Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 1, pp. 28-36, February 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.1.028

조건부 변분오토인코더 및 협업필터링을 활용한 복지 프로그램 추천

김성은[1] · 지민기[2] · 문일철[3] · 주원영[4†]

[1]한국과학기술원 데이터사이언스대학원 / [2]구글 코리아
[3]한국과학기술원 산업및시스템공학과 / [4]이화여자대학교 통계학과

Welfare Program Recommendation by Conditional Variational Autoencoder and Collaborative Filtering

Sungeun Kim[1] · Mingi Ji[2] · Il-Chul Moon[3] · Weonyoung Joo[4]

[1]Graduate School of Data Science, Korea Advanced Institute of Science and Technology
[2]Google Korea
[3]Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology
[4]Department of Statistics, EWHA Womans University

Recently, the government of South Korea has offered a variety of welfare programs that are customized to diverse demands, such as diabetes management, alcohol addiction rehabilitation, living condition improvement, etc. These welfare programs have become too diverse to be remembered and recommended by individuals, and the government now has a list matching program recipients and programs for further studies. This research investigates such welfare program recommendation with a conditional variational autoencoder merged with collaborative filtering, a.k.a. CVAE-CF. We use a natural language description to provide the program information, or item in the context, and we utilize the demographic information from potential recipients as the user information. Our results show agreeable performance for future application to recommendation tasks showing 63% recall and 13.1% precision on average.

Keywords: Welfare Program Recommendation, Collaborative Filtering, Variational Autoencoder

1. 개 요

최근 한국 정부가 다양한 복지 프로그램을 제공함에 따라 수급자들에게 적합한 복지 프로그램을 추천하는 것이 중요해지고 있다(Jung *et al.*, 2017). 복지 프로그램은 지역 주민의 건강 및 삶의 질 상승을 위해 각 구의 보건소를 통해 운영되고 있으며, 금연/금주 프로그램, 만성질환 관리 프로그램 등 신체의 건강 상태를 관리해주는 프로그램부터 아이 돌봄 서비스, 말벗 지원과 같이 전반적인 삶의 질 유지를 위한 프로그램까지 총 59개의 복지 프로그램이 제공된다(MOHW, 2022).

시간에 따라 증가하는 복지 프로그램을 수급대상자에 맞춤화하여 제공하기 위하여, 공중 보건 정보 시스템(PHIS, Public Health Information System)은 이러한 복지 프로그램을 지원하고, 수급자가 복지 프로그램을 제공받은 이력 데이터와 수급자의 나이, 성별과 같은 기본 인적 정보 및 건강 정보를 수집한다(MOHW, 2021). 이 중, 건강 정보는 일부 신체 계측 정보(혈

이 논문은 보건산업진흥원 국민건강 스마트관리 연구개발사업(HS20C005202)의 지원을 받아 수행되었음.
† 연락저자 : 주원영 교수, 03760 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과, Tel : 02-3277-2302,
E-mail : weonyoungjoo@ewha.ac.kr
2022년 8월 4일 접수; 2022년 12월 30일 수정본 접수; 2023년 1월 9일 게재 확정.

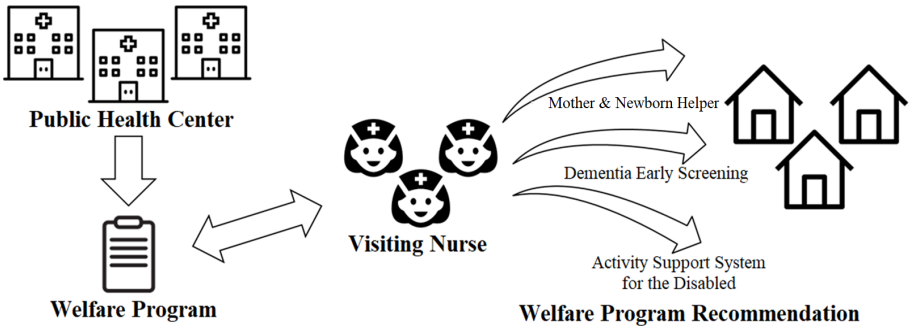


Figure 1. Welfare Program Recommendation Process

압, 혈당 등)와 건강 관련 설문 조사의 결과를 포함하고 있다.

복지 프로그램의 추천은 방문 건강관리 사업을 통해 이뤄진다. 방문 건강관리 사업은 지역 주민의 건강 수준을 향상시키는 목적으로 전국 258개(2020년 기준)(Whosaeng, 2021)의 보건소에서 운영 중이며, 각 보건소에 속한 간호사가 건강 위험요인이 큰 취약 계층을 직접 방문하여 건강 상태를 관리해주는 서비스를 시행한다(Kim *et al.*, 2022). 방문간호사는 기본 건강관리를 포함해 대상자의 특성에 따라 만성질환 증상 관리, 노인 허약 예방 관리, 임산부 건강관리, 장애인 재활관리 등을 제공하며, 방문 대상의 필요에 따라 보건소 내외에서 운영되는 복지 프로그램을 추천해줄 수 있게 된다. <Figure 1>은 현재 복지 프로그램이 방문 건강관리 사업을 통해서 전달되는 순차적 관계를 설명한다.

PHIS로부터 확보한 2018~2019년도 데이터를 기준으로, 복지 프로그램의 수급자는 약 130만 명이고, 각 복지 프로그램은 최소 8번부터 최대 약 25만 번까지 시행되었다. 구체적인 서비스 내용을 파악할 수 없는 기타 복지 프로그램을 제외하고는 고연령층을 대상으로 시행되는 치매조기검진 프로그램이 약 16만 번으로 가장 많이 시행되었는데, 같은 고연령층이어도 노인 우울증 위험이 더 크다면 치매조기검진 보다는 정신 보건 서비스가 우선하여 추천되어야 한다. 즉, 방문간호사는 방문 대상과 복지 프로그램 양측을 자세히 이해하고 적절한 복지 프로그램의 추천을 제공해야 하지만, 모든 방문간호사가 이를 제대로 파악하고 있기 어려운 실정이다. 그러한 차원에서, 본 연구에서는 추천을 제공하는 간호사의 의사결정을 돕기 위해 PHIS의 데이터와 복지 프로그램에 대한 설명을 제공하는 자연어 데이터를 활용하여 복지 프로그램 추천 모델을 개발하였다. 편의를 위해 복지 프로그램 수급자가 복지 프로그램을 제공받은 이력 정보, 복지 프로그램 수급자의 인적 및 건강 정보, 복지 프로그램 설명 정보를 각각 수급자-복지 프로그램 정보, 수급자 보조 정보, 복지 프로그램 보조 정보라고 지칭하겠다.

