특징 최소화와 데이터 선별을 활용한 영화 관객수 예측

양영보†, 유헌창‡
†롯데쇼핑 e 커머스, ‡고려대학교 정보대학 컴퓨터학과
e-mail : ybyang2@lotte.net, yuhc@korea.ac.kr

Prediction of Number of Movie Audience Using Feature Minimization and Data Selection

Youngbo Yang†, Heonchang Yu‡
†Lotte Shopping e-commerce, ‡Dept. of Computer Science and Engineering, Korea University

요 약

빅데이터 분석을 위해 많이 사용하고 있는 기계학습 알고리즘들 중 딥러닝 알고리즘이 많이 활용되고 있으며 분류와 예측에 높은 정확도를 나타내고 있다. 딥러닝 알고리즘의 적용에 따른 많은 장단점들이 있지만, 단점은 분석에 사용되는 특징들이 너무 많다는 것과 분석 모델을 만드는데 사용되는 알고리즘도 여러 가지를 적용하다 보니 분석 시간이 오래 걸린다는 것이다. 이런 단점들은 업무를 파악하면 특징을 최소화할 수 있고 필요로 하는 정보만 선별해서 대표적인 딥러닝 알고리즘 하나에 분석을 하게 되면 분석 시간을 단축시킬 수 있다. 이 실험은 [1], [2]에서 연구한 영화 관객수 예측 모델을 4개의 특징으로 최소화하고 선별된 데이터를 인공신경망 알고리즘 하나로 예측 모델을 생성하였을 때 유의미한 정보를 도출해 낼 수 있는지를 알아보기 위한 것이다. 실험결과는 최종 관객수를 1명 단위까지 정확하게 예측하지는 못했지만 비슷한 수준의 관객수 정보를 예측하였다. 학문적인 접근으로 보았을 때 예측 정확도가 높지 않으면 사용이 불가능한 모델이라고 판단할 수 있지만, 기업 입장으로 접근해 보았을 때 예측 정보가 [1]. [2] 연구 결과에 비해 부족한 수준은 아니다. 총 소요된 시간은 기획 3일, 데이터 수집 및 모델 개발 5일, 분석 시간 10분으로 개발 시간 단축, 업무 효율성 향상, 비용 절감을 기대할 수 있다.

1. 서론

오래 전부터 기업은 고객의 데이터를 분석하여 마케팅 전략에 사용해 왔다. 데이터 분석 초기에는 설문조사를 활용해서 단순하게 전통적 조사법에 의존한 데이터를 수집하여 일반 통계로 분석했다. 그리고, 인터넷이 발전하고 온라인에서 활동하는 고객들이 많아지면서 로그 데이터 분석을 통해 좀 더 의미있는 분석을 하게 되었다. 4차 산업혁명 시대에 들어서면서 빅데이터 분석을 하게 되었고 고객 분석뿐만 아니라 상품 추천, 트랜드 분석, 위험 요소 관리 등 다양한 영역에서 활용되고 있다. 이런 데이터분석 작업에는 오랜 시간 통계 이론을 활용해 왔으며 회귀분석이 보편적으로 많이 사용되어 왔다. 하지만, 사람이 쉽게 파악할 수 없는 빅데이터 분석 시대가 되면서 정교한 분석이 쉽지 않게 되었다. 이런 문제점을 보완하기 위해 기계학습 알고리즘을 이용하여 기계의 판단을 보고 분석하는 방법을 데이터 분석에 활용하게 되었다. 기계학습의 한 분야인 딥러닝이 빅데이터 분석에 많이 사용되고 있고 데이터 분류와 예측에 높은 정확도를 나타내고 있다. 하지만, 딥러닝 알고리즘의 단점은 분석에 사용되는 특징들이 너무 많다는 것과 분석 모델을 만드는데 사용되는 알고리즘도 여러 가지를 적용하다 보니 분석 시간이 오래 걸린다는 것이다. 이런 단점들은 업무를 파악하면 특징을 최소화할 수 있고 필요로 하는 정보만 선별해서 대표적인 딥러닝 알고리즘 하나에 분석을 하게 되면 분석 시간을 단축시킬 수 있다. 이 실험은 [1], [2]에서 연구한 영화 관객수 예측 모델을 4개의 특징으로 최소화하고 선별된 데이터를 인공신경망 알고리즘 하나로 예측 모델을 생성하였을 때 유의미한 정보를 도출해 낼 수 있는지를 알아보기 위한 것이다.

