The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 17, No. 3, pp.167-173, Jun. 30, 2017. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2017.17.3.167
JIIBC 2017-3-20

소셜 빅데이터 분석과 기계학습을 이용한 영화흥행예측 기법의 실험적 평가

An Experimental Evaluation of Box office Revenue Prediction through Social Bigdata Analysis and Machine Learning

장재영*

Jae-Young Chang*

요 약 인공지능으로 대표되는 4차 산업혁명에 대한 관심이 증가함에 따라 사회 전반에 빅데이터 및 머신러닝 활용하려는 움직임이 활발해지고 있다. 이러한 움직임은 다양한 분야에서의 예측 시스템 개발로 현실화되고 있다. 특히 영화 산업에서는 투자, 마케팅 등에 활용을 위해 흥행 여부를 사전에 예측하고자하는 여러 가지 시도가 있어왔다. 예전에는 영화에 대한 정적 데이터만을 고려한 예측이 주류를 이뤘으나, 최근에는 실시간으로 생성되는 소셜 데이터를 활용하여 예측하고자하는 노력이 진행되고 있다. 본 논문에서는 영화의 정적 데이터와 더불어 기사, 블로그, 영화평 등 다양한 피드백 정보를 활용한 예측 기법을 제안한다. 또한 제안한 기법을 활용하여 상대적으로 흥행에 성공한 영화만을 대상으로 이들의 흥행정도를 정량적으로 추정할 수 있는지의 여부를 실험적으로 평가하였다.

Abstract With increased interest in the fourth industrial revolution represented by artificial intelligence, it has been very active to utilize bigdata and machine learning techniques in almost areas of society. Also, such activities have been realized by development of forecasting systems in various applications. Especially in the movie industry, there have been numerous attempts to predict whether they would be success or not. In the past, most of studies considered only the static factors in the process of prediction, but recently, several efforts are tried to utilize realtime social bigdata produced in SNS. In this paper, we propose the prediction technique utilizing various feedback information such as news articles, blogs and reviews as well as static factors of movies. Additionally, we also experimentally evaluate whether the proposed technique could precisely forecast their revenue targeting on the relatively successful movies.

Key Words : Box office Revenue, Social Bigdata, Machine Learning, Prediction, Reviews

Ⅰ. 서 론

최근 들어 인공지능, 로봇기술, 생명과학으로 대표되는 4차 산업혁명에 대한 관심이 높아지고 있다. 특히 기계 및 컴퓨터 기술로 인해 대중화되었던 산업자동화 기술은 더 이상 첨단 기술로 대접받지 못하고, 예전에는 상상으로만 가능했던 지능형 시스템이 거의 모든 산업분야에서 관심을 받고 있다. 하지만 아직까지는 이러한 시스템이 성공적으로 활용되는 분야는 매우 제한적이다. 그

*정회원, 한성대학교 컴퓨터공학과
접수일자: 2017년 3월 21일, 수정완료: 2017년 4월 21일
게재확정일자: 2017년 6월 9일

Received: 21 March, 2017 / Revised: 21 April, 2017 /
Accepted: 9 June, 2017
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

이유는 머신러닝(machine learning), 빅데이터(bigdata), IOT 등 지능형 시스템을 위한 기반 연구가 아직 안정화되지 않고 있기 때문이다. 향후 이러한 분야들이 점차 안정화된 기술로 정착된다면 지능형 시스템은 산업 전반에 빠른 속도로 확산될 것이다. 지능형 시스템의 핵심을 구성하는 머신러닝과 빅데이터는 기존의 현상을 분석하는 기술에도 활용되지만 더욱 효과적인 분야는 분류로 대표되는 예측기술이다. 따라서 머신러닝과 빅데이터 연구의 대부분도 예측에 관한 것들이다[1-10].

본 논문에서는 머신러닝과 빅데이터 기술을 이용한 여러 가지 예측 가능한 분야 중에서 영화의 흥행성적을 예측하는 기법을 제안한다. 영화산업은 흔히 도박과 유사한 확률 게임이라고 부른다. 그만큼 성공여부가 매우 불투명하다. 현재 개봉되는 영화들의 대부분은 적자라고 알려져 있는데, 영화사들은 이러한 적자의 위험을 감수하더라도 하나의 큰 수익을 얻는 영화를 만들기 위해 노력한다. 하나의 큰 수익을 얻는 영화가 나머지 손실을 보전할 수 있는 가능성을 믿기 때문이다. 따라서 전적으로 확률에만 의존하는 흥행여부를 어느 정도 사전에 예측할 수 있다면 영화의 소비자는 물론 공급자에게 매우 큰 도움을 줄 수 있을 것이다. 아직까지 신뢰할만한 흥행예측 기법의 개발은 영화산업에 중요한 과제로 남겨져있다. 물론 지금까지 영화흥행 여부를 예측하고자하는 많은 시도가 있어왔다. 하지만 이들의 대부분은 결론이 서로 다르고 정확하지 않다. 그 이유는 영화흥행에 영향을 미치는 변수들이 매우 많을 뿐만 아니라 단기예측의 특성상 영화마다 흥행에 미치는 요소가 서로 다르기 때문이다.