본 연구는 자연어를 포함한 다양한 유형의 데이터를 활용하기 위해 데이터 분포 간의 조건부 종속성을 가정하여 데이터의 잠재 표현을 추론하는 조건부 변분 오토 인코더(Sohn *et al.*, 2015)가 결합한 협업 필터링 모델(CVAE-CF, Conditional Variational Autoencoder merged with Collaborative Filtering) (Lee *et al.*, 2017)을 활용하였다. 또한, PHIS 데이터의 수급자-복지 프로그램 정보는 3개 미만의 복지 프로그램을 제공받은 수급자가 전체의 99% 이상을 차지할 정도의 큰 희소성을 가지고 있어 일반적인 추천 알고리즘과 다른 학습/평가 데이터 분할 전략을 요구하였다. 따라서, 우리는 이 문제를 고려하여 재설계 및 최적화한 CVAE-CF 모델을 검증하였다.

본 논문의 구성은 다음과 같다. 제2장에서 연구 배경 및 사전 지식을 소개한다. 제3장에서는 복지 프로그램 추천 모델 외에도 본 연구에 활용한 데이터와 데이터 희소성 문제를 다루는 학습/평가 데이터 분할 방식을 설명된다. 제4장에서는 정량적/정성적 실험 결과를 요약하였으며, 마지막으로 연구의 결론과 기대 효과를 제5장에 제시하였다.

2. 연구 배경 및 사전 지식

2.1 연구 배경

본 장에서는 연구의 배경이 된 복지 프로그램에 대해 자세히 설명하고자 한다. 보건소 내에서는 건강 상태 관리를 위한 방문/내소 진료, 금연/금주 프로그램, 만성질환 관리 프로그램 등과 더불어 노인, 산모/신생아, 영유아 등의 특정 군을 대상으로 하는 프로그램까지 총 33개의 복지 프로그램이 운영되고 있다(MOHW, 2022). 예를 들어, 금연 프로그램은 흡연자들에게 상담이나 금연 보조제를 제공하고, 금주 프로그램은 알코올 중독자들을 위한 교육을 제공한다. 보건소 외에서는 아이 돌봄 서비스, 말벗 지원, 외출 보조와 같은 복지 영역의 서비스들과 함께 치매 보건, 정신 보건 등을 위한 각종 상담 및 검사 프로그램이 총 26개 운영되고 있다(MOHW, 2022). 이렇게 보건소 내외에서 운영되는 59개의 복지 프로그램은 방문 건강관리 서비스 운영 시 방문간호사를 통해 추천된다(MOHW, 2022). 이러한 다양한 복지 프로그램의 내용을 모든 방문간호사가 자세히 파악하고 있기 어려우며, 방문간호사의 숙련도에 따라 이해 정도가 다르므로 적합한 복지 프로그램을 추천하기 어려운 실정이다. 따라서 본 연구에서는 복지 프로그램과 수급자 사이의 관계를 학습하여 방문간호사가 적절한 추천을 할 수 있도록 돕는 복지 프로그램 추천 시스템을 제안하고자 한다.

2.2 사전 지식

(1) 협업 필터링

복지 프로그램 추천 모델 설명에 앞서 기존의 추천 모델 연구를 소개하고자 한다. 협업 필터링(Collaborative Filtering) (Schafer *et al.*, 2007)은 과거 사용자와 추천 항목 사이의 관계를 분석해 새로운 사용자와 추천 항목을 매칭하는 모델이다. 예를 들면, 영화 추천의 경우 사용자가 과거에 어떤 영화를 시청했는지, 상품 추천의 경우 어떤 상품을 구매했는지와 같은 정보를 통해 사용자에게 적합한 품목을 추천한다. 본 연구에서는 수급자가 복지 프로그램을 제공받은 이력이 사용자와 추천 항목 사이의 관계에 대응된다 볼 수 있다.

모델 기반 협업 필터링(Model-based Collaborative Filtering) (Su *et al.*, 2009)은 이러한 정보로부터 사용자와 추천 항목 사이의 관계에 영향을 미치는 잠재 요인을 추론하고 그로부터 새로운 추천을 만들어낸다. 대표적으로 행렬 분해 모델(MF, Matrix Factorization) (Koren *et al.*, 2019)이 있다. N명의 사용자, M개의 추천 항목으로 구성된 사용자-항목 평가 행렬을 $R=(r_{ui})\in\mathbb{R}^{N\times M}$라고 할 때, 본 논문에서는 사용자-항목 평가 행렬의 값이 사용자와 추천 항목 사이의 관측이 있는 경우 1, 그렇지 않은 경우 0의 값을 갖는 내재적 피드백(Implicit Feedback) 경우만 고려하도록 하겠다. MF는 평가 행렬을 d차원의 사용자/항목 잠재 요인 p_u, $q_i\in\mathbb{R}^d$를 가지는 낮은 차수의 행렬로 분해한다. 잠재 요인 학습을 위한 손실 함수는 관측된 평가 $r_{ui}\in\Omega^+$와 잠재 요인 벡터 내적 $q_i^T p_u$ 간의 평균 제곱오차이며, 정규화 항까지 포함한 식은 식 (1)과 같다.

$$L=\sum_{r_{ui}\in\Omega^+}(r_{ui}-q_i^T p_u)^2+\lambda_i\|q_i\|^2+\lambda_u\|p_u\|^2 \tag{1}$$

λ_i와 λ_u는 정규화 항의 세기를 조절한다. 이 알고리즘은 추론된 잠재 요인 p_u, q_i로 평가 행렬을 재구성하여 새로운 사용자-추천 항목 쌍을 생성한다.

