Journal of Convergence for Information Technology
Vol. 9. No. 6, pp. 83-90, 2019
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2019.9.6.083

협업필터링과 스태킹 모형을 이용한 상품추천시스템 개발

박성종[1], 김영민[2], 안재준[3*]
[1]연세대학교 정보통계학과 석사과정, [2]순천향대학교 빅데이터공학과 조교수, [3]연세대학교 정보통계학과 부교수

Development of Product Recommender System using Collaborative Filtering and Stacking Model

Sung-Jong Park[1], Young-Min Kim[2], Jae-Joon Ahn[3*]
[1]Student, Department of Information and Statistics, Yonsei University
[2]Assistant Professor, Department of Bigdata Engineering, Soonchunhyang University
[3]Associate Professor, Department of Information and Statistics, Yonsei University

요 약 사람들은 자신의 더 나은 선택을 위하여 끊임없이 노력한다. 이러한 이유로 추천시스템이 개발되었으며, 1990년대 초반부터 계속해서 발전하고 있다. 그 중, 협업필터링 기법은 추천시스템 분야에서 우수한 성능을 보였으며, 기계학습이 등장하면서 기계학습을 이용한 추천시스템에 관한 연구가 활발히 진행되었다. 본 연구는 앙상블 방법 중에서 스태킹 모형을 사용하여 추천시스템을 구축하며, 실제 고객의 상품 구매 데이터를 활용하여 협업필터링과 기계학습 기반 스태킹 모형으로 추천시스템을 개발하였다. 제시한 모형의 추천 성능은 기존의 협업필터링과 기계학습 기반 추천시스템과 비교하여 모형의 우수성을 확인하며, 연구결과는 스태킹 모형을 이용한 추천시스템 모형의 추천 성능이 개선됨을 확인하였다. 향후 본 연구에서 제안한 모형은 개인이나 기업이 더 나은 선택을 하여 상품을 추천할 때 도움을 줄 것으로 기대한다.

주제어 : 추천시스템, 협업필터링, 기계학습, 앙상블, 스태킹

Abstract People constantly strive for better choices. For this reason, recommender system has been developed since the early 1990s. In particular, collaborative filtering technique has shown excellent performance in the field of recommender systems, and research of recommender system using machine learning has been actively conducted. This study constructs recommender system using collaborative filtering and machine learning based on stacking model which is one of ensemble methods. The results of this study confirm that the recommender system with the stacking model is useful in aspects of recommender performance. In the future, the model proposed in this study is expected to help individuals or firms to make better choices.

Key Words : Recommender system, Collaborative filtering, Machine learning, Ensemble, Stacking

1. 서론

사람들은 다양한 선택을 함에 있어 개인적인 경험이 아닌 입소문 혹은 매체를 통한 타인의 추천에 의지하는 경향이 있다. 이러한 개인들의 심리적인 특성에 기인하여 많은 기업은 추천시스템에 집중하고 있다[1]. 추천시스템은 선호도에 따라 사람들에게 더 나은 선택을 제시한다. 이러한 추천시스템 모형들은 1990년대 초반부터 본격적으로 연구되었으며, 최근 아마존, 넷플릭스, 유튜브, 구글 그리고 페이스북 등 많은 기업에서 중요한 역할을 하고 있다[2].

*Corresponding Author : Jae-Joon Ahn(ahn2615@yonsei.ac.kr)
Received May 16, 2019 Revised May 31, 2019
Accepted June 20, 2019 Published June 28, 2019

협업필터링은 Goldberg[3]에 의해 처음으로 소개된 추천 시스템 기법으로 널리 사용되고 있다. 협업필터링은 추천대상이 되는 고객과 취향이 비슷한 고객을 선정하고 취향이 비슷한 고객들이 선호하는 아이템을 추천대상 고객에게 추천하는 사용자 기반 협업필터링[4]과 모든 아이템 간의 유사도를 기반으로 아이템 간의 선호도를 예측하여 고객에게 추천하는 아이템 기반 협업필터링 등 여러 가지 방법이 있다[5].

Lee[6]는 Movielens 데이터 셋을 이용하여 사용자 기반 협업 필터링으로 추천시스템을 구축하여, 기존에 소개되었던 사용자 기반의 알고리즘보다 좋은 성능이 나타남을 보였으며, Marlin[7]은 user rating profile 모델이라 불리는 평점 기반 협업필터링을 제안하였다. 또한, 아마존에서는 아이템 기반 협업필터링으로 추천시스템을 개발하여 높은 성능을 보이는 모델을 소개하였다. 하지만 협업필터링 알고리즘은 모든 품목 간의 유사도를 계산하여 유사도 행렬을 구축하는데 많은 처리 시간과 용량이 필요하기에 비효율적이라는 한계점이 있다[8].

