DBpia

논문 18-43-08-15 The Journal of Korean Institute of Communications and Information Sciences '18-08 Vol.43 No.08
https://doi.org/10.7840/kics.2018.43.8.1357

효과적인 협업 필터링을 위한 평점 정보 도움을 받는 잡음제거 오토인코더

김 현 진•, 신 동 진*, 신 원 용°, 황 창 하°

Rating Information-Aided Denoising AutoEncoder for Effective Collaborative Filtering

Hyun-Jin Kim•, Dong-Jin Shin*, Won-Yong Shin°, Changha Hwang°

요 약

추천 시스템은 사용자의 선호도를 기반으로 상품을 추천해 주는 것으로, 대표적으로 협업 필터링 방법이 있다. 그러나 협업 필터링은 사용자가 부여한 평점 데이터가 충분하지 않은 경우, 추천 정확도가 낮아지는 문제가 있다. 본 논문에서는 희소한 데이터에 주로 사용되는 기계 학습 방법인 잡음제거 오토인코더를 응용한 새로운 협업 필터링 방법을 소개하고, 이를 통해 개선된 추천 정확도를 보이고자 한다. 제안한 방법에서는 평점을 효과적으로 예측하기 위해서 해당 사용자의 평점 평균과 해당 상품의 평점 평균의 선형 결합을 고려한다. 즉, 각각의 평점 평균과 가중치의 곱으로 평점을 예측하며, 각 가중치는 잡음제거 오토인코더를 통하여 학습된다. 제안하는 모델은 Top-N 추천 시스템 환경에서 정밀도, 재현율, F-measure, nDCG의 측면에서 성능 검증된다. MovieLens 데이터셋 사용 시 기존의 잡음제거 오토인코더 기반 협업 필터링보다 nDCG의 측면에서 최대 210% 향상된 성능을 보이는 것을 확인한다.

Key Words : Collaborative Filtering (CF), Denoising AutoEncoder (DAE), Rating, Recommender System

ABSTRACT

Recommendation systems are the ones that recommend a preferred item based on user's preference, and typically include a collaborative filtering method. However, collaborative filtering has a shortcoming such that recommendation accuracy is degraded if there is no sufficient rating information assigned by users. In this paper, we introduce a new collaborative filtering method that employs denoising autoencoder which is one of machine learning techniques mainly used for sparse data, and show the improved recommendation accuracy. In the proposed method, to effectively predict ratings, a linear combination of the rating average of a target user and the rating average of a target item is considered. In other words, the rating is predicted by a weighted sum of each rating average and each weight is learned through denoising autoencoder. The performance of our model is demonstrated in terms of precision, recall, F-measure, and nDCG in the top-N recommendation system. When the MovieLens dataset is used, it is verified that the proposed method outperforms the conventional denoising autoencoder-based collaborative filtering by up to 210% in terms of nDCG.

※ 본 연구는 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(2017R1D1A1A09000835)이며, 2017학년도 단국대학교 대학연구비 지원으로 연구되었음.
• First Author : (ORCID:0000-0002-4069-7098)Department of Data Science, Dankook University, khj0417@dankook.ac.kr, 학생회원
° Corresponding Authors : (ORCID:0000-0002-6533-3469)Department of Computer Science and Engineering, Dankook University, wyshin@dankook.ac.kr, 종신회원
° Corresponding Authors : (ORCID:0000-0003-1262-5051)Department of Statistics, Dankook University, chwang@dankook.ac.kr, 정회원
* (ORCID:0000-0001-7182-4646)Department of Statistics, Dankook University, sdj1035@dankook.ac.kr
논문번호 : 201808-241-C-RN, Received August 13, 2018; Revised August 19, 2018; Accepted August 19, 2018

www.dbpia.co.kr

Ⅰ. 서 론

추천 시스템은 사용자의 취향, 소비 등의 데이터를 기반으로 선호할 만한 상품을 추천하는 것으로 여러 기업에서 광범위하게 사용되고 있다. 특히, 온라인을 기반으로 하는 기업에서는 우수한 추천 시스템이 핵심 경쟁력이라고 할 수 있으며 이를 위해 많은 투자를 하고 있다. 대표적인 예로, 세계 최대의 온라인 소매 업체인 아마존 (Amazon), 세계 최대 엔터테인먼트 네트워크 서비스 기업인 넷플릭스 (Netflix)가 있다. 아마존은 사용자가 구매 혹은 검색한 물건들을 중심으로 소비 패턴을 분석하여 상품을 추천한다. 자사의 추천 시스템인 A9[1]를 특허로 등록하여 기존의 온라인 쇼핑몰과의 차별성을 추구하였다. 넷플릭스는 소비자가 감상한 콘텐츠와 그에 부여한 평점 정보 등을 활용하여 선호할만한 콘텐츠를 추천한다. 또한, 추천 시스템의 성능을 높이기 위하여 넷플릭스 대회(Netflix Prize)[2]를 주최하였으며 최근 자사의 추천 시스템[3]을 발표하기도 하였다. 이는 사용자가 아직 접하지 않은 상품에 대한 선호도를 예측하는 추천 시스템의 비약적인 발전을 촉발시켰다.

1.1 사전 연구

여기서는 추천시스템에서 가장 널리 활용되는 기법인 협업 필터링은 상품 속성과 해당 상품의 속성에 대해 누적된 사용자의 평점 데이터를 활용하여, 특정 사용자에 대해 선호도가 높을 것이라고 예측되는 상품을 추천하는 방법이다. 협업 필터링은 크게 메모리 기반 협업 필터링과 모델 기반 협업 필터링으로 나눌 수 있다. 메모리 기반 협업 필터링은 기존의 구매 및 평점 데이터를 활용하여 접하지 않은 상품에 대한 선호도를 예측하는 기법이다. 반면, 모델 기반 협업 필터링은 기존의 평점 데이터 기반으로 모델링을 통하여 평점을 예측하는 기법인데, 최근 잠재 모델에 기반을 둔 방법이 주목을 받고 있다[4]. 여기에서 잠재 모델이란 사용자가 특정 상품을 선호하는 내재된 이유를 알고리즘을 사용하여 알아내는 기법으로, 이에 기계 학습 방법들을 활용할 수 있다.