(2) 변분 오토 인코더의 활용

MF를 통해 비확률적인 잠재 요인을 추론할 수 있지만, 현실 데이터에는 잡음이 섞여 있기에 좀 더 강건한 모델이 필요하다. 특히 내재적 피드백은 사용자와 추천 항목 사이 관측 여부 정보만 주고, 모든 관측값이 사용자의 항목에 대한 선호를 나타내는 것은 아니므로 잡음에 대한 노출이 크다. 잠재 요인을 확률 변수로 본다면 무작위로 발생하는 잠재 요인을 기반으로 모델을 학습하기 때문에 학습 데이터에 과적합(Overfitting)되는 것을 완화하고 좀 더 강건한 모델을 만들 수 있는데, 변분 오토 인코더(VAE, Variational Autoencoder)(Kingma *et al.*, 2013; Rezende *et al.*, 2014)는 잠재 요인을 확률 변수로 보고 그 분포의 파라미터(Parameter)를 학습하는 심층 생성 모델이다. VAE의 인식 모델은 데이터 r을 잠재 변수 z로 인코딩하고, 생성 모델은 인코딩된 잠재 변수를 다시 데이터 공간으로 디코딩한다. 데이터 r을 잘 재현할 수 있는 잠재 변수 z의 분포를 추론하고 싶을 때, 손실 함수는 식 (2)와 같이 데이터 r에 대한 로그 가능도 기댓값과 잠재 변수 z의 변분 사후 분포와 사전 분포 사이의 KL Divergence 합으로 이뤄진다. 로그 가능도의 기댓값은 본래 데이터를 잘 만들어낼 수 있는 잠재 변수의 분포를 추론할 수 있도록 하고, KL Divergence 항은 잠재 변수의 분포에 대한 정규화를 걸어준다.

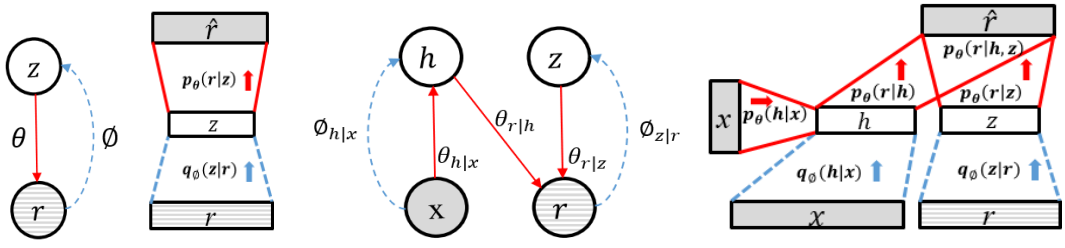
$$L=D_{KL}(q_\phi(z|r)\|p_\theta(z))-E_{q_\phi(z|r)}[\log p_\theta(r|z)] \tag{2}$$

인식 모델과 생성 모델은 θ와 ϕ를 파라미터로 하는 다층 인공 신경망(MLPs, Multilayer Perceptrons)으로 분할 상환 추론(Amortized Inference)된다.

일반적으로 VAE는 고밀도의 데이터를 고려하는데, 추천 시스템에서는 모든 가능한 사용자와 추천 항목 조합보다 관측되는 사용자와 추천 항목 사이 관계 정보가 희소하다. 이러한 데이터 희소성 문제를 해결하기 위해 변분 오토 인코더를 활용한 협업 필터링(VAE-CF, Variational Autoencoder merged wth Collaborative Filtering) (Lee *et al.*, 2017)은 negative sampling을 적용한다. 즉, 가능도를 최대화하는 과정 중, 관측된 양의 항목 $r_{ui}\in\Omega^+$외에도 관측되지 않은 값 일부를 샘플링하여 음의 항목 $r_{ui}\in\Omega^-$으로 활용한다. 사용되는 손실 함수는 식 (3)과 같다.

$$L=D_{KL}(q_\phi(z|r)\|p_\theta(z))-\sum_{r\in\Omega^+\cup\Omega^-}E_{q_\phi(z|r)}[\log p_\theta(r|z)] \tag{3}$$

VAE-CF의 graphical model과 신경망 구조는 <Figure 2(a)>와 같다. graphical model에서 흰색 노드는 잠재 변수를, 실선으로



(a) VAE-CF (b) CVAE-CF

Figure 2. Graphical Model and Neural Network Structure of VAE-CF and CVAE-CF

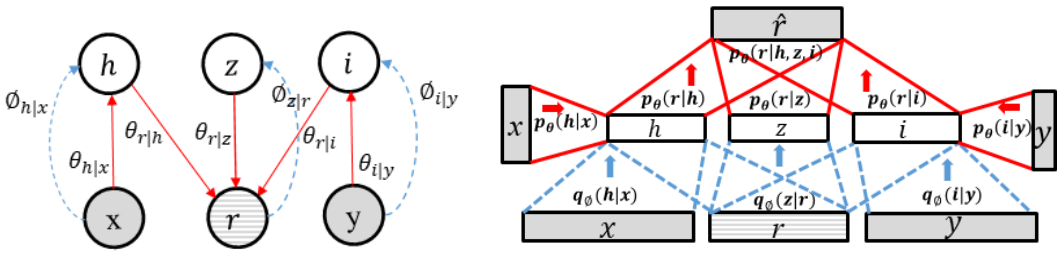


Figure 3. Graphical Model (left) and Neural Network Structure (right) of Modified CVAE-CF

채워진 노드는 희소한 관측 변수를 나타낸다. 빨간색 실선은 파라미터 θ의 생성 모델 MLP를, 파란색 점선은 파라미터 ϕ의 인식 모델 MLP를 나타낸다.

이때 사용자와 추천 항목 사이 관계 정보뿐만 아니라, 사용자와 추천 항목의 보조 정보를 같이 활용하기 위해서는 여러 데이터셋의 활용으로 인한 다중 모달리티 모델링이 필요하며, 조건부 변분 오토 인코더를 사용한 협업 필터링 (CVAE-CF, Conditional Variational Autoencoder merged with Collaborative Filtering) (Lee *et al.*, 2017)은 다중 모달리티 간의 종속 구조를 조건부 분포로 모델링 한다. 따라서, VAE-CF와 비교했을 때 보조 정보 x를 추가로 활용하기 위해 잠재 변수 h를 도입한다.

