Check for updates

Journal of KIIT. Vol. 21, No. 7, pp. 1-6, Jul. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.7.1

베스트셀러 도서 예측을 위한 머신러닝 알고리즘 성능평가

유지은*[1], 조솔비*[2], 유석종**

Performance Evaluation of Machine-Learning Algorithms for Bestseller Book Prediction

Ji-Eun Yu*[1], Sol-Bee Cho*[2], and Seok-Jong Yu**

요 약

베스트셀러 도서는 독자들이 책을 선택하는 가장 보편적인 방법이며, 이러한 이유로 베스트셀러의 예측과 선정은 출판 시장에서 중요한 마케팅 전략 지표이다. 본 연구에서는 도서의 메타 데이터를 활용하여 베스트셀러 순위 200위 내 유지 여부와 판매 지수 구간을 예측하는 모델을 제안하고, 다양한 머신러닝 알고리즘의 성능을 비교 평가하고자 한다. 이를 위하여 Yes24 사이트의 월간 베스트셀러 데이터를 크롤링하여 수집하고, 각 데이터 속성에 대해 적절한 전처리를 수행하였다. 순위 유지 여부 예측을 위해 다양한 분류 알고리즘을 활용하였고, 최종적으로 각 알고리즘의 예측 성능을 평가한 결과, 다중 퍼셉트론, CatBoost, 랜덤 포레스트의 순서로 정확도가 높게 나타났다. 본 연구는 베스트셀러 순위 유지 여부 예측 문제에 대해 주요 분류 알고리즘의 수행 성능을 종합적으로 비교했다는데 의미가 있다. 그러나 한계점으로 리뷰 수, 평점 등에 의존하는 예측 방법에서는 데이터가 부족한 신간 도서에서 cold start 문제를 극복하기 어려웠으며, 이에 대한 후속 보완 연구의 필요성을 제안한다.

Abstract

Bestsellers are the most common way for readers to choose books, and for this reason, the prediction and selection of bestsellers is an important marketing strategy indicator in the publishing market. In this study, we propose a model that predicts whether or not to remain in the top 200 bestseller rankings and sales index sections using metadata from bestsellers, and compare and evaluate the performance of various machine learning algorithms. To this end, monthly bestseller data on the Yes24 site were crawled and collected, and appropriate preprocessing was performed for each data attribute. Various classification algorithms were used to predict whether to maintain the ranking, and as a result of finally evaluating the prediction performance of each algorithm, the accuracy of MLP, CatBoost, and random forest was high. This study is meaningful in that it comprehensively compared the performance performance of various classification algorithms for predicting whether to maintain the bestseller ranking. However, in models that rely on the number of reviews and ratings as limitations, it was difficult to overcome the cold start problem in new books that lacked data, and the need for follow-up supplementary research is proposed.

Keywords

machine learning, prediction model, random forest model, ensemble model

* 숙명여자대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcid.org/0009-0001-9325-731X
- ORCID[2]: https://orcid.org/0009-0007-5777-1403
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: May 09, 2023, Revised: Jun. 27, 2023, Accepted: Jun. 30, 2023
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Women's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

최근 전세계 도서 출판 시장은 코로나 19이후 전반적인 하락세를 보이면서 급격한 침체를 겪고 있다[1]. 출판사들은 출판 시장을 활성화시키고 판매량을 높이기 위해 다양한 전략과 방법을 모색하고 있으며, 이 중 베스트셀러에 대한 분석 및 예측이 대표적이다. 베스트셀러 순위에 진입한 도서는 해당 출판사의 인지도를 높일 수 있으며, 대중에게 더 쉽게 노출되어 궁극적으로 책의 판매량을 높이는 등의 마케팅 효과가 있다[2][3].