2. 관련연구

영화의 역사가 오래된 만큼 흥행과 관련된 연구는 오랜 시간 이루어져 왔다. 주로 흥행에 영향을 끼칠 수 있는 특징을 찾아내거나 찾아낸 특징을 분석에 이용하여 예상 관객수를 예측하는 연구들이 많았다. [1]은 영화 흥행에 영향력이 미칠 것으로 예상되는 특징

29개를 이용하여 영화 관객수를 예측하는 연구를 했다. 데이터마이닝 분류 방법으로는 의사결정나무, 인공신경망, 다항로짓모형, SVM을 사용하였고 인공신경망 예측율이 가장 높게 나타났다. [2]에서는 GLS 모형과 Bass 모형을 결합한 Hybrid 모형을 이용하여 영화 관객수를 예측하는 연구를 하였고 사용된 특징은 21개이며 Hybrid 모형을 활용했을 때 예측율이 높았다.

3. 설계 및 구현

3.1 특징 최소화

[1], [2]에서 영화 관객수 예측에 영향을 끼칠 수 있는 특징들을 21~29개로 정리하였다. 감독, 배우, 개봉 시즌, 관람 등급, 배급, 관객 평가 등 영화 산업 전반적으로 중요한 특징들을 포함하고 있다. 하지만, 데이터 분석 업무 시 어디까지 고려해야 할지 데이터 수치화는 어떻게 해야 할지 경계가 모호하고 의견을 조율하는데 많은 시간을 소비하고 있다. 그래서, 수치화가 편리하고 영향력이 가장 높을 것으로 예상되는 특징으로 최소화해 주는 작업이 필요하다. 영화 흥행은 결국 관객의 선택이기 때문에 관객과 관련된 최소 특징으로 축소하였다. 온라인 트래픽이 가장 많은 곳은 [3]에서 포털 1위 네이버로 확인 하였고 [1],[2] 공통적으로 사용된 관객 평점과 조회 순위를 특징으로 사용 하였다. 실험에 사용된 특징은 <표 1>과 같다.

<표 1> 예측 모델에 사용된 특징

특징	설명
일차	개봉 후 경과 일수
별점	해당 일차 네이버 별점
조회 순위	해당 일차 네이버 조회 순위
누적 관객수	해당 일차 영화진흥위원회 누적 관객수
최종 관객수	해당 영화 영화진흥위원회 최종 관객수

3.2 데이터 수집

영화진흥위원회에서는 상업영화와 다양성영화로 구분하여 데이터를 제공하고 있다. 다양성영화는 개봉일과 종료일이 명확하지 않기 때문에 제외하였고 상업영화를 기준으로 2016년 1월 1일부터 2018년 12월 31일까지 총 3년간의 데이터를 수집하였다. [4]에서 제공해 주는 API 중에 일별 박스오피스를 활용하여 일자, 영화명, 누적 관객수 정보 10,960건을 수집하였고, 네이버 별점과 조회 순위는 [5],[6]에서 제공되는 정보를 크롤링하여 각각 51,139건, 54,796건을 수집하였다. API 연동과 크롤링 작업은 python 3.6과 requests, beautifulsoup4 라이브러리를 활용하였고 수집 이후에 데이터 선별을 위해 오픈소스 데이터베이스 PostgreSQL을 사용하여 수집한 데이터를 테이블로 저장하였다.

