 데이터는 5, 사후 데이터는 10이 되고, 전체 $\{R_1, R_2, R_3, R_4\}$ 운동 선택 횟수는 총 15회가 된다. 이처럼 R_4 가 선택되었을 때 각 운동 R_i에 대한 선호도 사후 확률은 식 (1)과 같이 계산된다[2].

$$P(R_i \mid X = R_4) = \frac{P(R_i)P(X = R_4 \mid P(R_i))}{\sum_{j=1}^{4} P(R_j)P(X = R_4 \mid R_j)} \quad (1)$$

예를 들어 R_3에 대한 선호도의 사후 확률은 식(2)와 같다.

$$P(R_3 \mid X = R_4) = \frac{\frac{3}{10} \times \frac{3}{15}}{\frac{1}{10} \times \frac{1}{15} + \frac{1}{10} \times \frac{1}{15} + \frac{3}{10} \times \frac{3}{15} + \frac{5}{10} \times \frac{10}{15}} \cong 0.15 \quad (2)$$

따라서 특정 운동의 선택에 의한 사전 선호도와 사후 선호도의 확률 값을 비교하면 <표 2>와 같다.

<표 2> 사전 선호도와 사후 선호도 비교

Rating / Users	R_1	R_2	R_3	R_4
사전	$\frac{1}{10}$ = 0.1	$\frac{1}{10}$ = 0.1	$\frac{3}{10}$ = 0.3	$\frac{5}{10}$ = 0.5
사후	0.02	0.02	0.15	0.82

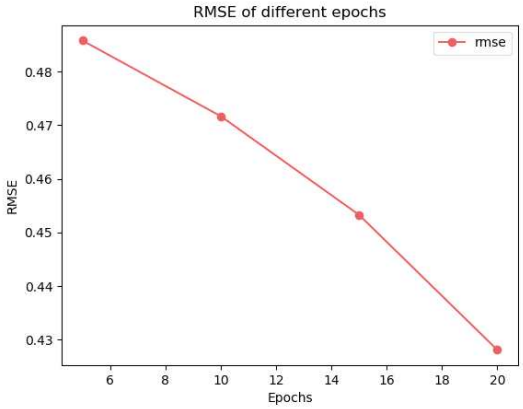
이렇게 구해진 사후 선호도를 이용해 User-Item matrix를 구성한다. 그리고 이 행렬에 특이값 분해 알고리즘을 적용하여 사용자에게 운동 추천을 제공한다. 다음 장에서 해당 시스템의 성능을 검증해보고자 한다.

4. 실험 및 성능 평가

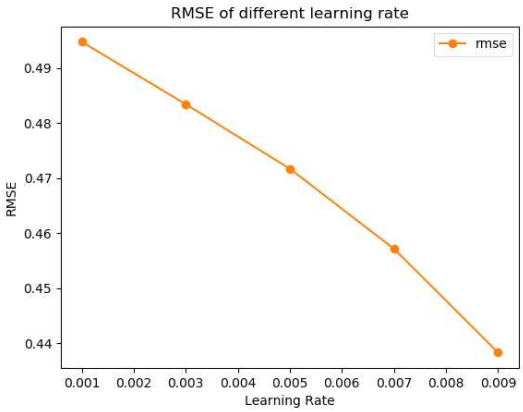
추천 시스템의 성능을 알아보기 위해 정확도, 재현율 등을 값을 계산해 보고, 다른 추천 기법과 비교해보고자 한다.

본 논문에서 제안하는 추천 시스템의 실험을 위해 Python Surprise 라이브러리를 활용하여 특이값 분해 알고리즘을 구현하였다[1]. 제안하는 시스템은 운동에 대한 사후 선호도 확률이 높을수록 높은 비율로 추천을 제공한다. 이 실험을 위해 User-Item matrix를 train-set 70%, test-set 30%로 나눈 후 특이값 분해 알고리즘을 통해 학습시키고 예측 결과를 살펴보았다.

먼저, 특이값 분해 알고리즘에 사용되는 epoch 수와 learning rate 값에 따른 평균 제곱근 오차인 RMSE 값을 비교해 보았다. 실험 결과는 그림 2, 그림 3과 같다.