출간 도서의 메타 데이터는 책의 특성을 정량적, 정성적으로 분석할 수 있는 중요한 자료이다. 메타 데이터는 책의 제목, 저자, 출판사, 장르 등의 정보를 포함하고 있으며, 기존 연구에서는 주로 작가, 평점, 리뷰에 대한 데이터를 활용하여 도서 판매량을 예측하였다. 본 연구에서는 다양한 머신 러닝 알고리즘을 활용하여 출간 도서가 베스트셀러 200위 내 유지 여부와 판매 지수 구간을 예측하는 모델을 제안하고자 한다. 이를 위해 Yes24의 월간 베스트셀러 페이지에서 각 도서의 메타데이터(제목, 저자, 출판사, 등)를 수집하고, 전처리하여 분석하였다. 구체적인 평가 지표로 특정 도서의 일정 기간 동안 베스트셀러 순위 200위 내 유지 여부와 판매 지수 구간을 예측하는 모델을 제안하고 각 머신러닝 알고리즘의 성능을 비교하였다. 본 논문의 구성은 다음과 같다. 2장은 관련 연구로 도서 판매량 예측에 대한 기존 연구를 리뷰하고, 주요 분류 알고리즘에 대해 소개한다. 3장에서는 데이터 수집과 전처리, 사용한 머신러닝 알고리즘에 대한 내용을 기술하고, 4장과 5장에서는 각각 성능 비교 실험 결과를 제시하고, 결론을 맺는다.

Ⅱ. 관련 연구

2.1 빅데이터를 활용한 도서 판매 예측

이종욱의 연구[4]에서는 베스트셀러 도서의 순위가 공공도서관에서의 평균 대출 수에 어떠한 영향을 미치는 지에 대한 연구를 수행하였다. 2018년~2019년까지 총 104주 간의 공공도서관 대출 데이터셋과 Yes24 사이트의 동일 기간의 주간 베스트셀러 데이터를 사용하였으며, 베스트셀러의 순위가 사람들의 도서관 대출에 영향을 미치고 있다는 점을 확인하였다. 결론적으로 베스트셀러 순위는 이용자의 요구를 예측하고, 도서관의 장서 구성이나 출판사의 마케팅 전략 활용에 유용한 정보임을 알 수 있다.

김나연의 연구[5]에서는 2016년~2020년까지 총 5년간의 해외 출간 도서 데이터를 활용하여 누적 5천부 이상 판매 여부를 예측하는 연구를 진행하였다. 판매 예측을 위해 작가, 출간 국가, 평점 평균, 평점 참여 수, 번역 만족도 등의 변수를 이용하였고 LightGBM 모델이 가장 좋은 예측 성능 모델임을 제시하였다. 해당 연구에서는 제목, 가격, 출판일, 장르 등 독자들의 도서 구매에 영향을 미치는 다른 메타 데이터를 활용하지 않았으며 5천 부 이상의 판매 여부에 대해서만 예측하였다는 점에서 본 연구와 차이가 있다.

2.2 랜덤 포레스트

랜덤 포레스트(Random forest)는 머신러닝 분야에서 널리 사용되는 앙상블 모델 중 하나로 다수의 결정 트리를 생성하고 이를 조합하여 더욱 강력한 모델을 만드는 방식이다[6]. 결정 트리 생성 시 특성(Feature)을 무작위로 선택하여 각각의 트리가 독립적으로 학습되도록 하며, 학습 데이터에서 무작위로 복원 추출하여 각 결정 트리를 학습시켜 오버피팅(Overfitting)을 방지하고 예측 성능을 높인다.

2.3 그래디언트 부스팅

그래디언트 부스팅(Gradient boosting)은 2001년 Friedman에 의해 처음 제안된 모델로, 결정 트리 여러 개를 결합한 모델이다. 이전 모델이 예측하지 못한 데이터에 대해 새로운 모델을 반복적으로 학습하여 이전 모델의 오차를 보완하며 오차 함수가 최소화되도록 모델을 학습한다. 그래디언트 부스팅은 오버피팅을 예방할 수 있지만 학습 속도가 느리고 하이퍼파라미터 설정의 어려움이 발생한다[7].

CatBoost, XGBoost, LightGBM, 알고리즘이 대표적인 그래디언트 부스팅 라이브러리라고 할 수 있다[8]-[10].