최근에는 높은 추천 성능 도출을 위해 기계학습 알고리즘을 사용하여 추천시스템을 구축하는 연구들이 활발히 진행되고 있다. Oh[9]는 군집 알고리즘을 이용하여 영화 추천시스템을 제안하였으며, 연구 결과는 추천시스템의 문제인 데이터 희소성과 콜드 스타트 문제를 해결하였다. Jeong[10]은 베이지안 네트워크를 이용하여 음식 쿠폰 추천시스템을 제안하였으며, Kim[11]은 베이지안 네트워크를 이용하여 음악 추천시스템에 관한 연구를 제안하였다. Tsuji 등[12]은 도서관 대출 기록을 비롯한 여러 요인을 변수로 설정하여 기계학습 기법인 서포트벡터머신(support vector machine; SVM), 랜덤포레스트(random forest)와 에이다부스트(adaboost)를 이용한 도서 추천시스템 구축 방법을 제안하였으며, 넷플릭스는 고객들의 영화 평점 정보를 기반으로 협업필터링과 기계학습 알고리즘으로 앙상블 모형을 개발하여 추천시스템을 구축하였다[13]. 그리고 Portugal 등[14]은 추천시스템에서 기계학습 적용의 연구 동향을 파악하고 기계학습의 접근법과 성능 지표를 분석하였다.

이러한 기계학습 기반의 추천시스템을 위한 기존 연구들은 추천모형을 구축할 때 단일 데이터 셋 하에서 여러 가지 파라미터에 의해 나온 추천정확도들을 비교하여 최적의 파라미터 선정에 집중하는데 그치고 있다. 하지만 본 연구에서는 여러 가지 파라미터들에 의해 생성된 서로 독립인 모델들의 결과를 모두 합쳐 하나의 데이터 셋으로 생성하는 스태킹 모형을 적용하여 추천시스템의 성능을 향상시키고자 한다. 앙상블 방법 중 하나인 스태킹 모형은 새로운 데이터 셋을 생성함으로써 모델 구축의 복잡성을 해결하고 데이터 속에 함축된 복잡하고 다양한 정보들을 반영할 수 있다는 장점이 있다[15]. 이러한 장점 때문에 스태킹 모형은 최근 다양한 분야의 연구에서 사용되고 있다. Ekbal[16]는 스태킹 모형을 사용하여 생물 의학적 개체 분류 모델에 관한 연구를 제안하였다. Thorne 등[17]은 5개의 독립적인 분류모형으로 구성한 스태킹 모형을 사용하여 가짜 뉴스 판별에 대한 해결책을 제시하였다. Tsai[18]는 이미지 분류를 위한 서포트벡터머신을 이용한 스태킹 모형을 제안하였으며, Huang 등[19]은 음성 신호에서 사람의 정서적인 상태를 위한 스태킹 모형을 제안하였다. 또한, Cao 등[20]은 스태킹 모형을 사용하여 기존 분류 모형인 랜덤포레스트, 서포트벡터머신, 그리고 앙상블 모형보다 분류 성능을 높일 수 있다고 주장하였는데, 최소 2%에서 최대 5%까지 정분류율이 높아짐을 실증분석을 통해 보여주었다.

본 연구에서 제안하는 추천시스템은 크게 두 단계로 구성되어 있다. 첫 번째 단계는 기존의 협업필터링 알고리즘을 이용하여 여러 가지 파라미터에 의한 서로 독립적인 협업필터링 결과를 생성한 후, 모든 결과의 품목에 대한 사후확률값을 합쳐 하나의 새로운 데이터 셋을 생성하는 것이다. 두 번째 단계는 생성된 데이터 셋으로 기계학습 모형을 이용한 추천시스템을 구축하는 것이다. 제안 모형의 추천 성능은 3개의 상품을 추천하였을 때, 추천대상 고객이 적어도 하나의 상품을 구매하였다면 추천을 성공한 것으로 평가하였다. 이는 기존의 많은 연구에서 추천모형의 성과평가 방법에 따른 것이다 [21,22]. 제안 모형의 성능은 협업필터링과 기계학습을 단일로 사용하여 구축한 상품 추천시스템의 결과를 제안한 모형과 비교하여 스태킹 모형의 우수성을 검증하고자 한다.

본 연구의 구성은 다음과 같다. 2절에서는 연구배경에 대하여 설명을 하고 3절에서는 구체적인 연구방법에 대한 소개한다. 4절에서는 실제 데이터를 이용한 연구 결과에 대하여 보이고 5절에서는 결론과 향후 연구에 대하여 제시한다.