한편, 협업 필터링은 누적된 사용자의 선호도를 나타내는 평점 데이터가 부족하면 추천 시스템의 성능을 보장할 수 없다는 한계점이 있다. 사용자가 부여한 기존의 평점 데이터를 활용하여 새로운 상품에 대한 선호도를 예측하므로, 기존의 평점 데이터가 희소하면 협업 필터링의 성능이 저하된다. 현실에서는 사용자가 모든 상품에 대한 평점을 부여할 수 없으므로, 평점 데이터에는 근본적으로 희소성[5]의 문제가 존재한다.

1.2 제안 방안

본 논문에서는 추천 정확도를 높이기 위하여 희소성의 문제를 해결하기 위해 기계 학습 방법론 중 하나인 잡음제거 오토인코더 (Denoising AutoEncoder, DAE)[6]를 활용한 새로운 협업 필터링 기반 추천 시스템 기술을 소개한다. DAE는 데이터에 손실이 발생하였을 때 이를 복원하기 위하여 사용되는 비지도 학습 방법으로, 주로 데이터가 희소할 경우 사용된다. 평점 데이터에서의 결측 값은 부정적인 사전 선호도의 표현으로 가정하여 0으로 대체하였기에, DAE의 비용함수에 결측 값을 포함한다. 또한, 기존 DAE 기술[6]과는 달리 하나의 평점을 예측하기 위해 해당 사용자 평점 평균과 해당 상품의 평점 평균의 선형 결합을 고려한다. 이 때, 두 평점 평균의 최적의 반영 비율 즉, 선형 결합의 가중치는 각 사용자와 상품에 따라 다름을 가정하는데, 각 가중치를 DAE로 학습함으로써 내재된 선호도를 찾아 해당 평점을 예측하게 된다. MovieLens 데이터셋을 활용하여 사용자가 부여할 것이라고 예측되는 평점의 상위 N개의 상품을 추천하고, 기존의 협업 필터링에서 사용되는 DAE와 성능 비교를 수행한다. 실험 결과로부터 기존 DAE 기반 협업 필터링 방법 대비 정밀도 (Precision), 재현율 (Recall), F-measure, nDCG (normalized Discounted Cumulative Gain) 측면에서 대략 130%, 165%, 140%, 173% 향상된 성능을 보임을 검증한다. 특히, 사용자 정보를 활용하여 MovieLens-1M 데이터셋에 적용하였을 때, nDCG 측면에서 최대 210% 향상된 성능을 보임을 확인한다.

본 논문의 구성은 다음과 같다. 2장 협업 필터링 개요에서는 협업 필터링의 개념, 종류 및 본 논문에서 활용한 모델에 관하여 설명한다. 그리고 3장에서는 본 논문에서 제안하는 알고리즘을 설명하며, 4장에서는 실험 결과와 함께 다른 알고리즘과 성능을 비교한다. 마지막으로 5장에서는 본 연구의 결과와 향후 연구 과제를 제시한다.

Ⅱ. 협업 필터링 개요

협업 필터링은 사용자가 기존에 부여했던 평점의 패턴을 통해 아직 평점을 부여하지 않은 상품에 대한 선호도를 예측하는 기술이다[7]. 많은 추천시스템에서는 사용자가 선호할 것으로 예측되는 상품을 추천하기 위해 협업 필터링을 기반으로 한 시스템을 구축하

www.dbpia.co.kr

고 있다.

협업 필터링은 크게 이웃 기반 (혹은 메모리 기반) 알고리즘과 모델 기반 알고리즘으로 나뉜다. 이웃 기반 협업 필터링 알고리즘은 수집한 정보를 이용해 사용자 간 또는 상품 간 유사도를 계산 후 해당 항목의 평점을 직접 예측하는 방법이며, 대표적인 알고리즘으로 사용자 기반 협업 필터링과 상품 기반 협업 필터링이 있다[5].

2.1 메모리 기반 협업 필터링

사용자 기반 협업 필터링[5]은 어떤 사용자의 평점 기록을 이용해 비슷한 성향을 가진 사용자들을 찾고, 이를 기반으로 상품의 평점을 예측한다. 이 때, 유사한 사용자를 찾기 위해서 유사도 기준을 사용하는데, 유사도는 피어슨 상관계수와 코사인 유사도[5]가 주로 사용된다. 피어슨 상관 계수와 코사인 유사도는 다음과 같이 계산한다.

$$\rho(x,y)=\frac{\sum_{i\in I_{xy}}\left(r_{xi}-\bar{r}_x\right)\left(r_{yi}-\bar{r}_y\right)}{\sqrt{\sum_{i\in I_{xy}}\left(r_{xi}-\bar{r}_x\right)^2}\sqrt{\sum_{i\in I_{xy}}\left(r_{yi}-\bar{r}_y\right)^2}} \quad (1)$$

$$\cos(x,y)=\frac{\sum_{i\in I_{xy}} r_{xi}r_{yi}}{\sqrt{\sum_{i\in I_{xy}} r_{xi}^2}\sqrt{\sum_{i\in I_{xy}} r_{yi}^2}} \quad (2)$$

여기에서 x, y는 서로 다른 두 사용자를 의미하며, r_{ui}는 사용자 u가 상품 i에 부여한 평점을 의미한다. I_{xy}는 두 사용자 x,y가 모두 평점을 부여한 상품의 집합이다. 위의 유사도는 [-1, 1]을 범위로 가지며, 1에 가까울수록 두 사용자의 평점 패턴이 유사함을 의미한다. 이러한 유사도를 기반으로 가중 합을 계산함으로써 r_{ui}을 예측한다.

그림 1은 사용자 u가 상품 i에 부여한 평점 r_{ui}의 집합을 행렬로 나타낸 예시이다. 예를 들어, r_{14}는 사용자 1이 상품 4에 부여한 평점을 의미한다. 사용자 기반 협업 필터링은 사용자 1이 상품 5에 부여한 평점인 r_{15}을 예측할 때, 사용자 1을 제외한 다른 사용자의 평점을 가중 합한다. 사용자 2가 사용자 1과 가장 비슷한 평점 패턴을 보인다고 가정하면, 사용자 2가 상품에 부여한 평점 패턴 r_{25}는 가장 높은 가중치가 부여된다. 그림 1에서는 높은 유사도를 보이는 사용자

User \ Item	item 1	item 2	item 3	item 4	item 5
user 1	5	3	4	4	?
user 2	3	1	2	2	3
user 3	4	3	4	3	5
user 4	3	3	1	5	4
user 5	1	5	5	2	1

Weighted Sum.