생성 모델 차원의 CVAE-CF를 통한 관계 정보 생성 과정은 다음과 같다. 1) 잠재 변수 z를 사전 확률 분포 $p(z)$에서 추출한다. 2) 보조 정보 x가 주어졌다는 가정하에 부가적인 잠재 변수 h를 조건부 확률 분포 $p_\theta(h|x)$에서 추출한다. 3) 마지막으로, 관계 정보 r을 조건부 확률 분포 $p_\theta(r|z,h)$에서 추출한다. 여기서 조건부 확률 분포 p_θ는 MLP로 이루어져 있다. 인식 모델 관점에서 CVAE-CF는 잠재 변수 z와 h를 추론하기 위해 MLP로 이루어진 분포 $q_\phi(z|r)$와 $q_\phi(h|x)$를 각각 도입한다. 추가로 활용되는 보조 정보 중 수치형 데이터는 값 자체로, 범주형 데이터는 임베딩 벡터로 변환되어 MLP의 입력값으로 활용된다. 학습 차원에서는 잠재 변수가 보조 정보 x보다 사용자와 추천 항목 사이 관계 정보 r에 더 초점을 맞출 수 있도록 정규화 항 $D_{KL}(q_\phi(h|x)\|q_\phi(h|r))$을 추가로 더해준다. 이때 손실 함수는 다음과 같다.

$$L = E_{q_\phi(z,h|r,x)}[\log p_\theta(r|z,h)] - D_{KL}(q_\phi(z|r)\|p_\theta(z)) \\ -\alpha_1 D_{KL}(q_\phi(h|x)\|p_\theta(h|x)) - \alpha_2 D_{KL}(q_\phi(h|x)\|q_\phi(h|r)) \quad (4)$$

α_1은 x와 관련된 정규화 항을 조절하고, α_2는 r과 관련된 정규화 항을 조절한다. CVAE-CF의 graphical model과 신경망 구조는 <Figure 2(b)>와 같다. 회색 노드는 관측 변수를 나타내며, 나머지 표기의 의미는 <Figure 2(a)>와 같다.

본 연구에서는 기존의 추천 모델 연구를 기반으로 복지 프로그램 추천 모델을 설계하고자 한다. 복지 프로그램 데이터의 특성을 고려하여 어떻게 모델을 설계하였는지는 다음 장에서 설명하도록 하겠다.

3. 복지 프로그램 추천 시스템

3.1 복지 프로그램 추천을 위한 조건부 변분 오토 인코더를 활용한 협업 필터링

본 연구에서 진행한 복지 프로그램 추천에서는 사용자에 해당하는 복지 프로그램 수급자와 추천 항목에 해당하는 복지 프로그램의 보조 정보를 모두 활용하기 위해 2.2.2장에 서술한 CVAE-CF에 추가 잠재 변수 i를 도입해 새로운 버전의 CVAE-CF를 제안하였다. 요약하자면, 잠재 변수 z, h, i는 각각 수급자-복지 프로그램 정보 r, 수급자 보조 정보 x, 복지 프로그램 보조 정보 y에 대응하며, 표기는 <Table 1>에 정리하였다. 최종적으로 사용하는 CVAE-CF 모델의 손실 함수는 식(5)와 같다.

$$L = E_{q_\phi}[\log p_\theta(r|z,h,i)] - D_{KL}(q_\phi(z|r)\|p_\theta(z)) \\ -\alpha_1 D_{KL}(q_\phi(h|x)\|p_\theta(h|x)) - \alpha_2 D_{KL}(q_\phi(h|x)\|q_\phi(h|r)) \\ -\alpha_3 D_{KL}(q_\phi(i|y)\|p_\theta(i|y)) - \alpha_4 D_{KL}(q_\phi(i|y)\|q_\phi(i|r)) \quad (5)$$

첫 번째 항은 수급자-복지 프로그램 정보를 재현하기 위한 데이터 r의 로그 가능도 기댓값이다. 두 번째, 세 번째, 다섯 번째 항은 각각 z, h, i의 사전 분포와 변분 사후 분포 사이의 정규화 항이다. 네 번째와 마지막 항은 수급자-복지 프로그램 정보 r에 더 초점을 맞추기 위해 추가된 정규화 항이다. α_2와 α_4는 r과 관련된 정규화 항의 세기를, α_1은 x와 관련된 정규화 항의 세기를, α_3은 y와 관련된 정규화 항의 세기를 조절한다. α_2와 α_4를 α_1과 α_3보다 크게 설정하면, 잠재 표현이 수급자와 복지 프로그램의 보조 정보 x, y보다 수급자-복지프로그램 정보 r에 더 초점을 맞추도록 강제하게 된다. graphical model과 신경망 구조는 <Figure 3>과 같고, 각 기호의 의미는 <Figure 2>와 같다.

Table 1. Variable Notation for Modified CVAE-CF

	Input	Latent Variable
Individual - Welfare Program Info.	r	z
Individual Auxiliary Info.	x	h
Welfare Program Auxiliary Info.	y	i

3.2 데이터 전처리 및 활용

(1) 복지 수급자 데이터 전처리

본 연구에서는 2018년 및 2019년에 수급자가 복지 프로그램을 제공받은 이력을 수급자-복지 프로그램 정보로, 수급자의 인적 및 건강 정보를 수급자 보조 정보로 활용하였다. 이 데이터들은 각 년도에 서로 다른 그룹으로부터 수집되었기 때문에 모든 정보를 가진 수급자를 추려내서 활용하였다.

복지 프로그램 추천은 수급자와 복지 프로그램의 특성에 크게 좌우되기 때문에 이를 반영하기 위하여 수급자와 복지 프로그램의 보조 정보를 모델 학습에 활용하였다. 수급자의 보조 정보는 나이, 성별과 같은 인적 정보와 건강 정보를 포함하며, 이는 변수의 유형에 따라 전처리되었다. 운동이나 건강 검진 시행 여부 같은 객관식 질문에 대한 답변인 범주형 변수는 원-핫 인코딩(one-hot encoding) 처리하였다. 범주형 변수더라도 음주나 흡연의 빈도와 같이 그 값에 순서가 있는 경우와 나이, 키, 몸무게 같은 연속형 변수는 가우시안 정규화를 하였다. 이렇게 전처리 된 변수들을 연결하여 총 340차원의 수급자 보조 정보를 얻었다.