2. 연구배경

2.1 협업필터링

협업필터링은 상품추천시스템에서 널리 사용되는 추천 알고리즘으로써, 추천대상 사용자의 행동을 다른 사용자의

행동과 비교하고 가장 가까운 이웃을 선정하여 이웃의 선호도를 기반으로 추천대상 사용자의 관심 또는 선호도를 예측한다[1]. 특정 고객에게 구매확률이 높을 것으로 기대되는 상품을 추천하기 위한 협업필터링 알고리즘은 크게 3단계로 구성하였다.

1단계 - 협업필터링 알고리즘의 첫 번째 단계는 사용자의 구매 기록을 바탕으로 혼돈행렬(confusion matrix)을 도출하는 것이다. 혼돈행렬은 사용자가 구매한 모든 상품을 대상으로 전체 구매 횟수에 대한 개별 상품구매 횟수의 비율 값으로 구성되어 있다.

Table 1. Example of confusion matrix

Id	Item 1	Item 2	Item 3
A00001	0.7	0.1	0.2
A00002	.	0.6	0.4
A00003	0.1	0.9	.
A00004	0.5	0.2	0.3

Table 1에서 Id는 각 사용자를 나타내며, 각 Item은 각 상품을 의미한다. 행과 열이 교차하는 지점의 값은 사용자의 해당 상품에 대한 구매 비율을 나타낸다. 교차하는 지점에 값이 없을 경우는 사용자가 해당 상품의 구매 이력이 없는 경우이다.

2단계 - 두 번째 단계는 이전 단계에서 도출한 혼돈행렬을 이용하여 사용자 간의 유사도를 계산하고 가장 가까운 이웃을 찾는 것이다. 사용자 간의 유사도를 측정하는 대표적인 방법은 상관관계 유사도 측정방법, 코사인 유사도 측정방법, 고어 유사도 측정방법, 그리고 자카드 유사도 측정방법 등이며, 이외에도 여러 가지 방법이 존재한다. 본 연구에서는 추천시스템 구축 시 가장 널리 사용되는 코사인 유사도 측정방법을 사용하였다[9].

계산된 코사인 유사도 값은 -1부터 1 사이의 값을 가지게 된다. -1에 가까운 값일수록 서로 유사하지 않은 경우이며, 값이 0이면 서로 독립인 경우이고 1에 가까운 값일수록 서로 유사한 경우이다.

이웃은 코사인 유사도 측정방법을 이용하여 도출한 유사도 값으로 선정한다. 유사한 이웃을 선택하는 것은 이웃의 수가 결과에 영향을 주기 때문에 예측 정확도에 중요하다. 유사한 이웃을 선택하는 방법은 임계치 설정 방법과 최근접 이웃 방법 등이 있다. 임계치 설정 방법은 대상 고객과 이웃 고객의 임계치를 미리 설정하여 임계치를 넘는 값을 갖는 고객을 이웃 고객으로 선정하는 방법이다. 최근접 이웃 방법은 k명의 이웃 고객을 설정하여 최대 k명의 이웃 고객의 정보를 반영하는 방법이며, k를 설정하는 방법은 예측 정확도에 따라 설정하며 모든 고객에게 공통적인 값을 부여한다[23].

3단계 - 세 번째 단계는 고객의 품목에 대한 예상 등급을 구하는 것이다. 예상 등급은 고객별 품목에 대한 구매의 정도이며, 예상 등급이 높은 순서로 정렬하여 등급이 높은 품목을 추천 상품으로 선정할 수 있다. 예상 등급은 이전 단계에서 구한 이웃들의 유사도 값을 가중하여 항목에 대한 구매 비율의 합으로 계산된다[24].

2.2 스태킹

앙상블 모형은 훈련용 데이터에서 여러 개의 분류기를 만든 후, 소속집단을 모르는 데이터를 분류할 때 각 분류기를 적용하여 집단을 분류한 후 결과 집단의 다수결로서 최종 집단을 결정하는 방법이다. 앙상블 모형은 각각의 분류기보다 더 좋은 분류결과를 얻을 수 있다. 하지만 각각의 분류기가 서로 독립이어야 하며 오분류율이 적어도 50%보다는 적어야 한다. 앙상블 모형에 사용하는 분류기는 어떠한 분류모형이라도 사용할 수 있다[15,25].

스태킹은 이러한 앙상블 방법 중 하나이며, 일반화된 높은 정확도를 위한 방법이다. 스태킹은 크게 두 단계로 구성된다. 첫 번째 단계는 다양한 기본 분류모형을 선택하여 학습시키는 것이다. 다양한 기본 분류모형을 선택하는 이유는 다양한 기본 분류모형으로 구성된 앙상블 모형이 우수한 성능을 보이며, 스태킹은 다양한 파라미터에 의한 결과를 모두 반영할 수 있기 때문이다[25,26]. 선택된 기본 분류모형들은 학습을 통하여 사후확률을 도출한다. 다양한 파라미터에 의해 도출된 사후확률을 합쳐서 새로운 데이터 셋을 생성한다. Fig. 1은 스태킹의 첫 번째 단계를 통하여 생성된 새로운 데이터 셋의 예시이다. $y_n^{(k)}$는 k번째 분류모형에서의 n번째 출력값이다[27].