그림 1. 사용자 기반 협업 필터링의 예
Fig. 1. An example of user-based collaborative filtering

의 평점에 더욱 진하게 음영을 표시하였다. 사용자 4와 사용자 5는 상대적으로 사용자 1과의 유사도가 낮아 가중치가 낮다.

상품 기반 협업 필터링[8]은 각 상품의 평점 기록을 통해 비슷한 평점 패턴을 가지는 상품을 찾고, 이를 기반으로 상품의 평점을 예측하는 방법이다. 상품 기반 협업 필터링 역시 사용자 기반 협업 필터링과 마찬가지로 유사도를 기반으로 가중 합을 계산함으로써 상품의 평점을 예측한다.

그림 2에서 상품 기반 협업 필터링은 상품 5가 사용자 1로부터 부여받은 평점인 r_{15}을 예측할 때, 상품 5를 제외한 다른 상품의 평점을 가중 합한다. 상품 1이 상품 5와 가장 비슷한 평점 패턴을 보인다고 가정하면, 상품 1이 부여받은 평점 패턴은 가장 높은 가중치가 부여된다. 그림 2에서는 높은 유사도를 보이는 사용자의 평점에 더욱 진하게 음영을 표시하였다. 상품 3은 상대적으로 상품 5와의 유사도가 낮아 가중치가 낮다.

User \ Item	item 1	item 2	item 3	item 4	item 5
user 1	5	3	4	4	?
user 2	3	1	2	2	3
user 3	4	3	4	3	5
user 4	3	3	1	5	4
user 5	1	5	5	2	1

Weighted Sum.

그림 2. 상품 기반 협업 필터링의 예
Fig. 2. An example of item-based collaborative filtering

2.2 모델 기반 협업 필터링

한편, 모델 기반의 협업 필터링은 평점 정보를 모델링에 사용한다. 대표적인 모델 기반의 협업 필터링 기술은 특이 값 분해 (Singular Value Decomposition,

www.dbpia.co.kr

SVD)[9],[10]를 활용한 방법이다. 각 평점이 저차원의 잠재 공간으로 표현할 수 있다고 가정하는 경우, SVD를 통해 잠재 공간을 찾을 수 있는데, 이러한 모델은 다음과 같이 표현할 수 있다.

$$\hat{R} \approx USV^T \tag{3}$$

여기에서 R은 사용자가 상품에 부여한 평점의 집합인 평점 행렬을 나타내며, S는 잠재 요인 행렬을 의미한다. 잠재 요인 행렬은 저차원이면서도 R을 대표하는 정방 행렬이다. U와 V는 각각 좌 특이 벡터와 우 특이 벡터를 나타낸다. 좌 특이 벡터는 사용자와 잠재 요인의 관계를, 우 특이 벡터는 상품과 잠재 요인의 관계를 의미한다. 결측 값이 포함된 평점행렬을 위해 기댓값 최대화 (Expectation Maximization, EM) 알고리즘을 통해 SVD를 수행하기도 한다[9].

또 다른 모델 기반 방법으로 오토인코더 (AutoEncoder, AE)가[11] 사용된다. AE는 인공 신경망 모델 중 하나로 입력변수를 그대로 출력하는 신경망을 의미한다. AE는 입력 변수를 저차원의 잠재 변수로 사상하는 인코더와, 잠재 변수를 원 입력 변수로 사상하는 디코더로 구성된다. 즉, AE는 인코더와 디코더를 학습함으로써, 입력 변수를 대표하는 저차원의 잠재 변수를 얻을 수 있는 모델이다. 가장 단순한 형태의 AE는 하나의 은닉층을 가지는 인공 신경망이며, 이 경우 잠재변수와 신경망의 출력은 다음과 같다.

$$Z = f^{(1)}\left(XW^{(1)} + B^{(1)}\right) \tag{4}$$

$$\hat{X} = f^{(2)}\left(ZW^{(2)} + B^{(2)}\right) \tag{5}$$

여기에서 $X = \{x_1, \ldots, x_p\}$는 입력 변수들을 나타내며, $Z = \{z_1, \ldots, z_q\}$는 인공 신경망의 은닉층인 잠재 변수들을 나타낸다. 또한, $f^{(1)}, f^{(2)}$와 $W^{(1)}, W^{(2)}, B^{(1)}, B^{(2)}$는 각각 인공 신경망의 활성 함수와 변수를 나타낸다. 한편, AE의 변수들은 역전파 알고리즘을 통해 학습된다.

2.3 협업 필터링에서의 희소성 해결 방안

추천 시스템에서 협업 필터링을 활용할 때의 가장 큰 문제점은 데이터의 희소성이다[7]. 협업 필터링은 사용자가 상품에 부여했던 기존의 평점 정보를 이용해 새로운 상품에 대한 평점을 예측하므로, 데이터의 희소성은 협업 필터링의 성능을 저하시키는 원인이 된다. 그러나 실제 데이터셋에서의 평점 행렬에서는 대부분 많은 결측 값이 존재한다.

그림 3은 희소성의 문제가 있는 경우의 예시이다. 전체 데이터셋의 크기에 비하여 많은 결측 값이 존재하는 것을 확인할 수 있다. 최근 연구에서는 희소성의 문제를 해결하기 위하여 아래와 같이 여러 방법이 제안되었다.

이웃 기반 협업 필터링 방법에서의 희소성 해결 방안으로 적응적 최대화 방법 (Adaptive-Maximum imputation method, AdaM)이 제안되었다[12]. 사용자 u에 대하여 사용자 기반 협업 필터링을 통해 상품을 추천해 준다고 가정할 경우 일반적인 메모리 기반 알고리즘에서의 방법론은 모든 사용자의 유사도를 이용한다. 반면, AdaM은 그 대신 상품 i에 평점을 부여한 사용자들의 유사도만을 이용한다.