2.4 CatBoost

CatBoost는 GBDT(Gradient Boosting Decision Tree) 기반 알고리즘 중 하나로, 결정 트리 여러 개를 연속으로 연결한 앙상블 모델의 일종이다. 다른 GBDT 모델과는 달리, 카테고리형 데이터 처리에 효과적인 Ordered target encoding 기법을 사용한다. 또한 오버피팅 방지를 위해 스태킹(Stochastic gradient boosting)과 폴딩 기법을 사용하여 모델의 성능을 높인다[8].

2.5 다중 퍼셉트론

다중 퍼셉트론(MLP)은 인공신경망의 일종으로, 입력층, 은닉층, 출력층으로 구성된 모델이다. 각 층은 여러 개의 뉴런으로 이루어져 있으며, 노드 간의 가중치와 편향을 학습하여 입력층과 출력층 사이의 관계를 모델링한다. MLP는 비선형 관계를 근사할 수 있어 복잡한 문제에 대해서도 성능이 우수하다는 장점이 있지만 최적의 하이퍼파라미터를 찾는 것이 어렵고 오버피팅의 확률이 높다는 단점이 있다.

III. 베스트셀러 예측을 위한 머신러닝 알고리즘 성능 비교

3.1 데이터 수집

본 연구에서는 월별로 1,000위까지의 순위를 제공하는 Yes24 사이트의 월간 베스트셀러 페이지를 웹크롤링하여 2018년~2022년까지 5년간의 데이터를 수집하였다. 크롤링은 Python 라이브러리인 BeautifulSoup을 활용하여 각 도서의 제목, 작가, 출판사, 출판일, 판매 지수, 가격, 리뷰 개수, 평점, 장르 데이터를 총 59,290개 수집하였다. Pandas와 Numpy를 사용하여 하나의 데이터프레임으로 병합하여 실험 데이터셋을 구축하였다.

표 1. 수집한 베스트셀러 데이터 정보
Table 1. Information of collected bestseller data

Field	Value
Period	2018. 01 ~ 2022. 12
Size	7.77 MB
Number of bestseller data	59,290

3.2 베스트셀러 도서의 순위 변동 분석

도서의 순위 변동 유형과 기간별로 몇 권의 도서가 베스트셀러 상위 200위를 유지했는지 확인하기 위한 사전 실험을 진행하였다. 베스트셀러 순위 내 도서 중 '언어의 온도'와 '코스모스', 2권의 책에 대해 순위 변동 양상을 그림 1과 같이 그래프로 나타내었다. 2016년 8월에 출간된 '언어의 온도'는 꾸준히 상위권을 유지하다 점점 하락하여 2020년 6월 이후부터는 200위 권 밖에 랭크되는 양상을 보이고 있다. 2006년 12월에 발매된 '코스모스'는 15년 동안 베스트셀러 순위를 유지하고 있다.

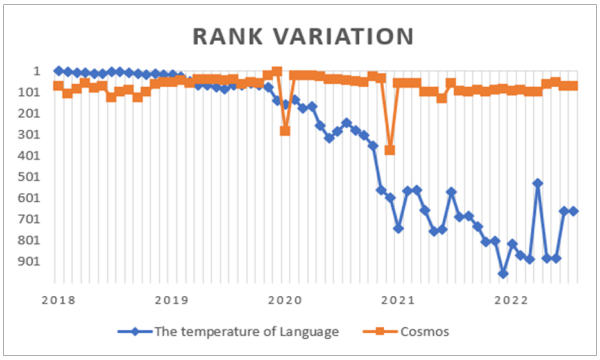


그림 1. 베스트셀러 도서의 순위 변동 양상
Fig. 1. Ranking variation of bestseller books

다음은 상위 200위 이내에 머문 도서의 수에 대한 분석 실험 결과이다.

표 2. 기간별 순위 200위 이내 유지한 도서의 수
Table 2. Number of books ranked in top 200 by period

Number of months ranked in top 200	Number of books
0	8,594
1	1,599
2 ~ 3	1,015
4 ~ 6	357
more than 6	309

실험 결과, 200위 내에 진입하지 못한 도서가 전체의 72% (8,594권)에 해당되었다. 전체의 약 3%만이 6개월 이상 200위 이내를 유지하였고 순위 유지 기간은 매우 짧았다. 베스트셀러 상위권을 꾸준히 유지하는 도서는 매우 적었고 변화 속도가 심했다.