$$
\begin{array}{cc}
Input & output \\
y_1^{(1)}, y_1^{(2)}, \cdots, y_1^{(i)}, \cdots, y_1^{(k)}, & y_1 \\
y_2^{(1)}, y_2^{(2)}, \cdots, y_2^{(i)}, \cdots, y_2^{(k)}, & y_2 \\
\vdots & \vdots \\
y_n^{(1)}, y_n^{(2)}, \cdots, y_n^{(i)}, \cdots, y_n^{(k)}, & y_n
\end{array}
$$

Fig. 1. Example of new data set generated by stacking

두 번째 단계는 이전 단계에서 생성된 새로운 데이터 셋을 입력변수로 사용하는 분류모형을 학습하여 최종 분류 문제를 해결하는 것이다. 이 단계에서 적용되는 분류모형은 이전 단계에서 얻은 예측값과 최종 결정 사이의 관계를 학습하여 일반적으로 단일 분류모형보다 좋은 성능을 보인다고 알려져 있다[28]. Fig. 2는 일반적인 스태킹 프로세스를 도식화한 것이다.

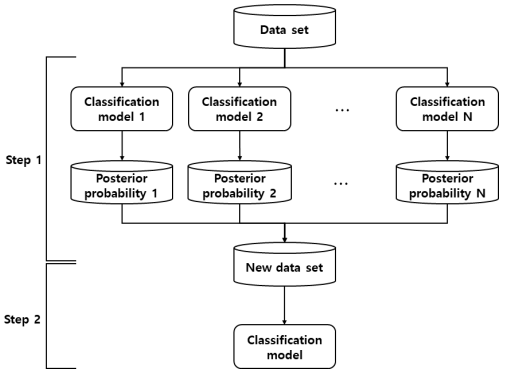

Fig. 2. Diagram of general stacking process

3. 연구방법

본 연구에서 제안하고자 하는 방법론은 스태킹 기반 추천시스템이며, 크게 두 단계로 이루어져 있다. 1단계는 사용자 기반 협업필터링 알고리즘을 통하여 고객의 상품에 대한 예상 구매 평점을 구하는 것이며, 2단계는 이전 단계에서 구해진 구매확률을 기계학습 분류모형에 학습시키는 것이다.

1단계 – 협업필터링 알고리즘의 첫 번째 단계는 구매 비율 행렬을 얻는 것이다. 구매 비율 행렬은 각 품목을 구매한 전체 품목 수의 합으로 나눠준 값으로 구성된다. 도출한 구매 비율 을 바탕으로 아래와 같은 식을 통해 사용자의 품목에 대한 평점이 5점 척도로 계산된다.

$$\cdot\ Rating_{n,i} = 1 + 4 \times \frac{(r_{n,i} - \min_n)}{(\max_n - \min_n)}$$

여기서 $Rating_{n,i}$은 고객 n의 품목 i에 대한 구매 평점, $r_{n,i}$은 고객 n의 품목 i에 대한 구매 비율, $\max_n$은 고객 n의 구매 비율 중 가장 큰 값, 그리고 $\min_n$은 고객 n의 구매 비율 중 가장 작은 값이다.

다음으로 고객 간의 유사도는 앞서 구한 평점 행렬을 이용하여 구할 수 있다. 본 연구에서는 코사인 유사도 측정방법을 이용하며 이를 통해 유사 고객 순위가 결정된다. 고객 간 유사도는 이웃 고객들의 품목에 대한 평점과 함께 고객별 상품의 선호도 예측에 사용된다. 본 연구에서 상품 예측 선호도는 고객별 상품의 예상 평점을 의미하는데 아래의 식으로부터 도출할 수 있다.

$$\cdot\ E\_R_{n,i} = \frac{\sum_{u=1}^{n} Cosine_{n,u} * R_{u,i}}{\sum_{u=1}^{n} Cosine_{n,u}}$$

여기서 $E_R_{n,i}$는 고객 n의 품목 i에 대한 예상 선호도, $Cosine_{n,u}$는 고객 n과 이웃 고객 u의 유사도, 그리고 $R_{u,i}$는 이웃 고객 u의 품목 i에 대한 평점이다.