한편, 모델 기반 협업 필터링 방법에 대해 결측 값을 일종의 선호도의 표현으로 가정한 영 삽입 (Zero Injection) 기법이 제안되었다[13]. Zero Injection은 사용자의 선호도를 사전 선호도와 일반적인 평점 정보인 사후 선호도로 구분하였다. 사전 선호도는 평점을 부여하기 전 상품에 대한 선호도로 평점이 부여되었는지 여부를 통해 나타난다. 즉, 사전 선호도는 평점이 부여되었는지를 의미하는 지시 함수로 표현된다. Zero injection은 이러한 지시 함수를 통해 사전 선호 행렬 P를 생성하고, 사전 선호 행렬에 기존의 결측 값 대체 방법을 사용하여 $\hat{P}$를 얻는다. 대체된 사전 선호도 중 일정 이하의 값에 해당하는 u, i의 쌍에 대하여, 평점 행렬의 r_{ui}을 0으로 대체한다. 이러한 평점 행렬을 이용한 추천 시스템의 정확도는 기존의 방법에 비해 크게 개선됨을 보였다.

User \ Item	item 1	item 2	item 3	item 4	item 5
user 1	5	?	?	?	?
user 2	?	1	?	?	?
user 3	?	3	?	?	5
user 4	?	?	1	?	4
user 5	?	?	?	2	?

그림 3. 일부 평점 정보가 없는 경우의 평점 행렬
Fig. 3. A rating matrix having partially masked ratings

www.dbpia.co.kr

2.4 잡음제거 오토인코더

AE는 잡음 혹은 결측 값이 포함된 자료를 복원하기 위해 기계 학습에서 주로 사용되는 비 지도학습 모델이다. 자료의 복원은 다른 변수들을 이용해 복원하고자 하는 변수를 예측함으로써 이루어진다. 평점 자료에서의 AE는 관측된 평점들을 이용해 관측되지 않은 평점을 예측하는 역할을 한다. DAE는 AE의 응용 모델로써, 데이터의 희소성이 높은 경우에 더욱 효과적인 모델로 알려져 있다. 본 절에서는 AE와 DAE를 협업 필터링에 적용하는 과정을 소개한다.

2.4.1 오토인코더

협업 필터링을 위한 AE인 AutoRec[14]은 한 사용자의 구매 평점의 집합인 $R_u = \{r_{u1}, \ldots, r_{un}\}$ 혹은 한 상품의 구매 평점의 집합인 $R_i = \{r_{1i}, \ldots, r_{mi}\}$을 AE의 입력 벡터로 사용한다. 사용자의 구매 평점을 입력 벡터로 사용한 경우인 U-AutoRec의 비용함수는 다음과 같다.

$$\| R_u^O - \hat{R}_u^O \|^2 + \frac{\lambda}{2}\left(\| W^{(1)} \|_F^2 + \| W^{(2)} \|_F^2 \right) \quad (6)$$

여기에서 O는 사용자 u가 평점을 부여한 상품의 집합을 나타낸다. R_u^O는 사용자 u가 부여한 평점 벡터를, $\hat{R}_u^O$은 실제 관측되었던 평점에 대한 예측 평점 벡터를 나타낸다. 한편, AutoRec의 비용 함수를 식 (6)과 같이 구성함으로써 결측 값은 모델의 학습에 영향을 주지 않게 된다. 학습이 끝난 후 예측하고자 하는 평점은 다음과 같이 표현된다.

$$\begin{aligned} &\hat{R}_u = f^{(2)}\left(Z W^{(2)} + B^{(2)}\right), \\ &\text{where } \; Z = f^{(1)}\left(R_u W^{(1)} + B^{(1)}\right) \end{aligned} \quad (7)$$

이 때, 예측하고자 하는 평점과 모든 결측 값은 0으로 입력한다. 다음은 AutoRec을 통해 평점을 예측하는 과정을 도식화한 그림이다.

그림 4는 사용자 3이 부여한 평점 벡터가 {5, 1, r_{33}, 2, 4}일 때, r_{33}을 예측하는 과정이다. r_{33}과 입력 벡터 안의 주홍색 원으로 표시된 노드는 예측하고자 하는 평점과 결측 값을 의미한다. r_{33}는 0으로 대체하여 입력되고, U-AutoRec을 통해 3으로 예측된다. 이로써 사용자 3이 부여한 평점 벡터의 출력은 {5, 1, 3, 2, 4}로 주어진다.

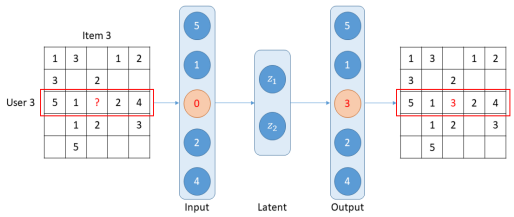

그림 4. U-AutoRec을 통해 평점을 예측하는 과정
Fig. 4. The process of predicting a rating of user u through U-AutoRec

2.4.2 잡음제거 오토인코더

DAE는 AE의 비용 함수를 수정한 모델이다. DAE의 학습에는 인위적으로 잡음을 추가하는 과정이 포함되며, Salt-and-Pepper 잡음과 마스킹 잡음[6] 등을 사용한다. Salt-and-Pepper 잡음은 관측된 입력 변수 중 일부를 해당 입력 벡터의 최댓값 혹은 최솟값으로 대체함을 의미하며, 마스킹 잡음은 결측 값과 같이 대체함을 의미한다. 그리고 DAE의 비용 함수는 잡음이 추가된 입력 변수에 대한 비용 함수와, 잡음이 추가되지 않은 입력 변수에 대한 비용 함수의 선형 결합으로 구성된다. 즉, "잡음이 포함된 입력 변수를 얼마나 실제 값과 유사하게 복원하는지"를 의미하는 잡음제거 오차와 "입력 변수를 얼마나 원래대로 복원하는지"를 의미하는 복원 오차의 가중 합으로 구성된다. 협업 필터링을 위한 DAE[15]의 비용 함수는 다음과 같다.

$$\alpha\left(\| R_u^{O\cap J} - \hat{R}_u^{O\cap J} \|^2 \right) + \beta\left(\| R_u^{O\cap J^C} - \hat{R}_u^{O\cap J^C} \|^2 \right) \quad (8)$$

여기에서 J는 잡음이 추가된 상품의 집합을 나타낸다. 즉, $R_u^{O\cap J}$, $\hat{R}_u^{O\cap J}$는 각각 사용자 u가 평점을 부여한 상품 중 잡음이 추가된 상품에 대한 평점 벡터와 그 예측 값이며, $R_u^{O\cap J^C}$, $\hat{R}_u^{O\cap J^C}$는 잡음이 추가되지 않은 상품에 대한 평점 벡터와 그 예측 값이다. 이러한 방법을 사용하면 α와 β의 비율에 따라서 잡음의 여부에 따른 두 비용의 반영 비율이 결정된다. 한 편, 예측하고자 하는 평점을 얻을 때에는 식 (7)과 같은 과정을 따른다.