전처리되기 전 수급자의 건강 정보 객관식 질문 및 답변의 구체적 예시는 다음과 같다. 먼저 수급자를 연령(노인, 성인, 청소년, 어린이, 신생아) 및 특정 군(장애인, 임산부, 재가암 등)에 속하는 지로 카테고리를 나누고, 각 카테고리에 속하는 수급자들에게 설문 조사를 진행한 결과 데이터다. 예를 들어, 노인에 속하는 수급자에게는 '삶이 허무하다고 느끼는가?', '보통 기분이 좋은 편인가?'에 예/아니오로 답변하는 노인 우울증과 관련된 질문부터 '100에서 7을 빼면 얼마인가?', '오늘은 몇 월 며칠입니까?'에 올바르게 대답했는지를 확인하는 노인 치매와 관련된 질문 등으로 구성되어 있다. 장애인에 속하는 수급자에 대해서는 식사나 목욕, 휠체어 이동 등을 혼자 수행 가능한지와 같이 장애 정도를 측정할 수 있는 질문지로 구성되어 있다. 이처럼 설문지는 각 카테고리와 관련된 사항들로 구성되어 있으며, 앞서 설명한 것처럼 설문지의 답변 유형에 따라 전처리를 다르게 진행하였다.

(2) 복지 프로그램 데이터 전처리

총 59개의 복지 프로그램 중 제공 횟수가 적은 10개의 복지 프로그램(철분제 엽산제 지원, 산전 진료비 지원, 선천성 대사이상 검사, 미숙아/선천성 이상아 의료비 지원, 신생아 청각 선별 검사, 아동건강관리 바우처사업, 다문화가족지원센터, 보육/교육비 지원, 아이 돌보미, 통번역 서비스)은 제거하였다.

복지 프로그램의 보조 정보는 복지 프로그램을 설명해주는 자연어 데이터이다. 따라서 별도의 임베딩이 필요하였는데, FastText(Bojanowski *et al.*, 2017) 모델을 사용하였다. FastText는 기존의 단어 임베딩 방법에서 최소 단위가 단어인 것과 달리 글자 또는 자모 단위의 최소 단위로 임베딩을 표현한다. 따라서 기존 학습 데이터에 없는 단어 또한 단어 임베딩을 표현할 수 있다는 장점이 있다. 본 연구에서 사용한 복지 프로그램의 용어는 단어 임베딩을 학습할 공개된 데이터셋이 부족하다. 따라서 본 연구에서는 글자(또는 자모) 단위의 FastText를 사용한 단어 임베딩을 사용했고, 단어 임베딩 학습을 위한 learning rate는 0.1, 단어 임베딩 차원은 300, 윈도우 사이즈는 5, epoch은 5로 설정하였다. 각 복지 프로그램은 여러 키워드로 표현된다. 가령, 운동 프로그램'의 경우 키워드로 '교육, 신체활동, 운동, 건강증진'을 보유하고 있다. 본 연구에서는 각 복지 프로그램의 단어 임베딩을 키워드의 단어 임베딩의 평균으로 구하여 활용했다.

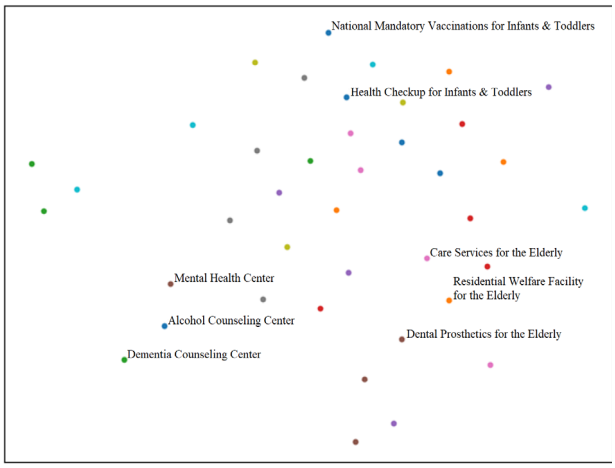


Figure 4. Welfare Program Word Embedding t-SNE

<Figure 4>는 복지 프로그램의 단어 임베딩을 t-SNE (Van der Maaten, 2008)를 사용하여 시각화한 것이다. 단어 임베딩이 복지 프로그램을 잘 표현한다면 유사한 복지 프로그램의 단어 임베딩은 가까이 분포하여 시각화되어야 한다. 영유아 관련 복지 프로그램, 노인 관련 복지 프로그램, 그리고 상담 관련 복지 프로그램의 단어 임베딩들이 가까이 분포하는 것을 <Figure 4>에 나타난다. 이를 통해 본 연구에서 FastText를 사용하여 표현한 단어 임베딩이 복지 프로그램의 특성을 반영하였음을 확인할 수 있다. <Figure 4>에 사용된 t-SNE를 위해서 perplexity는 30, 학습 반복 횟수는 1000으로 설정했다.

최종 모델 학습에 활용되며 보건소 내외에서 운영되는 복지 프로그램의 시행 횟수는 <Table 2>와 <Table 3>에서 각각 확인할 수 있다. 특히 기타 복지 프로그램이 다른 복지 프로그램보다 제공 횟수가 높은 것을 알 수 있는데, 단순히 제공이 많이 되었다는 이유로 모델이 수급자의 특성과 무관하게 기타 복지 프로그램을 추천해주는 것만으로 학습이 될 수 있다. 이는 수급자와 복지 프로그램 사이의 관계를 학습하는 데 방해가 될 수 있다고 판단하여, 이 영향을 살피고자 기타 복지 프로그램 포함 여부에 따라 두 가지 경우로 나눠서 실험하였다. 기타 복지 프로그램만을 제공받은 수급자는 기타 복지 프로그램을 포함하지 않은 경우에서 모두 제거되었다.