2단계 – 이전 단계에서 얻어진 상품의 예상평점을 입력변수로 하고, 고객별 상품의 실제 구매여부를 출력변수로 하는 새로운 훈련 데이터 셋을 생성하여 기계학습 분류모형에 학습시킨다. 본 연구에서는 분류 문제에서 높은 성능을 보인다고 알려진 의사결정나무 모형과 신경망 모형 그리고 두 모형의 결과를 결합한 앙상블 모형을 추천시스템을 위한 분류모형으로 고려하였다. 각 모형은 각 상품의 구매에 대한 사후확률을 도출하며, 도출된 사후확률을 바탕으로 상품별 순위를 매겨 상위 3개의 상품을 추천하였다. 모형비교는 모형별로 3개의 상품 중 하나의 상품 이상 적중 시 추천 성공으로 간주하게 된다. Fig. 3은 본 연구에서 제안하는 추천시스템 구조를 도식화해서 보여주고 있다.

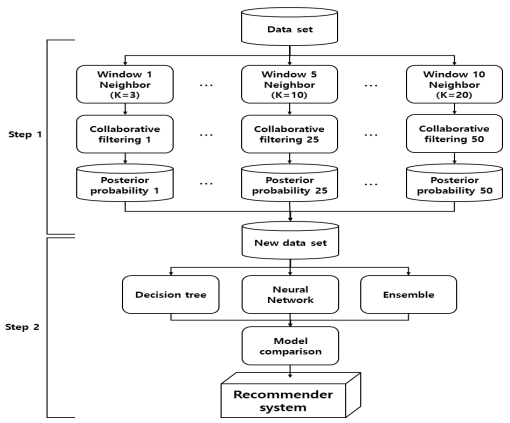

Fig. 3. Flowchart of proposed stacking architecture

4. 실증 분석

본 연구의 실증분석을 위해 사용된 데이터는 2014년 1월부터 2015년 12월까지 'L'사 19,383명 고객의 일별 구매 이력 정보와 이 기간 동안 구매된 4,386개의 상품 정보를 포함하고 있다. 또한, 분석을 위한 최종 데이터 셋은 19,383명의 고객이 각자 구매한 상품들을 고려하여 구성하였으며 총 약 2,850만개의 샘플로 구성되어 있다. 구매 고객들의 특성을 살펴보면 여성 고객이 80%로 높은 비율을 차지하고 있으며, 40대 고객이 43%로 가장 높았다. 구매 상품들은 대, 중, 소분류로 구분되어 있는데 대분류는 593개, 중분류는 807개, 그리고 소분류는 4,386개이다. 본 연구에서는 상품에 대한 정보의 손실을 최소화하기 위해 소분류로 구분된 상품정보를 사용하였다. 또한, 고객들의 최근 구매 이력의 중요도를 반영하기 위해 슬라이딩 윈도우 방법을 적용하였으며 학습 기간은 1년, 검증 기간은 1개월로 설정하였다[29]. 실증 분석은 통계분석 도구 SAS 9.4와 SAS Enterprise Miner Workstation 14.1을 사용하여 진행하였다.

Table 2. Training period and test period assigned to each window

Window no.	Training period	Test period
Window 1	2014.01.01.~2014.12.31.	2015.01.01.~2015.01.31.
Window 2	2014.02.01.~2015.01.31.	2015.02.01.~2015.02.28.
Window 3	2014.03.01.~2015.02.28.	2015.03.01.~2015.03.31.
Window 4	2014.04.01.~2015.03.31.	2015.04.01.~2015.04.30.
Window 5	2014.05.01.~2015.04.30.	2015.05.01.~2015.05.31.
Window 6	2014.06.01.~2015.05.31.	2015.06.01.~2015.06.30.
Window 7	2014.07.01.~2015.06.30.	2015.07.01.~2015.07.31.
Window 8	2014.08.01.~2015.07.31.	2015.08.01.~2015.08.31.
Window 9	2014.09.01.~2015.08.31.	2015.09.01.~2015.09.30.
Window 10	2014.10.01.~2015.09.30.	2015.10.01.~2015.10.31.

Table 2는 본 연구에서 설정한 10개의 윈도우에 할당된 실험 기간을 보여주고 있다. 각 윈도우별 학습 기간과 테스트 기간은 해당 기간에서의 구매이력 데이터가 입력변수로 사용되며, 상품의 구매여부가 출력변수로 사용되는 데이터 셋으로 구성되어 있다. 각각의 윈도우에서 협업필터링 알고리즘을 통하여 고객의 예상 선호도가 생성되는데, 이때 유사 이웃의 수에 따라 추천되는 상품이 바뀌게 되면서 최종 상품 추천 성공률이 달라지게 된다. 따라서, 본 연구에서는 이웃의 수를 3명부터 40명까지 바꿔가며 상품 추천 오분류율을 확인하여 이웃의 수를 결정하였다.