3.3 데이터 분석 및 실험 과정

수집된 도서 데이터에서 중복 및 판매 지수 결측 행을 제거하여 총 11,874개의 데이터셋을 생성하였다. 모델 학습에 사용된 데이터프레임은 표 3과 같다. weighted_rating은 리뷰 개수에 가중치를 부여한 평점 데이터이다. title은 제목에 포함된 단어 데이터에 대한 tf-idf 계산값이다. genre, publisher, author의 범주형 데이터는 원 핫 인코딩(One-hot encoding)으로 수치형 데이터 변환 후 사용하였다. 범주형 데이터를 PCA 기법으로 차원 축소 후 실험에 사용하였다.

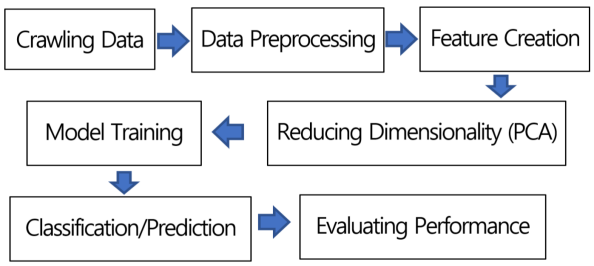


그림 2. 데이터 처리 및 실험 과정
Fig. 2. Data processing and experimental procedure

표 3. 모델 학습에 사용할 데이터 프레임
Table 3. Information of collected bestseller data

Features	Meaning
reviewnum	Number of reviews
weighted_rating	Rating data weighted according to the number of reviews
price	Selling price
diff_month	Number of months passed since publication
title	Tf-idf value for title of book
genre	The genre of books
publisher	The publisher of books
author	The author of books

본 실험의 목표는 특정 도서의 베스트셀러 200위권 내 3개월/6개월 이상 유지 여부 예측이다. 학습 데이터가 200위 순위권 내에 한 번도 진입하지 못한 경우는 클래스 0, 3개월 이상 진입한 도서는 클래스 1로 분류하였다. 또한 4단계의 판매 지수 구간을 예측하는 실험을 K-means 기법을 사용하여 진행하였다. 구간별 데이터에 발생한 불균형 분포를 해소하기 위해 과소 구간 데이터에 가중치를 두는 ADASYN 기법으로 데이터를 오버샘플링하였다.

3.4 실험에 사용한 머신러닝 알고리즘

표 4는 실험에 사용한 분류 알고리즘의 목록이고, 표 5는 GridSearchCV 수행 결과를 참고하여 설정한 각 모델별 하이퍼파라미터 내용이다.

표 4. 실험에 사용한 분류 알고리즘 목록
Table 4. Experimental classification algorithms

Symbol	Classification model
A	Decision tree
B	Random forest
C	Gradient boosting
D	XGBoost
E	LightGBM
F	CatBoost
G	Ensemble(Random forest+Gradient boosting)
H	MLP

표 5. 분류 알고리즘 별 하이퍼파라미터
Table 5. Hyperparameters for classification algorithms

Model	Hyperparameter
A	max_depth = 10, min_samples_leaf = 10
B	n_estimators = 100, max_depth = 10
C	n_estimators = 50, max_depth = 5
D	n_estimators = 50, max_depth=3
E	'objective': 'multiclass', 'metric': 'multi_logloss', 'num_leaves': 15, 'learning_rate': 0.01, 'feature_fraction': 0.5
F	iterations=1000, learning_rate=0.05, depth=3, loss_function='MultiClass'
G	rf : n_estimators=100, max_depth=10, gb : n_estimators=100, max_depth=3
H	hidden_layer_sizes=(20,20), activation='relu', solver='adam', max_iter=1000

Ⅳ. 실험 및 성능 평가

4.1 실험 환경

본 실험에서 사용된 데이터 분석과 예측 모델은 Pandas, Numpy, scikit-learn, konlpy 등의 라이브러리를 이용하여 구현하였다. 또한 실험을 위한 훈련과 평가를 위한 데이터는 크롤링 및 전처리를 통해 수집한 총 11,874개의 데이터를 랜덤으로 8:2 비율로 나누어 각각의 데이터셋을 구축하였다.