Table 2. Welfare Programs in the Health Center

Welfare Programs	Count	Welfare Programs	Count
Visiting Treatment	2,606	Eye Examination & Eyesight Recovery Operation for the Elderly	371
Inpatient Treatment	3,800	Dementia Early Screening	23,691
Smoking Cessation Program	1,319	Dementia Treatment Management Charge Support	135
Alcohol Moderation Program	59	etc. (the Elderly)	21,492
Exercise Program	2,923	Mother & Newborn Helper	68
Nutrition Program	1,592	Nutrition Plus Program	84
Obesity Program	103	National Mandatory Vaccinations for Infants & Toddlers	134
High Blood Pressure Class	1,694	Health Checkup for Infants & Toddlers	43
Diabetes Class	963	Nutrition Plus Program	113
etc. (Chronic Disease Management)	11,334	Medical Charge Support for Cancer Patients & Rare Intractable Disease	517
Inpatient Oral Care	485	Free Surgery Support	36
etc. (Oral Health)	9,469	Various Examination (Blood, Osteoporosis, etc.)	9,245
Dental Prosthetics for the Elderly	371	Palliative Therapy Service	604
Fluoride Spread Scaling for the Elderly	4,362	etc. (Others)	24,580

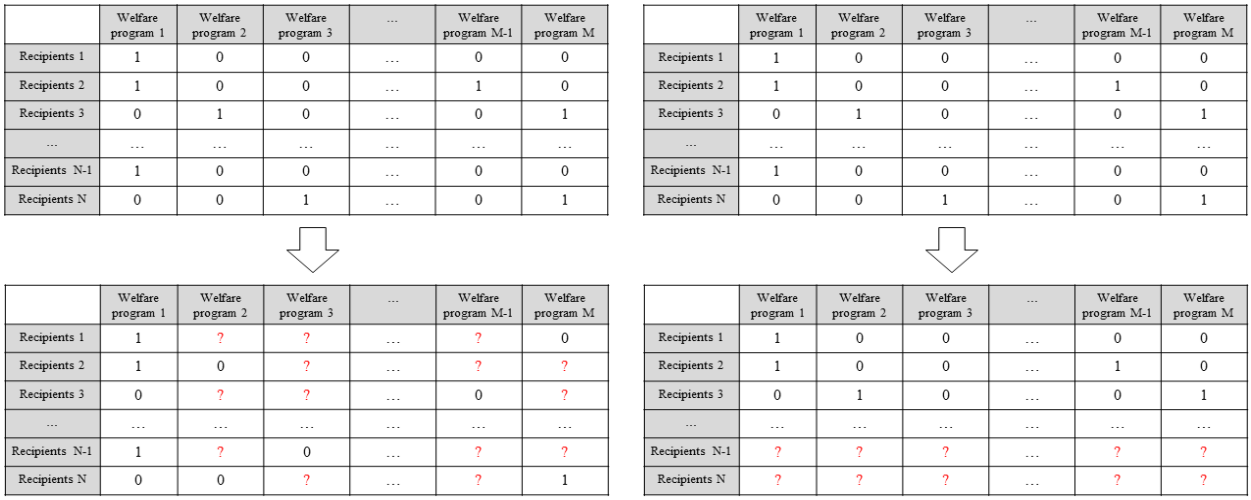
Table 3. Welfare Programs Outside the Health Center

Welfare Programs	Count	Welfare Programs	Count
Medical Institution Treatment	12,836	Activity Support System for the Disabled	73
Health Medical Examination	8,927	Housekeeping & Nursing Services Support	1,394
Oral Care Center for the Disabled	10,672	Bath Service Support	39
Palliative Therapy Service	9	Living Environment Improvement	358
Dementia Counseling Center	135	Learning Guidance	30
Mental Health Center	6,017	Vehicle & Companion Support, Outing Assistance	151
Alcohol Counseling Center	4,982	Other Voucher Program	404
Link to Other Services	51	etc. (Welfare)	21,743
Residential Welfare Facility for the Elderly	37	Medical Benefits Case Management	2,989
Dream Start Program	46	Long-term Care Insurance Transfer for the Elderly	406
Care Services for the Elderly	612		

3.3 데이터 희소성 문제에 기반한 학습 접근법

<Figure 5>에서 검은색 숫자는 학습에 사용 되고, 빨간색 물음표는 평가에 사용되는데, 추천 시스템의 일반적인 학습/평가 데이터 분할 방법은 <Figure 5(a)>와 같이 동일한 수급자에 대해 관측된 복지 프로그램 정보를 수급자에 관계없이 임의로 분할한다. 그러나 수급자-복지 프로그램 정보는 희소성이 크기 때문에 일반적인 추천 시스템의 학습/평가 데이터 분할 방법을 활용할 경우, 각 수급자가 어떤 복지 프로그램을 제공받았는지 모델이 충분히 학습할 수 없는 문제가 발생한다.

따라서, 데이터셋을 <Figure 5(b)>와 같이 수급자를 기준으로 학습/평가 데이터를 8:2 비율로 나누되, 각 수급자가 제공받은 복지 프로그램의 수에 따라 학습 단계에 활용하는 정보를 다르게 설정하였다. <Figure 5(b)>의 수급자 1과 같이 한 개의 복지 프로그램만 제공받은 수급자는 수급자 보조 정보만을 활용하여 해당 복지 프로그램을 예측할 수 있도록 하였다. <Figure 5(b)>의 수급자 2와 같이 두 개 이상의 복지 프로그램을 제공받은 수급자는 수급자 보조 정보와 제공받은 복지 프로그램 보조 정보를 같이 학습 단계에 제공하였다. 학습 단계에서 관측된 수급자-복지 프로그램 정보는 양의 항목으로, 관측되지 않은 것 중 임의로 샘플링을 하여 음의 항목으로 취급하였다. 평가 단계에서도 학습 단계와 같이 수급자가 제공받은 복지 프로그램 개수에 따라 보조 정보의 활용을 다르게 하였다. 전체적인 데이터 통계치는 <Table 4>와 같다.



(a) General training/test split method (b) Our training/test split method

Figure 5. Dataset Split Method

Table 4. Statistics of Dataset

Case	Individuals	Welfare Programs	Observation	Sparsity
w/ etc.	181,733	49	194,104	97.82%
w/o etc.	99,869	44	105,486	97.59%

4. 실 험

4.1 모델 선정 및 평가

본 연구에서는 수급자-복지 프로그램 정보, 수급자 보조 정보, 그리고 복지 프로그램 보조 정보를 활용하여 복지 프로그램 추천 실험을 수행하였다. 추천을 위하여 변분 오토 인코더 기반 협업 필터링 모델(VAE-CF), 그리고 조건부 변분 오토 인코더 기반 협업 필터링 모델(CVAE-CF)을 사용하였다. 특히 CVAE-CF는 수급자-복지프로그램 정보뿐만 아니라 수급자 및 복지 프로그램 보조 정보를 이용할 수 있다. 따라서, 1) 복지 프로그램 수급 정보가 없는 사람의 경우, 인적 정보를 활용하여 유사한 인적 정보를 갖춘 수급자들이 보편적으로 제공받은 복지 프로그램을 추천받을 것과 2) 기존에 복지 프로그램을 제공받은 경험이 있는 사람의 경우, 제공받은 복지 프로그램의 보조 정보를 활용하여 이와 유사한 복지 프로그램 추천을 기대할 수 있다. CVAE-CF-1은 수급자 보조 정보만 사용한 모델이며, CVAE-CF-2는 수급자 보조 정보와 복지 프로그램 보조 정보를 모두 사용한 모델을 나타낸다.