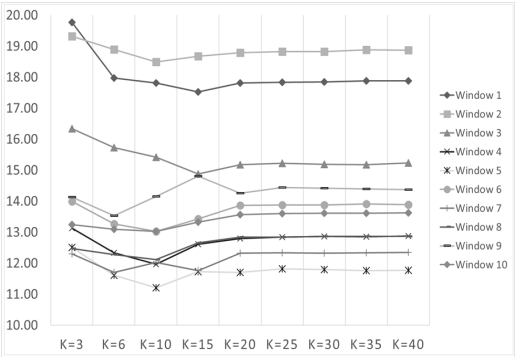

Fig. 4. Misclassification rate(%) according to the number of neighbors in each window

Fig. 4는 이웃의 수에 따른 협업필터링 기반의 추천시스템 오분류율을 보여주고 있다. 오분류율은 이웃 수(K) 20 값을 기준으로 커질수록 큰 변동이 없기에 20보다 작은 이웃의 수를 새로운 데이터셋 생성을 위한 사후확률 도출에 사용하기로 하였다.

Table 3. Misclassification rate(%) of collaborative filtering when K value is under 20

Window no.	Number of neighbors				
	K=3	K=6	K=10	K=15	K=20
Window 1	19.77	17.97	17.81	17.53	17.81
Window 2	19.31	18.89	18.49	18.67	18.79
Window 3	16.34	15.73	15.41	14.83	15.17
Window 4	13.13	12.33	11.97	12.61	12.79
Window 5	12.51	11.61	11.21	11.73	11.71
Window 6	13.99	13.27	13.03	13.43	13.87
Window 7	12.30	11.70	12.02	11.76	12.32
Window 8	12.48	12.28	12.12	12.66	12.84
Window 9	14.13	13.53	14.15	14.81	14.25
Window 10	13.25	13.09	13.03	13.33	13.57
Avg.	14.72	14.04	13.93	14.14	14.31

Table 3은 각 이웃 수에 따른 협업필터링 알고리즘의 추천 결과를 나타낸 표이다. 각 윈도우와 각 이웃의 수마다 다른 오분류율 값을 보이지만, 이웃의 수를 기준으로 오분류율의 평균을 보면 비슷한 값을 보인다. 이를 통해, 협업필터링 단독으로 추천시스템을 구축하였을 경우에도 낮은 오분류율을 도출할 수 있다는 것을 알 수 있다. 이러한 과정을 통해 도출된 50개의 상품구매의 사후확률 값들은 새로운 데이터 셋 생성을 위해 입력변수로 사용되게 된다.

Table 4. Misclassification rates(%) of proposed stacking model

Decision tree	Neural network	Ensemble
11.64	11.08	11.22

Table 4는 본 연구에서 제안한 스태킹 모델을 적용한 추천시스템의 추천 오분류율 결과이다. 기계학습 분류모형 중에서는 신경망이 가장 좋은 성능을 보였으며 세 가지 분류모형 모두 협업필터링을 단독으로 사용한 결과보다 높은 추천 성능을 보였다. 이는 제안 모형이 단순히 구매 이력만을 이용하여 도출한 사후 구매확률보다 이 확률값들을 기계학습방식으로 구축하는 모형이 더 높은 추천 성능을 가질 수 있다는 것을 보여준다.

Table 5는 제안 모형과의 성능을 비교하기 위해 스태킹 모형의 적용 없이 구매이력 정보를 기계학습 분류모형에 학습시킨 결과를 나타낸 표이다. 제안 모형과의 공평한 비교를 위해 학습 기간의 입력변수와 출력변수를 제안 모형과 동일하게 설정하였다. 설정한 학습 기간의 입력변수와 출력변수는 2014년 1월부터 2015년 10월의 구매 이력 그리고 2015년 11월의 구매 여부로 설정하였으며, 검증 기간은 2015년 12월의 구매 여부로 설정하였다. 본 연구에서 사용된 의사결정나무와 신경망 모형은 다양한 탐색을 통해 최적 파라미터들이 결정되었다. 의사결정나무 모형의 경우 분리기준으로 카이제곱 통계량을 사용하였으며, 최대가지는 2, 최대 깊이는 6으로 설정하였다. 신경망 모형의 경우 은닉층의 수는 1개, 은닉마디의 수는 3개, 그리고 활성함수는 sigmoid 함수로 설정하였다.

Table 5. Misclassification rates(%) of machine learning classifiers without applying stacking model

Decision tree	Neural network	Ensemble
13.58	12.36	12.50

스태킹 모형의 적용 없이 기계학습 분류모형만으로 추천시스템을 구축한 경우 협업필터링 모형의 결과보다는 높은 추천 성공률을 보였지만 스태킹 모형을 적용한 경우보다는 다소 떨어지는 추천 성공률을 나타냈다. 이는 제안 모형이 단순히 구매 이력만을 학습하여 추천하는 알고리즘보다 더 많은 정보를 담고 있는 훈련 데이터를 학습하기 때문이라고 판단된다. 즉, 스태킹 모형으로부터 생성된 새로운 데이터 셋이 기계학습 분류모형의 훈련 데이터 셋으로써 매우 유용한 역할을 할 수 있는 것이다.