Ⅲ. 제안 방법

본 논문에서 제안하는 방법은 평점 평균의 선형 가정[16]을 적용한 DAE이다. 제안하는 방법 중 사용자의 평점 패턴을 입력 벡터로 사용하는 방법의 평점 예측

www.dbpia.co.kr

은 다음과 같이 이루어진다.

$$\hat{R}_u = \overline{R}_i \times \gamma(R_u) + \overline{R}_u \times \delta(R_u) \quad (9)$$

여기에서 $\overline{R}_i$, $\overline{R}_u$는 각각 상품 i의 평점 평균과 사용자 u의 평점 평균을 의미하며, $\gamma(R_u)$와 $\delta(R_u)$는 DAE로 출력되는 선형 결합의 가중치를 의미한다. $R_u = \{r_{u1}, \ldots, r_{un}\}$은 사용자 u의 평점 벡터를 의미하며, 신경망의 입력 벡터로 사용된다. 한편, 상품의 평점 패턴인 $R_i = \{r_{1i}, \ldots, r_{mi}\}$을 입력 벡터로 사용하는 경우 다음과 같이 평점 예측이 이루어진다.

$$\hat{R}_i = \overline{R}_i \times \gamma(R_i) + \overline{R}_u \times \delta(R_i) \quad (10)$$

제안하는 방법은 일반적인 협업 필터링을 위한 DAE와 두 가지 차이점을 가지고 있다. 첫 번째는 제안 방법에서는 신경망에서 평점 평균으로 이루어진 선형 결합에서의 가중치를 출력한다는 점이다. 예측하고자 하는 하나의 평점을 사용자와 상품의 두 평점 평균의 선형 결합으로 가정하는 경우, 제안 방법은 두 평점 평균의 적절한 반영 비율을 찾는 작업을 자동으로 수행한다는 장점이 있다. 또한, 본 논문의 제안 방법은 기존 DAE 기술[16]과 달리, 사용자 혹은 상품의 성질에 따라 두 평점 평균에 대한 최적의 반영 비율이 다름을 고려한다. 두 번째는 DAE의 비용함수에 결측값에 대한 오차도 포함한다는 점이다. 즉, 본 논문에서 제안하는 방법은 평점이 부여되지 않은 경우가 낮은 선호도를 가지고 있다고 가정하였다. 제안하는 방법의 비용 함수는 다음과 같다.

$$\alpha\left(\| R_u^I - \hat{R}_u^I \|^2 \right) + \beta\left(\| R_u^{I^C} - \hat{R}_u^{I^C} \|^2 \right) + \frac{\lambda}{2} \| W \|_F^2 \quad (11)$$

여기에서 W는 신경망에 포함된 가중치 행렬을 나타낸다. 그림 5와 그림 6은 본 논문에서 제안하는 방법을 통해 평점을 예측하는 과정을 도식화한 그림이다.

그림 5는 사용자 기준의 평점 패턴을 입력 및 예측하는 모델을 나타낸다. 사용자 3이 부여한 평점 벡터가 {5, 1, r_{33}, 2, 4}일 때, r_{33}을 예측하는 과정이다. r_{33}은 사용자 3이 상품 3에 대해 평점을 부여하지 않은 결측 값을 의미한다. 이는 0으로 대체되어 {5, 1, 0, 2, 4}로 입력되고, 은닉층을 지나 가중치를 출력하

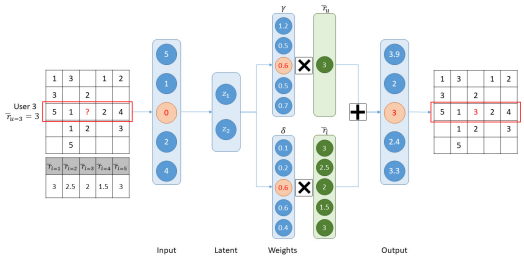

그림 5. 제안하는 모델을 통해 사용자 u의 평점을 예측하는 과정
Fig. 5. The process of predicting a rating of user u through the proposed model

게 된다. 가중치는 사용자 혹은 상품의 성질에 따라 두 평점 평균에 대한 최적의 반영 비율이 다름을 가정하였기 때문에, 하나의 평점은 두 개의 가중치를 갖게 된다. 예측하고자 하는 평점인 r_{33}에 대해서도 마찬가지로 가중치 벡터 안의 주홍색 원으로 표시된 0.6, 0.6 이렇게 두 개의 가중치가 출력된다. 출력된 가중치는 각각 사용자 3의 평점 평균인 3, 상품 3의 평점 평균인 2와 곱해진다. 최종적으로 r_{33}은 가중치 0.6과 사용자 3의 평점 평균인 3의 곱, 가중치 0.6과 상품 3의 평점 평균인 2의 곱의 합인 3으로 예측된다. 이로써 사용자 3이 부여한 평점 벡터는 {5, 1, 3, 2, 4}로 예측된다.

그림 6은 상품 기준의 평점 패턴을 입력 및 예측하는 모델을 나타낸다. 상품 3이 부여받은 평점 벡터가 {r_{13}, 2, r_{33}, 2, r_{53}}일 때, 평점을 예측하는 과정이다. 각 r_{13}, r_{33}, r_{53}는 상품 3이 사용자 1, 3, 5에 의해 부여받지 못한 결측 값을 의미한다. 이들은 0으로 대체되어 {0, 2, 0, 2, 0}로 입력되고, 은닉층을 지나 가중치를 출력하게 된다. 가중치는 사용자 혹은 상품의 성질에 따라 두 평점 평균에 대한 최적의 반영 비