3.3 데이터 선별

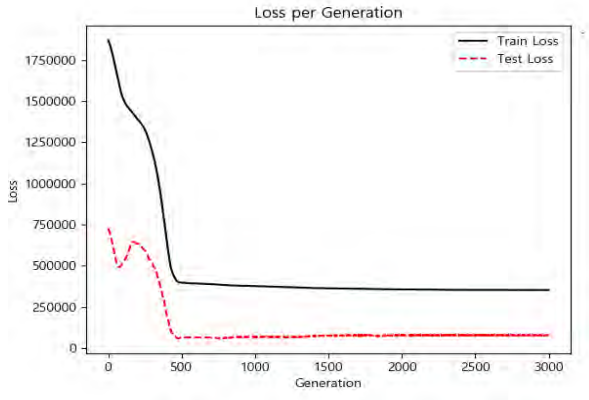
PostgreSQL에 저장된 누적 관객수, 별점, 조회 순위 정보를 일자 정보와 영화명을 기준으로 Join 하여 7,365건 데이터로 1차 선별을 하였으며 별점 또는 조회 순위 정보가 없는 영화 정보는 제외하였다. [7]에서 주간/주말 스크린점유율을 확인 해보면 우리나라 상업영화는 영화 개봉 이후 1주일이 지나면 흥행 여부를 판단하고 극장에서 상영 횟수를 줄이는 경우가 많이 있다. 그래서, 개봉 첫 주 누적 관객수 정보가 최종 관객수에 영향을 끼칠 수도 있다고 판단하고 2차 선별을 하였다. 일반적으로 영화는 매주 수요일에 개봉을 많이 하고 있고 주말에 관객이 많이 관람하기 때문에 주말을 포함하는 개봉 5일차부터 7일차까지 일차 별로 데이터를 선별하였다. 개봉 5일차 305건, 6일차 304건, 7일차 295건 데이터로 2차 선별하여 총 3개의 CSV 데이터를 생성하였다.

3.4 분석 알고리즘 최소화

[1], [2]에서 분석할 때 의사결정나무, 인공신경망, SVM, 다중 회귀, Bass 모형 등 다양한 통계와 기계학습 알고리즘이 사용되었다. 데이터 분석 시 너무 많은 알고리즘을 구현하고 분석에 해당 알고리즘이 적합한지 테스트 하는데 많은 시간을 소비하고 있다. 이번 실험에서는 딥러닝에 베이스가 되는 인공신경망 알고리즘 하나만 사용하여 많은 알고리즘 구현과 테스트에 소요되었던 시간을 단축하였다. 인공신경망 알고리즘 개발 시간도 단축하기 위하여 구글에서 제공하는 딥러닝 라이브러리인 Tensorflow를 이용하여 [8]에 공개되어 있는 소스 코드를 이번 실험에 맞게 수정하여 사용하였다.

3.5 모델 생성

2차 선별된 데이터 3건을 인공신경망 알고리즘을 이용하여 개봉 일차별 예측 모델 3개를 생성하였다. 일반적으로 트레이닝 데이터와 테스트 데이터에 비율을 80%와 20%로 분리하지만 2차 선별된 데이터에 양이 많지 않아 95%와 5%로 분리하여 사용하였다. 예측 모델 생성 방법은 트레이닝 데이터를 3000번 학습을 시키면서 테스트 데이터를 적용하였을 때 에러율이 가장 낮은 학습 시점에 예측 모델을 저장하였다. (그림 1)은 개봉 7일차 예측 모델 에러율 그래프이다.



(그림 1) 개봉 7일차 예측 모델 에러율

이번 실험에 분석 프로세스는 (그림 2)와 같으며 예측 모델은 ckpt 파일로 저장하여 샘플데이터 적용시 사용하였다.