5. 결론

본 연구는 'L'사 고객의 대규모 구매이력 정보인 구매 이력 데이터를 활용하여 스태킹 모형을 적용한 추천시스템을 구축하는 방법을 제안하고자 하였다. 현재 'L'사는 다른 온라인 쇼핑 업체와 마찬가지로 고객 빅데이터를 활용한 맞춤형 고객추천모형 개발에 집중하고 있다. 이는 본 연구에서 개발하고자 하는 새로운 상품 추천 알고리즘의 목적과 부합한다고 할 수 있다. 스태킹 모형은 초기 단계에 참여한 모델들이 얼마나 잘 학습이 되었는지에 관계없이 각 모델이 동일한 양을 기여한다는 단점[30]이 존재하지만 새로운 데이터 셋을 생성함으로써 모델 구축의 복잡성을 해결하고 데이터 속에 함축된 복잡하고 다양한 정보들을 반영할 수 있다는 장점이 있다. 스태킹 모형의 유용성을 검증하기 위해 새롭게 생성된 데이터 셋은 기계학습 분류모형인 의사결정나무 모형, 신경망 모형 그리고 앙상블 모형을 사용하여 훈련되었다. 실험 결과는 제안 모형이 단순히 협업필터링만을 이용한 추천시스템과 스태킹 모형의 적용 없이 기계학습 모델로 구축한 추천시스템에 비해 추천 성능이 향상됨을 보였다. 본 연구의 의의는 추천시스템의 추천 성능 향상을 위해 모델 구축의 복잡성을 해결하고 데이터 속에 함축된 정보를 최대한 반영했다는 것이다. 또한, 본 연구에서 제안하는 모형은 슬라이딩 윈도우 방법을 적용하여 최근 구매이력 정보의 중요도를 반영하고 있다. 그리고 스태킹 모형으로부터 생성된 새로운 데이터 셋은 여러 시점의 모든 상품에 대한 구매 정보를 가지게 되므로 기계학습 모형이 다양한 정보를 학습할 수 있게 한다. 따라서 기업에서 본 연구에서 제안한 추천시스템을 이용한 고객 맞춤형 서비스를 제공한다면 새로운 이익을 창출할 수 있는 기회를 만들 수 있을 것이다.

그럼에도 불구하고 본 연구는 고객의 구매 이력 데이터 이외에 고객의 인구통계학적 데이터와 같은 상품 추천에 영향을 줄 수 있는 추가적인 데이터를 활용하지 못했다는 한

계점을 가지고 있다. 협업필터링 알고리즘을 적용함에 있어서 구매 이력 데이터 이외에 고객의 인구 통계학적인 데이터를 추가적으로 사용한다면 추천 성능이 더욱 좋은 추천시스템을 구축할 수 있을 것으로 기대된다. 또한, 대용량 고객 데이터라는 특성에 기인하여 협업필터링은 많은 연산시간을 필요로 한다. 본 연구에서는 협업필터링 방법을 적용하여 스태킹을 통한 새로운 입력 데이터 셋을 생성하였는데, 이 과정에서도 협업필터링이 아닌 기계학습 모형을 적용한다면 이러한 문제점을 어느 정도 해결할 수 있을 것으로 기대한다.

REFERENCES

[1] P. Resnick & H. R. Varian. (1997). Recommender systems. *Communications of the ACM*, 40, 56-58.

[2] J. Bennett & S. Lanning. (2007). The netflix prize. *Proceedings of KDD cup and workshop*, 35.

[3] D. Goldberg, D. Nichols, B. M. Oki & D. Terry. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35, 61-70.

[4] B. S. Kang. (2019). A Study on the Accuracy Improvement of Movie Recommender System Using Word2Vec and Ensemble Convolutional Neural Networks. *Journal of Digital Convergence*, 17(1), 123-130.

[5] Z. D. Zhao & M. S. Shang. (2010). User-based collaborative-filtering recommendation algorithms on hadoop. *IEEE, 2010 Third International Conference on Knowledge Discovery and Data Mining*, 478-481.

[6] H. C. Lee. (2006). Improved algorithm for user based recommender system. *Journal of the Korean Data & Information Science Society*, 17, 717-726.

[7] B. M. Marlin. (2003). Modeling user rating profiles for collaborative filtering. *Advances in neural information processing systems, 16*, 627-634.

[8] G. Linden, B. Smith & J. York. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*, 76-80.

[9] S. Lee. (2014). A New Collaborative Filtering Method for Movie Recommendation Using Genre Interest. *Journal of Digital Convergence*, *12(8)*, 329-335.