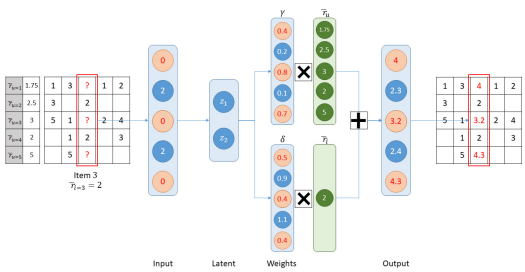

그림 6. 제안하는 모델을 통해 상품 i의 평점을 예측하는 과정
Fig. 6. The process of predicting a rating of item i through the proposed model

www.dbpia.co.kr

율이 다름을 가정하였기 때문에, 하나의 평점은 두 개의 가중치를 갖게 된다. 예측하고자 하는 평점인 r_{13}, r_{33}, r_{53}에 대해서도 마찬가지로 각각 가중치 벡터 안의 주홍색 원으로 표시된 두 개씩의 가중치가 출력된다. 출력된 가중치는 각각 사용자와 상품의 평점 평균에 곱해진다. 최종적으로, r_{13}은 가중치 0.4와 사용자 1의 평점 평균인 1.75의 곱, 가중치 0.5와 상품 3의 평점 평균인 2의 곱의 합인 4로 예측된다. 같은 방법으로 r_{33}, r_{53}은 각각 3.2, 4.3으로 예측된다. 이로써 상품 3이 부여받은 평점 벡터는 {4, 2, 3.2, 2, 4.3}으로 예측된다.

Ⅳ. 실험 결과

본 논문에서 제안하는 방법을 검증하기 위해 공개된 데이터셋인 MovieLens[17]를 사용한다. MovieLens는 영화의 평점에 대한 데이터셋이며, 사용자는 영화에 1, 2, 3, 4, 5점을 부여하게 된다. 본 논문에서는 MovieLens-100K와 MovieLens-1M을 사용한다. MovieLens-100K는 943명의 사용자가 1,682개의 영화에 대하여 100,000개의 평점을 부여한 데이터셋이고, MovieLens-1M은 6,040명의 사용자가 3,952개의 영화에 대하여 1,000,208개의 평점을 부여한 데이터셋이다. 아래 표는 각 데이터셋에 대한 세부사항이다.

제안하는 방법의 성능은 Top-N 추천 시스템[18]에서 사용하는 정밀도, 재현율, F-measure, nDCG 측면에서 검증하였다[19]. Top-N 추천 시스템은 사용자에게 적절할 것으로 예측되는 N개의 상품을 추천하는 것을 의미한다. 본 논문에서는 모든 결측 값에 대한 예측 평점을 출력하며, 각 사용자에 대해 평점을 부여하지 않은 상품 중 예측된 평점이 높은 순으로 N개를 추천한다. 한편, 사용자 u에 대해 정밀도, 재현율, F-measure는 각각 다음과 같다.

표 1. MovieLens의 세부사항
Table 1. The details of MovieLens

	MovieLens-100K	MovieLens-1M
The number of ratings	100,000	1,000,208
Sparsity	95%	96%
The number of users	943	6,040
The number of items	1,682	3,952
Range of ratings	{1, 2, 3, 4, 5}	{1, 2, 3, 4, 5}

$$P_u@N = \frac{|rel_u \cap rec_u|}{|rec_u|} \tag{12}$$

$$R_u@N = \frac{|rel_u \cap rec_u|}{|rel_u|} \tag{13}$$

$$F_u@N = \frac{2 \times P_u@N \times R_u@N}{P_u@N + R_u@N} \tag{14}$$

여기에서 $P_u@N$, $R_u@N$, $F_u@N$은 N개의 상품을 추천할 경우 각각 정밀도, 재현율, F-measure를 의미한다. 여기에서 rec_u는 사용자 u에게 추천한 N개의 상품의 집합을, rel_u는 사용자 u에게 적절한 상품의 집합을 나타낸다. 또한, 해당 데이터셋에서 적절한 상품은 사용자가 부여할 수 있는 최고 점수인 5점을 부여한 상품임을 의미한다. 즉, 정밀도는 추천한 상품 중 적절히 추천된 상품의 비율을, 재현율은 사용자에게 적절한 상품 중 추천된 상품의 비율을 의미한다. 또한, F-measure는 두 지표의 조화 평균을 나타낸다. nDCG는 예측 평점의 순위에 따른 상품과 적절한 상품의 상관을 의미한다. i_k를 추천한 상품 중 예측 평점이 높은 상위 k번째 상품이라 할 때, i_k가 사용자에게 적절한 상품인지를 의미하는 지시 함수를 y_k라 하면 nDCG는 $Y = \{y_1, \ldots, y_N\}$를 이용해 계산될 수 있다. 따라서, 사용자 u에 대한 nDCG는 다음과 같이 표현된다.

$$nDCG_u@N = \frac{DCG_u@N}{IDCG_u@N} \tag{15}$$

$$DCG_u@N = \sum_{k=1}^{N} \frac{2^{y_k} - 1}{\log_2(k+1)} \tag{16}$$

$$IDCG_u@N = \sum_{k=1}^{|rel_u|} \frac{2^{y_k} - 1}{\log_2(k+1)} \tag{17}$$

nDCG는 예측 평점의 순위가 높은 상품이 적절한 경우 1에 가깝고, 반대의 경우 0에 가깝다. 본 논문에서의 성능평가 지표는 모든 사용자에 대한 각 성능 평

www.dbpia.co.kr

가 지표의 평균을 사용하였다. 모든 성능 평가는 5겹 교차 검증을 통해 진행되었다. 본 논문에서 제안하는 방법의 성능은 비용함수에 결측 값을 반영하지 않은 DAE, 이러한 DAE와 같은 구조를 가지며 비용함수에 결측 값을 반영하는 DAE 등과 비교하였다. 결측 값인 NA (not available)를 반영하지 않은 것은 "without NA", 결측 값을 0으로 반영한 것은 "with NA"로 표기하였다.

그림 7과 8은 각각 MovieLens-100K 데이터셋을 사용 시 본 논문에서 제안하는 방법과 기존에 제안된 DAE 기반 협업 필터링 방법의 추천 정확도를 비교한 것이다. 또한, 그림 9와 10은 MovieLens-1M 데이터셋 하에서 측정한 추천 정확도를 나타낸다. 그림 7과 9는 사용자 정보인 R_u를, 그림 8과 10은 상품의 정보인 R_i를 입력으로 한다. 구체적으로, 기존에 제안된 DAE 기반 협업 필터링 (DAE without NA), 비용 함수에 결측 값을 0으로 반영한 DAE 기반 협업 필터링 (DAE with NA) 그리고 본 논문에서 제안하는 방법 (Our approach) 이렇게 세 가지 방법을 비교하였다. (a), (b), (c), (d)는 각각 정밀도, 재현율, F-measure, nDCG에 대한 결과를 보여준다.