(그림 2) 분석 프로세스

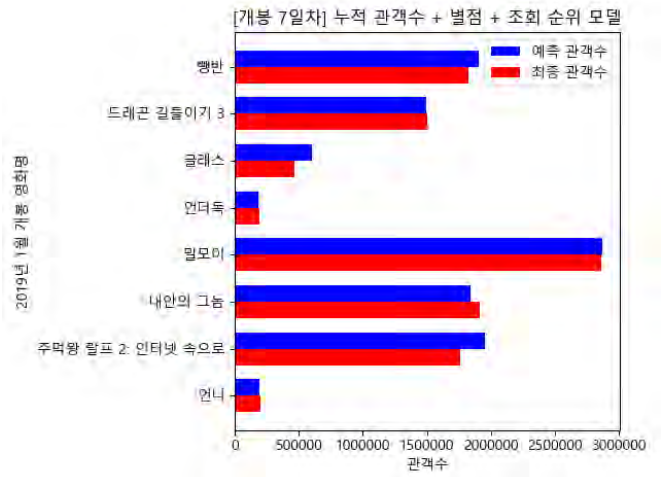
4. 분석

샘플데이터는 2019년 1월 개봉한 영화 중 극장 상영을 종료한 상업영화를 기준으로 5일차, 6일차, 7일차 데이터가 있는 영화로 추출하였다. 영화 극한직업은 분석 시점 현재까지 개봉 상태이기 때문에 제외하였다. 샘플데이터 중에서 최종 관객수 정보가 [4]로 수집된 정보와 영화진흥위원회 통계 정보가 다른 경우 영화진흥위원회에 등록된 최종 관객수 정보로 수정하였다. 샘플데이터 정보는 <표 2>와 같으며 별점, 조회 순위, 누적 관객수 정보는 생략하였다.

<표 2> 예측 모델에 사용한 샘플데이터

영화명	개봉일 (2019년)	최종 관객수
언니	1월 1일	198,377명
주먹왕 랄프 2	1월 3일	1,758,880명
내안의 그놈	1월 9일	1,916,851명
말모이	1월 9일	2,861,365명
언더독	1월 16일	190,128명
글래스	1월 17일	466,475명
드래곤 길들이기 3	1월 30일	1,504,353명
뺑반	1월 30일	1,826,276명

샘플데이터를 예측 모델에 적용한 결과 8개 영화 중 최종 관객수와 예상 관객수에 오차가 가장 적은 모델은 개봉 7일차 모델로 나타났으며 예측 결과는 (그림 3)과 같다.



(그림 3) 개봉 7일차 모델 예측 결과

개봉 7일차 모델 예측 결과를 [2]에 예측률 검증 방법을 활용하여 비교해 보면 <표 3>,<표 4>와 같다.

<표 3> [2] 관객수 예측 결과

영화명	예상 관객수 최종 관객수	RMSE	예측률
베리 굿 걸	124,526명 103,318명	21,208명	79.47%
제보자	1,227,764명 1,258,835명	31,071명	97.53%
드라큘라	492,809명 555,409명	62,600명	88.73%

RMSE = Root Mean Square Error

<표 4> 개봉 7일차 모델 관객수 예측 결과

영화명	예상 관객수 최종 관객수	RMSE	예측률
언니	189,935명 198,377명	8,442명	95.74%
주먹왕 랄프 2	1,951,866 명 1,758,880 명	192,986명	89.03%
내안의 그놈	1,842,265명 1,916,851명	74,586명	96.11%
말모이	2,869,027명 2,861,365명	7,662명	99.73%
언더독	181,755명 190,128명	8,373명	95.60%
글래스	603,940명 466,475명	137,465명	70.53%
드래곤 길들이기 3	1,496,685명 1,504,353명	7,668명	99.49%
뺑반	1,908,558명 1,826,276명	82,282명	95.49%

RMSE = Root Mean Square Error

<표 3>,<표 4>를 비교해 보면 [2]에 결과와 비슷하게 예측률이 높은 영화와 낮은 영화가 섞여 있으며 영화 8개 평균 예측률은 92.71%로 기존 연구에 부족한 수준은 아닌 것을 확인할 수 있었다.
영화 말모이와 뺑반을 기준으로 개봉 일차, 별점, 조회 순위, 누적 관객수, 예상 관객수, 최종 관객수를 정리한 <표 5>, <표 6>를 비교해 보았다. 누적 관객수는 두 영화가 비슷한 수준으로 증가하고 있지만 예상 관객수는 많은 차이를 나타내고 있다. 어떤 특징에 영향을 받아서 예상 관객수 정보가 다른 것인지 확인해 보면 조회 순위는 두 영화 모두 상위권이지만 별점에 차이가 많은 것을 확인할 수 있다. 하나에 비교 샘플로 상관관계를 알아낼 수는 없지만 누적 관객수가 비슷한 경우 별점에 따라 최종 관객수에 영향을 끼칠 수도 있다는 것을 알게 되었다.