(그림 2) epoch 수에 따른 RMSE 값



(그림 3) learning rate에 따른 RMSE 값

마지막으로, 본 논문과 같이 차원 축소 모델인 특이값 분해 알고리즘을 사용한 모델과 제안하는 기법을 비교해보았다. 개인 성향을 고려한 운동 추천 연구 중 k-최근접 이웃 알고리즘을 통해 데이터를 분류한 후, 모델 기반 matrix factorization 을 통해 운동을 추천하는 방법(k-nn - SVD) [1] 과 아이템 기반 협업 필터링을 기반으로 특이값 분해 알고리즘을 적용한 추천 방법(IBCF-SVD)[7], 그리고 본 논문에서 제안하는 추천 기법을 비교하는 실험을 진행하였다. 이 실험 또한 RMSE 값을 사용해 추천 기법의 성능을 비교하였고 결과는 <표 3> 과 같다.

<표 3> 다른 모델과의 RMSE 값 비교

System	RMSE
Proposed System	0.4281
K-NN – SVD [1]	0.6044
IBCF – SVD [7]	0.8199

세 방법 모두 특이값 분해 알고리즘을 이용해 추천 시스템을 구현하였지만, 본 논문에서 제안하는 시스템이 [1]의 방법보다 약 18% 만큼 더 나은 성능을 보여주며, [7]의 방법보다는 약 39% 만큼 더 나은 성능을 보여주고 있음을 확인할 수 있다. 제안하는 방법은 선형 회귀 실험 결과에 따라 상관 계수가 큰 컬럼으로 데이터 셋을 재구성하였고, 베이지안 네트워크를 통해 사용자의 사후 선호도에 따라 추천을 제공함으로써 보다 적절한 추천 정보를 제공한다.

5. 결론

본 논문에서는 홈 트레이닝을 하는 사용자들의 성향 정보와 사후 선호도를 고려하여 효율적인 운동 추천을 제공하고자 한다. 사용자의 비만 정보, 신체 정보 등을 토대로 비만 정도를 분류하고, 운동 데이터셋 또한 운동의 소모 칼로리를 기준으로 k-means 클러스터링 기법을 적용하였다.

사용자의 운동 패턴은 여러 요인에 따라 달라질 수 있기 때문에 선호도 계산에 있어 모든 운동에 대해 동일한 가중치를 부여하는 것은 비현실적이라는 문제를 해결하고자 베이지안 네트워크를 통해 사용자의 선택에 따라 사후 선호도를 유연하게 계산함으로써 적절한 추천이 이루어지도록 하였고, 특이값 분해 알고리즘을 통해 실시간 추천을 가능하게 하였다.

성능 평가를 위해 다른 추천 방법과의 비교 실험을 진행하였고 본 논문에서 제안하는 시스템이 기존 방법보다 우수한 성능을 나타냄을 확인하였다.

향후 연구에서는 사용자의 효과적인 건강관리를 위해 운동의 레벨을 결정함에 있어 칼로리뿐만 아니라 다른 요인들도 함께 고려하여 추천할 수 있도록 확장하고자 한다.

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1047113).

참고문헌

[1] H.-Y. Lee and O.-R. Jeong "A personalized exercise recommendation system using dimension reduction algorithms," Journal of the Korea Society of Computer and Information, vol. 26, no. 6, pp. 19-28, 2021.

[2] W.-B. Park, Y.-S. Cho, and H.-H. Ko "Clustering method of weighted preference using k-means algorithm and bayesian network for recommender system," Journal of Information Technology Applications & Management, vol. 20, no. 3, pp. 219-230, 2013.

[3] J.-Y. Hyun, S.-Y. Ryu, and S.-Y. Lee, "How to improve the accuracy of recommendation systems: Combining ratings and review texts sentiment scores," Journal of Intelligence and Information Systems, vol. 25, no. 1, pp. 219-239, 2019.

[4] S.-Y. Jeong and H.-J. Kim, "A recommender system using factorization machine," Journal of Digital Contents Society, vol. 18, no. 4, pp. 707-712, 2017.

[5] Kaggle, "Obesity based on eating habits & physical cond.", https://www.kaggle.com/ankurbajaj9/obesity-levels

[6] Kaggle, "Calories Burned During Exercise and Activities", https://www.kaggle.com/aadhavvignesh/calories-burned-during-exercise-and-activities

[7] D.-W. Kim, S.-G. Kim, and J.-Y. Kang, "An empirical study on hybrid recommendation system using movie lens data," Korea Bigdata Society, vol.2, no.1, pp.41-48, 2017