[10] J. T. Oh & S. Y. Lee. (2017). A Movie Recommendation System based on Fuzzy-AHP with User Preference and Partition Algorithm. *Journal of Digital Convergence, 15(11),* 425-432.

[11] N. K. Kim & S. Y. Lee. (2013). Bayesian network based Music Recommendation System considering Multi-Criteria Decision Making. *Journal of Digital Convergence, 11(3),* 345-352.

[12] K. Tsuji, F. Yoshikane, S. Sato & H. Itsumura. (2014). Book Recommendation Using Machine Learning Methods Based on Library Loan Records and Bibliographic Information. *IEEE, 2014 IIAI 3rd International Conference on Advanced Applied Informatics*. (pp. 76-79).

[13] M. Piotte & M. Chabbert. (2009). *The pragmatic theory solution to the netflix grand prize*. Netflix prize documentation, Canada

[14] I. Portugal, P. Alencar & D. Cowan. (2017). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205-227.

[15] R. Polikar. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine, 6,* 21-45.

[16] A. Ekbal & S. Saha. (2013). Stacked ensemble coupled with feature selection for biomedical entity extraction. *Knowledge-Based Systems, 46,* 22-32.

[17] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang & A. Vlachos. (2017). Fake news stance detection using stacked ensemble of classifiers. *In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism.* (pp. 80-83).

[18] C. F. Tsai. (2005). Training support vector machines based on stacked generalization for image classification. *Neurocomputing, 64,* 497-503.

[19] Y. Huang, G. Zhang, & X. Xu. (2009, November). Speech emotion recognition research based on the stacked generalization ensemble neural network for robot pet. *In 2009 Chinese Conference on Pattern Recognition* (pp. 1-5). IEEE.

[20] Z. Cao, X. Pan, Y. Yang, Y. Huang & H. B. Shen. (2018). The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics, 34(13),* 2185-2194.

[21] J. L. Herlocker, J. A. Konstan, L. G. Terveen & J. T. Riedl. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS), 22(1),* 5-53.

[22] H. Kim, G. Yang, H. Jung, S. H. Lee & J. J. Ahn. (2019). An intelligent product recommendation model to reflect the recent purchasing patterns of customers. *Mobile Networks and Applications, 24(1),* 163-170.

[23] S. J. Lee. (2009). A study on neighbor selection methods

in k-NN collaborative filtering recommender system. *Journal of the Korean Data & Information Science Society, 20,* 809-818.

[24] J. Herlocker, J. A. Konstan & J. Riedl. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information retrieval, 5(4),* 287-310.

[25] L. Rokach. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33,* 1-39.

[26] K. S. Eo & K. C. Lee. (2019). Investigating Opinion Mining Performance by Combining Feature Selection Methods with Word Embedding and BOW (Bag-of-Words). *Journal of Digital Convergence, 17(2),* 163-170.

[27] J. Yan & S. Han. (2018). Classifying Imbalanced Data Sets by a Novel RE-Sample and Cost-Sensitive Stacked Generalization Method. *Mathematical Problems in Engineering, 2018.*

[28] F. Gunes, R. Wolfinger & P. Y. Tan. (2017). *Stacked Ensemble Models for Improved Prediction Accuracy.* SAS Global Forum 2017, SAS0437-2017.

[29] C. Kim, T. Y. Kim, I. Park & J. J. Ahn. (2015). A study on the improvement of the economic sentiment index for the Korean economy. *Journal of the Korean Data & Information Science Society, 26,* 1335-1351.

[30] D. H. Wolpert. (1992). Stacked Generalization. *Neural networks, 5(2),* 241-259.

박 성 종(Sung-Jong Park) [학생회원]

· 2018년 2월 : 연세대학교 정보통계학과(이학사)
· 2018년 3월 ~ 현재 : 연세대학교 정보통계학과 석사과정
· 관심분야 : 고객추천시스템, 머신러닝
· E-Mail : tjdwhdgo@yonsei.ac.kr

김 영 민(Young-Min Kim) [정회원]

· 2009년 8월 : 국민대학교 비즈니스IT학과(경영학사)
· 2015년 8월 : 연세대학교 산업공학과(공학박사)
· 2018년 3월 ~ 현재 : 순천향대학교 빅데이터공학과 교수
· 관심분야 : 금융빅데이터분석, 지능형정보시스템
· E-Mail : kimym38@sch.ac.kr

안 재 준(Jae-Joon Ahn) [정회원]

· 2005년 2월 : 연세대학교 산업공학과(공학사)
· 2008년 2월 : 연세대학교 산업공학과(공학석사)
· 2013년 2월 : 연세대학교 산업공학과(공학박사)
· 2014년 3월 ~ 현재 : 연세대학교 정보통계학과 교수
· 관심분야 : 고객추천시스템, 데이터마이닝, 금융통계
· E-Mail : ahn2615@yonsei.ac.kr