실험 결과로부터 제안하는 방법이 기존에 제안된 DAE 기반 협업 필터링 (DAE without NA) 방법에 비해 각 정밀도, 재현율, F-measure, nDCG의 측면에서 평균적으로 약 130%, 165%, 140%, 173% 향상된 성능을 보였다. 특히, 사용자 정보를 활용하여 MovieLens-1M 데이터셋에 적용하였을 때는 최대 210% 향상된 성능을 보였다. 전반적으로, 정밀도는 추천하는 상품 수인 N이 증가함에 따라 점차 감소하는 경향을 보이며, 재현율과 nDCG는 증가하는 경향이 있는 것으로 나타났다. 이러한 경향은 N이 증가함

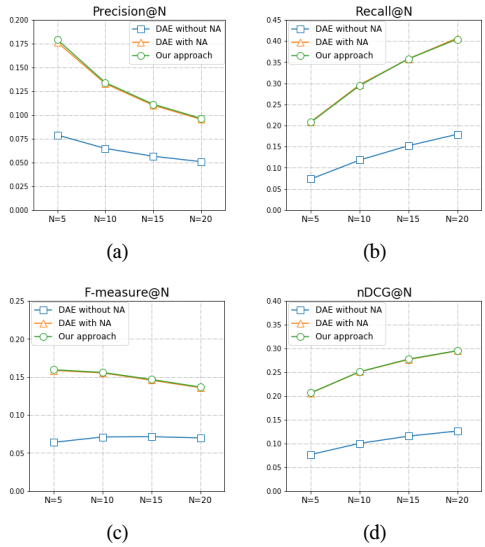

그림 8. 상품 정보를 이용하는 모델들의 Top-N 추천 정확도 (MovieLens-100K)
Fig. 8. The accuracy of top-N recommendations with ratings of items (MovieLens-100K)

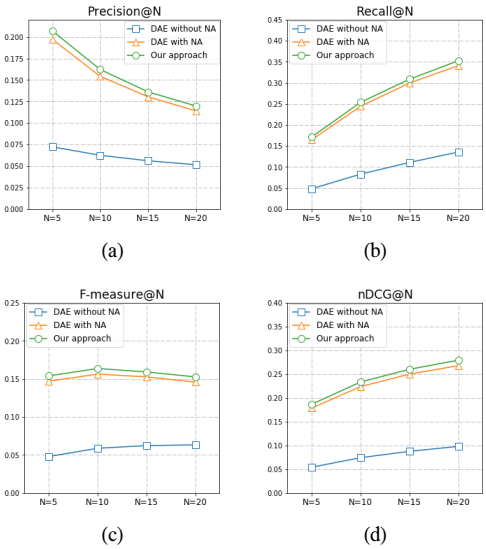

그림 7. 사용자 정보를 이용하는 모델들의 Top-N 추천 정확도 (MovieLens-100K)
Fig. 7. The accuracy of top-N recommendations with ratings of users (MovieLens-100K)

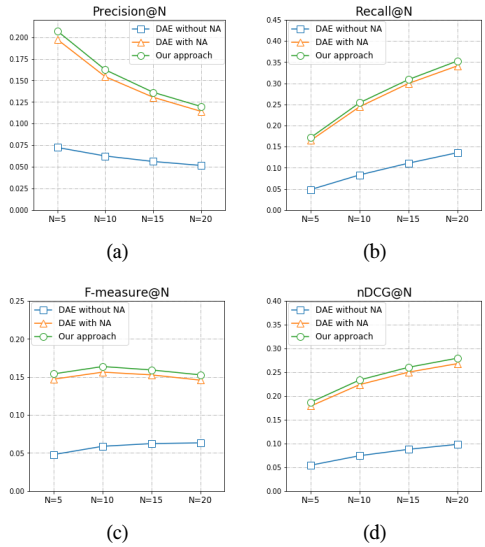

그림 9. 사용자 정보를 이용하는 모델들의 Top-N 추천 정확도 (MovieLens-1M)
Fig. 9. The accuracy of top-N recommendations with ratings of users (MovieLens-1M)

www.dbpia.co.kr

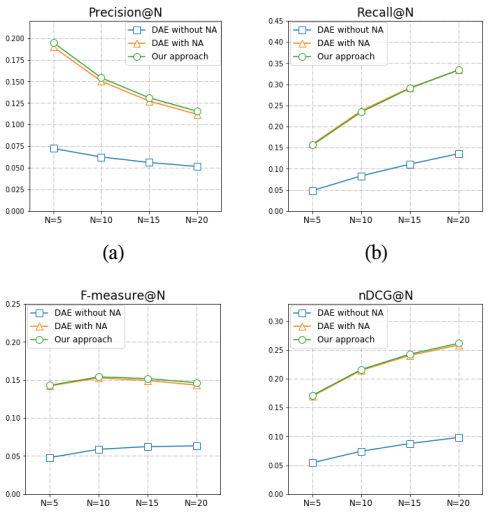

그림 10. 상품 정보를 이용하는 모델들의 Top-N 추천 정확도 (MovieLens-1M)
Fig. 10. The accuracy of top-N recommendations with ratings of items (MovieLens-1M)

에 따라 추천하는 상품 중 사용자에게 적절하게 추천된 비율은 줄어들며, 사용자에게 적절한 상품 중 추천된 상품의 비율은 늘어나는 것을 의미한다. 또한, 상대적으로 재현율과 nDCG는 정밀도에 비해 전반적으로 높은 수치를 보였다. 한 편, F-measure는 각 경우에 따라 상이한 결과를 보였다. 본 논문에서 제안하는 방법은 MovieLens-100K 데이터셋에 대해 N=5에서, MovieLens-1M 데이터셋에 대해 N=10에서 가장 높은 F-measure를 보였다.

한편, 제안하는 방법은 비용 함수에 결측 값을 0으로 반영한 DAE 기반 협업 필터링 (DAE with NA) 방법과 비교 시 약 5% 증가된 정확도를 보였다. 또한, 모든 실험에서 결측 값을 DAE의 비용 함수로 반영하지 않은 방법이 가장 낮은 정확도를 보였다. 이는 결측 값을 가장 낮은 선호도로 가정하지 않고, 비용 함수에서 무시하는 것이 상위 N개의 상품을 추천해주는 것에 있어서 성능이 저하됨을 의미한다.