<표 5> 영화 말모이 일차별 예측 정보

개봉 일차	별점	조회 순위		
5 일차	9.05	1	누적 관객수	1,184,920명
			예상 관객수	3,153,482명
			최종 관객수	2,861,365명
6 일차	9.04	1	누적 관객수	1,296,957명
			예상 관객수	2,919,429명
			최종 관객수	2,861,365명
7 일차	9.04	1	누적 관객수	1,408,844명
			예상 관객수	2,869,027명
			최종 관객수	2,861,365명

<표 6> 영화 뺑반 일차별 예측 정보

개봉 일차	별점	조회 순위		
5 일차	6.92	2	누적 관객수	963,001명
			예상 관객수	1,510,624명
			최종 관객수	1,826,276명
6 일차	6.81	2	누적 관객수	1,143,906명
			예상 관객수	1,709,167명
			최종 관객수	1,826,276명
7 일차	6.75	2	누적 관객수	1,306,425명
			예상 관객수	1,908,558명
			최종 관객수	1,826,276명

5. 결론 및 향후 연구

본 연구는 데이터분석 프로세스를 최소화 하였을 때 유의미한 정보를 도출해 낼 수 있는지를 알아보기 위함이다. 실험결과는 최종 관객수를 1명 단위까지 정확하게 예측하지는 못했지만 비슷한 수준에 관객수 정보를 예측하였고 별점이 최종 관객수 예측에 영향을 미칠 수 있다는 것을 확인할 수 있었다. 학문적인 접근으로 보았을 때 정확도가 높지 않으면 사용이 불가능한 모델이라고 판단할 수 있지만, 기업의 입장으로 접근해 보았을 때 예측 정보가 [1], [2] 연구 결과에 부족한 수준이 아니고 총 소요된 시간은 기획 3일, 데이터 수집 및 모델 개발 5일, 분석 시간 10분으로 개발 시간 단축, 업무 효율성 향상, 비용 절감을 기대할 수 있다. 따라서 데이터 분석을 할 때 많은 딥러닝 알고리즘을 적용하기 보다는 분석 작업을 위한 업무 분석을 먼저 수행하고 꼭 필요로 하는 특징과 데이터 선별을 통해 빠른 결과를 도출해 보아야 한다. 그리고, 분석된 데이터에 활용 가능 여부를 판단한 후 수정 보완해 가는 반복 작업이 필요하다고 본다.
향후 연구로는 최소 특징에 개수를 늘려가면서 예측 모델을 만들었을 때 어떤 유형에 특징 조합이 높은 예측률을 나타내는지, 최적의 결과를 보여주는 특징에 최소 개수는 몇 개 인지를 알아 보기 위한 연구를 수행할 것이다.

참고문헌

[1] 전성현,손영숙. 데이터마이닝을 이용한 박스오피스 예측. 2016.
[2] 김보경,임창원. GLS와 Bass모형을 결합한 하이브리드 모형을 이용한 영화 관객수 예측. 2018.
[3] 랭키닷컴. http://www.rankey.com
[4] 영화진흥위원회 API. http://www.kobis.or.kr/kobisopenapi/
[5] 네이버 영화 별점. https://movie.naver.com/movie/sdb/rank/rmovie.nhn?sel=cnt
[6] 네이버 영화 조회 순위. https://movie.naver.com/movie/sdb/rank/rmovie.nhn?sel=cur
[7] 영화관 입장권 통합전산망. http://www.kobis.or.kr
[8] tensorflow cookbook neural network. https://github.com/nfmcclure/tensorflow_cookbook/tree/master/06_Neural_Networks