4.2 실험 결과 및 성능 평가

각 실험 모델을 사용하여 Yes24 베스트셀러 데이터를 활용하여 3개월 이상과 6개월 이상 베스트셀러 200위 내에 머무는지에 대한 분류 예측을 각각 수행하고 각 모델의 예측 정확도를 비교하였다. 베스트셀러 순위 200위 내 유지 여부에 대한 실험 결과, 훈련 및 테스트 정확도(train/test accuracy) 각각 0.992과 0.96으로 다중 퍼셉트론 모델이 가장 높았다. 랜덤 포레스트와 CatBoost 모델의 정확도는 0.937과 0.930의 높은 정확도를 보였다. 다중 퍼셉트론으로 학습 데이터의 20%를 검증 데이터로 사용하여 학습한 결과, epoch 20 시점에서 손실 값이 크게 상승하여 오버피팅의 가능성을 확인하였다.

6개월 이상 상위 200위 내 유지 여부에 대한 실험 결과에서도 다중 퍼셉트론 모델이 가장 높았으며, 랜덤 포레스트와 CatBoost의 정확도가 0.969, 0.971로 그다음 높은 정확도를 보였다. 판매 지수 구간 예측 실험에서도 다중 퍼셉트론 모델이 높은 정확도를 보여주었다.

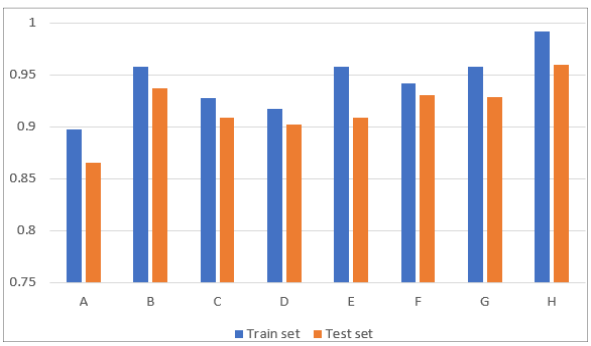


그림 3. 순위 예측 정확도 (3개월 이상 200위 내 유지)
Fig. 3. Prediction accuracy by experiment model (3 months ranked in the top 200)

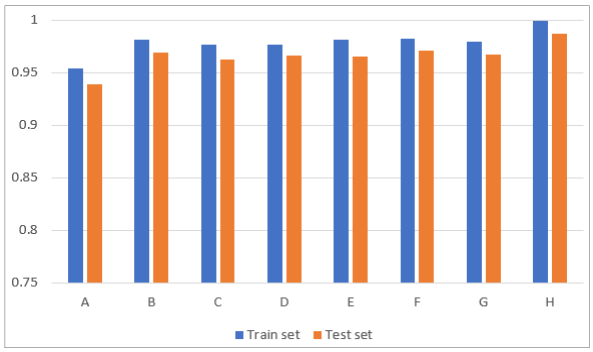


그림 4. 순위 예측 정확도 (6개월 이상 200위 내 유지)
Fig. 4. Prediction accuracy (Over 6 months more ranked in top 200)

2023년 이후 출판되어 순위 데이터가 없는 신간 도서 3권('세이노의 가르침', '메리골드 마음 세탁소', '도둑맞은 집중력')에 대해 베스트셀러 순위 200위 내 유지 여부 실험을 진행하였다. A, C, D, F 실험 모델은 3권 모두 순위를 유지하지 못할 것으로 예측했으며 B, E, G, H 모델은 '세이노의 가르침'만 순위를 유지할 것으로 예측하였다.