실험은 기타 복지 프로그램의 포함 여부에 따라 두 가지 데이터셋으로 나누어 수행하였다. 본 연구에서는 성능에 대한 평가를 위한 지표로 재현율(recall, R@k), 정밀도(precision, P@k), mean reciprocal rank(MRR), normalized discounted cumulative gain(NDCG)를 사용하였다. 재현율은 실제로 참인 경우에 모델이 참으로 예측한 것의 비율이며, 정밀도는 모델이 참으로 예측한 것 중에 실제 참인 것의 비율이다. k는 모델이 참으로 예측한 확률이 가장 높은 k개에 대한 평가를 의미한다. MRR과 NDCG는 우선순위를 고려한 평가 지표이며, 값이 더 클수록 좋은 성능을 의미한다.

수급자 정보 x, 복지 프로그램 정보 y의 차원은 각각 340, 300이 활용되었고, 수급자-복지 프로그램 정보 r의 경우 49(기타 포함) 또는 44(기타 미포함) 차원이 활용되었다. 또한, 각 정보에 해당하는 잠재 변수 h, i, z는 모두 64차원으로 지정하였으며, 본 연구에서는 실험의 공정성을 위해 모든 모델에 같은 설정을 사용하였다. learning rate는 0.003, 그리고 L2 정규화를 위한 파라미터는 0.001, batch의 크기는 256, 그리고 VAE-CF와 CVAE-CF의 잠재 변수 모델링을 위한 뉴런의 수는 64개로 지정하였다. 그리고 학습은 총 20 epoch으로 동일하게 수행하였다. 각 모델의 학습을 위한 하이퍼 파라미터(hyper-parameter)는 각 모델별로 가장 좋은 결과를 보인 값을 토대로 설정하였다.

4.2 정량 분석

<Table 5>는 전체 모델의 복지 프로그램 추천 성능을 보여준다. <Table 5>의 성능은 각 모델별로 5회 반복 실험한 결과값의 평균이다. VAE-CF와 CVAE-CF-1를 비교했을 때, 기타 복지 프로그램을 포함한 경우와 포함하지 않은 경우 모두 조건부 변분 오토 인코더(CVAE) 기반 모델이 더 좋은 성능을 보였다. 이는 복지 프로그램 추천에 있어 수급자의 보조 정보를 사용하는 것이 큰 효과가 있었음을 보여준다. 더불어 FastText를 사용한 임베딩을 이용해 복지 프로그램의 보조 정보를 사용한 CVAE-CF-2 가 CVAE-CF-1 보다 더 좋은 성능을 보여주었다. 위 실험을 통하여 본 연구는 복지 프로그램 이력 데이터를 통한 실험으로 조건부 변분 오토 인코더를 사용한 수급자와 복지 프로그램의 보조 정보 활용의 효과성을 확인하였다.

Table 5. Summary of Average Results for Five Repeated Experiments

Case	w/ etc.						w/o etc.					
Model	R@1	R@5	P@1	P@5	MRR	NDCG	R@1	R@5	P@1	P@5	MRR	NDCG
VAE-CF	0.050	0.273	0.057	0.061	0.192	0.361	0.029	0.127	0.035	0.029	0.122	0.290
CVAE-CF-1	0.122	0.540	0.127	0.114	0.318	0.470	0.237	0.597	0.242	0.123	0.406	0.538
CVAE-CF-2	0.170	0.551	0.180	0.116	0.362	0.503	0.241	0.630	0.247	0.131	0.420	0.551

4.3 정성 분석

(1) 잠재 변수 시각화

기타 복지 프로그램의 유무에 따른 잠재 변수의 차이를 확인하기 위하여 <Figure 6>에 CVAE-CF-2의 잠재 변수를 t-SNE (Van der Maaten, 2008)를 사용하여 시각화하였다. <Figure 6>의 서로 다른 색은 서로 다른 복지 프로그램을 의미한다. 기타 복지 프로그램을 포함하지 않은 경우같은 복지 프로그램을 접한 수급자의 잠재 변수를 시각화하였을 때, 클러스터링이 기타 복지 프로그램을 포함한 경우에 비해 비교적 잘 되어 있음을 확인할 수 있다. 이 결과를 통하여 기타 복지 프로그램을 포함하지 않은 경우의 잠재 변수가 비교적 표현력이 좋음을 확인할 수 있었고, 정량 결과를 보았을 때도 기타 복지 프로그램을 포함하지 않은 경우의 성능이 더 좋음을 확인할 수 있다. <Figure 6>에 사용된 t-SNE 파라미터 값은 <Figure 3>에서 사용한 값과 동일하다.

Table 6. Hyper-parameter Values of CVAE-CF Models Corresponding for <Table 5>

Case	α_1	α_2	α_3	α_4
CVAE-CF-1 w/ etc.	0.1	1	-	-
CVAE-CF-1 w/o etc.	0.01	0.1	-	-
CVAE-CF-2 w/ etc.	0	0.01	0.1	1
CVAE-CF-2 w/o etc.	0.01	0.1	0	10

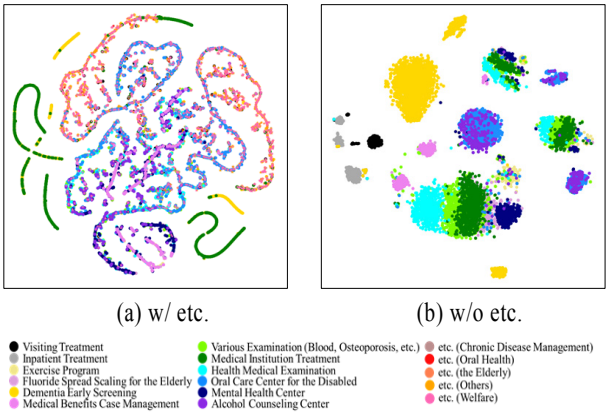


Figure 6. t-SNE Plot of Latent Factors of Users from CVAE-CF with Users and Items Auxiliary Information