V. 결 론

본 논문에서는 추천 시스템을 위한 협업 필터링 방법으로써, 평점 정보를 활용한 새로운 DAE를 제안하였다. 제안한 방법은 각 평점이 사용자와 상품의 평점 평균들의 선형 결합으로 이루어져 있음을 고려하였다. 또한 결측 값을 부정적인 사전 선호도의 표현으로 가정하고, 기존 DAE 기반 협업 필터링 방법과 달리 비용 함수에 결측 값을 포함하였다. 그 결과 기존 방법에 비해 모든 정확도 면에서 높은 성능을 보이는 것을 확인하였다. 특히, 사용자 정보를 이용하여 MovieLens-1M 데이터셋에 적용하였을 때, nDCG의 측면에서 최대 210% 향상된 성능을 보였다.

최근에는 추천 항목들이 서로 얼마나 다른지를 의미하는 추천의 다양성, 사용자가 경험한 것들과 비교하여 얼마나 다른지를 의미하는 참신성이 대두되고 있다.[20] 향후에는 이러한 문제들에 대해서도 고려할 수 있는 협업 필터링 방법에 대한 연구가 필요하다.

References

[1] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: item-to-item collaborative filtering,” in *IEEE Internet Computing*, vol. 7, no. 1, pp. 76-80, Jan. 2003.

[2] Netflix Prize, Retrieved Jul., 30, 2018, from http://www.netflixprize.com

[3] C. A. G.-Uribe and N. Hunt, “The Netflix recommender system: algorithms, business value, and innovation,” *ACM TMIS*, vol. 6, no. 4, Jan. 2015.

[4] B. W. Seo, *The evolution of contents recommendation algorithm*(2016), Retrieved Jul., 31, 2018, from http://www.kocca.kr/insight/vol05/vol05_04.pdf

[5] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*, Cambridge University Press, 2010.

[6] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proc. ICML*, pp. 1096-1103, Helsinki, Finland, Jul. 2008.

[7] C. C. Aggarwal, *Recommender systems: The textbook*, Springer International Publishing, 2010.

[8] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in *Proc. 10th Int. Conf. WWW*, pp. 285-295, Hongkong, Hongkong, May 2010.

www.dbpia.co.kr

[9] S. Zhang, W. Wang, J. Ford, F. Makedon, and J. Pearlman, "Using singular value decomposition approximation for collaborative filtering," in *Proc. 7th IEEE Int. Conf. E-Commerce Technol.*, pp. 257-264, Washington DC, USA, 2005.

[10] M. G. Vozalis and K. G. Margaritis, "Using SVD and demographic data for the enhancement of generalized collaborative filtering," *Information Sci.*, vol. 177, pp. 3017-3037, 2007.

[11] M. A. Kramer, "Autoassociative neural networks," *Comput. Chem. Eng.*, vol. 16, no. 4, pp. 313-328, Apr. 1992.

[12] Y. Ren, et al., "AdaM: adaptive-maximum imputation for neighborhood-based collaborative filtering," in *Proc. IEEE Int. Conf. Advances in Soc. Netw. Anal. and Mining*, pp. 628-635, 2013.

[13] W. S. Hwang, J. Parc, S. W. Kim, J. Lee, and D. Lee, "Told you i didn't like it": Exploiting uninteresting items for effective collaborative filtering," *2016 IEEE 32nd ICDE*, pp. 349-360, Helsinki, 2016.

[14] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie, "AutoRec: autoencoders meet collaborative filtering," in *Proc. Int. Conf. WWW*, pp. 111-112, Florence, Italy, May 2015.

[15] F. Strub and J. Mary, "Collaborative filtering with stacked denoising autoencoders and sparse inputs," *NIPS Workshop on Machine Learning for eCommerce*, Montreal, Canada, Dec. 2015.

[16] Z. Qian, L. Qing, and Z. Xue, "A collaborative filtering recommendation algorithm based on correlation and improved weighted prediction," in *Proc. ICEE*, pp. 1-3, Shanghai, China, May 2011.

[17] Retrieved Jul. 30, 2018, from https://grouplens.org/datasets/movielens/

[18] J. Lee, et al., "Improving the accuracy of top-N recommendation using a preference model," *Information Sci.*, pp. 290-304, 2016.

[19] J. W. Ha, et al., "Top-N recommendation through belief propagation," in *Proc. ACM Int. Conf. Inf. and Knowledge Management*, pp. 2343-2346, 2012.

[20] S. J. Yu, "A study of improvement of individual item diversity in collaborative filtering-based recommendation," *J. KIIT,* vol. 14, no. 8, pp. 89-94, Aug. 2016.

김 현 진 (Hyun-Jin Kim)

2017년 2월 : 단국대학교 수학, 컴퓨터과학과 졸업

2017년 2월~현재 : 단국대학교 데이터사이언스학과 석사과정

<관심분야> 기계학습, 딥러닝, 추천시스템

신 동 진 (Dong-Jin Shin)

2016년 2월 : 단국대학교 정보통계학과 졸업

2018년 2월 : 단국대학교 정보통계학과 석사

<관심분야> 기계학습, 딥러닝, 통계학, 컴퓨터비전

www.dbpia.co.kr

신 원 용 (Won-Yong Shin)

2002년 : 연세대학교 기계전자공학부 학사
2004년 : KAIST 전자전산학과 석사
2008년 : KAIST 전자전산학부 박사
2009년 5월~2011년 : 10월 Harvard University Postdoctoral Fellow
2011년 10월~2012년 2월 : Harvard University Research Associate
2012년 3월~2017년 2월 : 단국대학교 컴퓨터학과 조교수
2017년 3월~현재 : 단국대학교 컴퓨터학과 부교수
<관심분야> 정보이론, 통신이론, 신호처리, 빅데이터분석, 소셜네트워크분석

황 창 하 (Changha Hwang)

1982년 2월 : 경북대학교 수학교육과 졸업
1984년 2월 : 서울대학교 통계학과 석사
1991년 2월 : 미시간대학교 통계학과 박사
2005년 3월~현재 : 단국대학교 응용통계학과, 데이터사이언스학과 교수
<관심분야> 기계학습, 계산통계

www.dbpia.co.kr