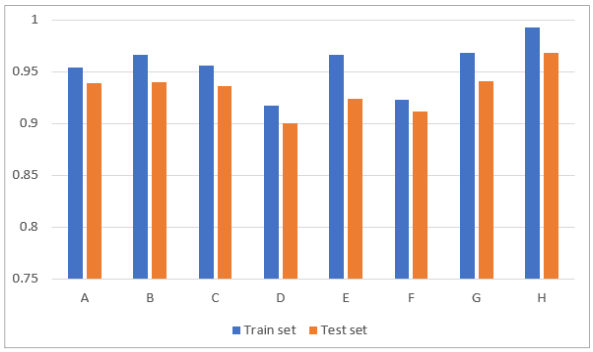


그림 5. 판매 지수 구간 예측 정확도
Fig. 5. Prediction accuracy (Sales classification)

Ⅴ. 결론 및 향후 연구

본 연구는 도서 메타데이터를 활용하여 베스트셀러 순위 유지 여부와 판매 지수 구간 예측을 목표로 하였다. 총 8가지 알고리즘을 사용하여 분류 예측 정확도를 비교하였다. 다중 퍼셉트론 모델의 정확도가 가장 좋았으며, 랜덤 포레스트와 CatBoost 모델이 그 다음으로 나타났다. 다중 퍼셉트론 모델은 오버피팅의 가능성이 있었으며, 랜덤 포레스트와 CatBoost 모델은 정확도가 평균 94.5%로 적절한 성능을 보여주었다.

본 연구의 한계점으로 신간 도서의 분류 예측 시 연관 데이터 부족으로 인한 cold start 문제가 발생할 수 있다. 추후 연구에서 회귀 모델을 활용한 성능 비교 연구로 확장되기를 기대한다.

References

[1] Q. Wei, H. J. Lee, H. S. Lee, and G. W. Lee, "Global Digital Publishing Market and Concept Evolution", The Korea Digital Publishing, Society, Vol. 15, pp. 10-25, Aug. 2020. http://doi.org/10.30580/kdips.2020.15.1.11.

[2] Y. J. Nam, "A Study on the Utilization of Librarian Recommendation System and Bestseller List", JKOSIM, Vol. 38, No. 3, pp. 311-334, 2021. https://doi.org/10.3743/KOSIM.2021.38.3.311.

[3] J. D. Kim and C. O. Kim, "A Study on the Best Seller Strategy of Publishers", pp. 165-182, 2010.

[4] J. W. Lee, W. J. Kang, and J. K. Park, "The Effects of the Bestseller Ranks on Public Library Circulation: Based on Panel Data Analysis", JKOSIM, Vol. 38, No. 4, pp. 1-23, 2021. https://doi.org/10.3743/KOSIM.2021.38.4.001

[5] N. Kim, D. Kim, M. Kim, J. Jung, and H. H. Kim, "Prediction of Good Seller in Overseas sales of Domestic Books Using Big Data", Korea Information Processing Society, pp. 401-404, 2022.

[6] L. Breiman, "Random Forests", Machine Learning, pp. 5-32, Oct. 2001. https://doi.org/10.1023/A:1010933404324.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", Annals of statistics, Vol. 29, No. 5, pp. 1189-1232, Oct. 2001. https://doi.org/10.1214/aos/1013203451.

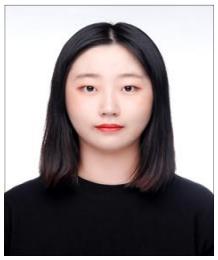
[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 785-794, Aug. 2016. https://doi.org/10.1145/2939672.2939785.

[9] D. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features", Advances in Neural Information Processing Systems, Vol. 31, pp. 6638-6648, 2018.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems, Vol. 30, 2017.

저자소개

유 지 은(Ji-Eun Yu)



2021년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심 분야 : 데이터 마이닝, 머신러닝, 딥러닝

조 솔 비(Sol-Bee Cho)

2021년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심 분야 : 딥러닝, 머신러닝

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 컴퓨터과학과(이학사)

1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)

2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)

2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수

관심분야 : 데이터마이닝, 추천시스템, 정보시각화