(2) 사례 분석

CVAE-CF-2와 CVAE-CF-1의 차이는 복지 프로그램 보조 정보 사용 유무의 차이이다. 본 장에서는 복지 프로그램 보조 정보 사용 여부에 따른 추천 성능 차이를 확인하기 위한 사례 분석을 진행하였다. '알코올 상담센터'를 이용한 이력이 있는 수급자에게 '알코올 상담센터'의 보조 정보를 사용한 경우에는 '치매 조기 검진, 의료기관 진료, 장애인 구강진료센터, 정신보건센터, 건강검진' 순으로 복지 프로그램이 추천되었다. 보조 정보를 사용하지 않았을 때 치매 조기 검진, 장애인 구강 진료센터, 의료기관 진료, 건강검진, 혈액, 골다공증 등 각종 검사' 순으로 복지 프로그램이 추천되었다. 실제 해당 수급자가 수행한 복지 프로그램은 '정신보건센터'와 '건강검진'으로 복지 프로그램 보조 정보를 사용한 경우보다 정확한 추천을 한 사례를 확인할 수 있었다.

5. 결 론

복지 프로그램은 전반적인 건강과 삶의 질 향상을 달성하기 위한 국가 운영 서비스이다. 제한된 자원으로 운영되기 때문에 적절한 추천이 필요하며, 이는 사용자와 복지 프로그램 특성 간의 복잡한 관계에 의존한다. 따라서 사용자가 과거에 어떤 복지 프로그램을 제공받았는 지와 더불어 사용자/복지 프로그램의 보조 정보 데이터를 활용하여 CVAE-CF 모델을 재설계하고, 실험을 통해 향후 현장에서 이를 적용할 때 바람직한 추천 성능을 낼 수 있음을 검증하였다. 우리는 이번 연구의 도움으로 개인의 필요에 맞춰 다양한 복지 프로그램이 올바르게 제공될 것을 기대한다.

그러나 본 연구에서 재설계한 모델은 수급자가 복지 프로그램을 제공받은 이력이 있는 경우에만 복지 프로그램의 보조 정보를 활용할 수 있다는 제한이 있다. 따라서 처음 복지 프로그램을 수급받는 사람에게도 적절한 추천이 이뤄질 수 있는 모델로 개선이 필요할 것이다.

참고문헌

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017), Enriching Word Vectors with Subword Information, *Transactions*

of the Association for Computational Linguistics, **5**, 135-146.
Jung, H. S. and Choi, E. H. (2017), A Qualitative Research on the Characterization of Visiting Healthcare in Public Health Centers - Identification of Characteristics, Service Process and Recipient Change, *Korean Journal of Health Education and Promotion*, **34**(5), 107-119.
Kim, H. G., Jang, S. N., Chin, Y. R., Hur, J., and Lee, R. S. (2022), Contract Employment Experiences of Visiting Nurses at Public Health Centers in the Metropolitan Area: Focused on Employment Type and Treatment, *Journal of Korean Academy of Community Health Nursing*, **33**(2), 175-187.
Kingma, D. P. and Welling, M. (2013), Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.
Koren, Y., Bell, R., and Volinsky, C. (2009), Matrix Factorization Techniques for Recommender Systems, *Computer*, **42**(8), 30-37.
Lee, W., Song, K., and Moon, I. C. (2017), Augmented variational autoencoders for collaborative filtering with auxiliary information, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1139-1148.
Ministry of Health and Welfare [MOHW] (2021), Health center and Health branch operations status, http://www.mohw.go.kr/react/jb/sjb1101vw.jsp?SEQ=101&MENU_ID=03320101&page=1&PAR_MENU_ID=03.
Ministry of Health and Welfare [MOHW] (2022), Guide to the 2022 Home care visit support project, http://www.mohw.go.kr/react/jb/sjb030301vw.jsp.
Park, W. B. (2021), Need to Improve Visit Management for Vulnerabel People Compared to Infectious Diseases, whosaeng, http://www.whosaeng.com/130850.
Rezende, D. J., Mohamed, S., and Wierstra, D. (2014), Stochastic Backpropagation and Approximate Inference in Deep Generative Models, *International Conference on Machine Learning*, 1278-1286.
Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007), Collaborative Filtering Recommender Systems, The adaptive web, Springer, Berlin, Heidelberg, 291-324.
Sohn, K., Lee, H., and Yan, X. (2015), Learning Structured Output Representation Using Deep Conditional Generative Models, *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 3483-3491.
Su, X. and Khoshgoftaar, T. M. (2009), A survey of collaborative filtering techniques, Advances in artificial intelligence.
Van der Maaten, L. and Hinton, G. (2008), Visualizing data using t-SNE, *Journal of Machine Learning Research*, **9**(11), 2579-2605.

저자소개

김성은 : 홍익대학교 산업공학과에서 2021년 학사학위를 취득하고 한국과학기술원 데이터사이언스대학원 석사과정에 재학중이다. 연구분야는 머신러닝이다.

지민기 : 서울대학교 조선해양공학과에서 2016년 학사, 한국과학기술원 산업및시스템공학과에서 2018년 석사, 2022년 박사학위를 취득하고, 구글 코리아에 재직 중이다. 연구분야는 머신러닝이다.

문일철 : 서울대학교 컴퓨터공학과에서 2004년 학사, 카네기 멜런 대학교 정보시스템학과에서 2005년 석사학위를 취득하고, 카네기 멜런 대학교 컴퓨터 과학과 박사학위를 2008년 취득하였다. 2011년부터 한국과학기술원 산업및시스템공학과 교수로 재직하고 있다. 연구분야는 컴퓨터 공학, 경영학, 사회학, 작전 연구, 군사 지휘 및 통제 분석, 대테러 분석, 정보 분석, 재난 관리 등이다.

주원영 : 한국과학기술원 수리과학과에서 2014년 학사, 2016년 석사학위를 취득하고, 한국과학기술원 산업및시스템공학과에서 박사학위를 2020년 취득하였다. 이후, 삼성전자에서 책임연구원으로 근무하였고, 2022년부터 이화여자대학교 통계학과 교수로 재직하고 있다. 연구분야는 머신 러닝, 데이터 마이닝, 통계분석, 조합론 등이다.