예전에는 배우, 감독, 제작비 등과 같이 영화의 제작단계부터 개봉직후 까지 해당 영화와 관련된 정적 데이터(static data)만으로 예측이 이루어졌다. 여기서 정적 데이터란 정량적인 변화가 없는 고정된 변수들을 의미한다. 물론 이러한 정적 데이터만으로는 만족스러운 예측은 거의 이루어지지 않았다. 일반적으로 영화선택의 가장 큰 기준은 구전효과(word of mouth effect)로 알려져 있다[2]. 구전효과란 흔히 입소문을 말하는데, 주변으로 부터 얻은 비공식적인 정보가 영화선택의 가장 큰 기준이 되는 것이다. 그 이외에도 마케팅이나 언론 보도 등도 흥행에 영향을 미치는 요소로 알려져 있다. 이러한 요소들은 개봉전후 부터 그 이후까지 여러 가지 요인에 의해 그 값들이 유동적으로 변할 수 있기 때문에 동적 데이터(dynamic data)라고 정의한다.

앞서 언급한 바와 같이 예전에는 동적 데이터들을 정량적으로 판단하는 기준이 없어 정적 데이터만으로 예측이 이루어졌다[1]. 최근 들어 인터넷과 SNS의 발달로 인해 구전효과를 간접적으로 정량화할 수 있는 방법들이 개발되고 있다[2-5]. 이에 따라 대부분의 최신 연구들에서도 정적 데이터와 동적 데이터를 결합한 예측 기법들이 주를 이루고 있으며, 그 결과로 예측정확도 측면에서도 많은 발전을 가져왔다[2-10]. 그럼에도 불구하고 이러한 연구들의 문제점은 영화흥행에 영향을 미치는 요소와 예측정확도 측면에서 일관성이 떨어진다는데 있다. 그 이유는 연구 방법의 잘못보다는 개발된 방법을 평가하는데 있어서 실험 데이터를 어떻게 수집했느냐에 따라 그 결과가 다르기 때문이다. 즉, 예측 모델을 만드는데 필요한 영화를 선택하는데 있어서 연구마다 시기, 장르, 규모 등 영화 정보에 대한 수집 데이터에 큰 차이를 보이고 있다.

본 논문에서는 기존의 연구와 유사하게 영화에 대한 정적 데이터 및 동적 데이터를 모두 활용하고 이들 중에서 어떠한 요소가 흥행에 가장 관련되어 있는지를 탐색하고 예측 정확도를 측정한다. 다만 본 논문에서는 비교적 이름이 알려진 영화만을 대상으로 한다. 즉, 일정기간 동안의 관람객 기준 상위 100개 영화만을 대상으로 예측 모델을 생성한다. 그 이외의 영화들은 선택된 상위 100여 개의 영화들에 비해서 투자규모 등 정적 요인 면에서 많은 차이가 나므로 이들을 예측 모델에 포함하는 것은 큰 의미가 없다. 기존에는 이러한 것들까지 포함하여 정확도가 높다고 주장하였으나, 규모면에서 편차가 큰 영화들을 대상으로 한 예측은 큰 의미가 없다. 왜냐하면, 투자비용 측면에서 투자비나 마케팅 비용이 큰 영화는 손익분기점만큼의 관객을 모으지 못해 흥행에 실패했다하더라도, 마케팅이나 언론에 자주 오르내리기 때문에 개봉초기에 어느 정도 규모의 관객을 모으는 것이 가능하기 때문이다. 하지만 본 논문에서는 상위 100개의 영화만으로 대상으로 하였으므로 흥행요인이 무엇인지를 좀 더 정교하게 분석할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 기술하고 3장에서는 예측모델을 만들기 위한 데이터 수집과 전처리 기법을 소개한다. 4장에서는 예측 모델을 생성하는 기법과 실험 결과를 기술하고, 마지막으로 5장에서는 결론을 맺는다.

II. 관련연구

국내외에서 영화흥행을 사전에 예측하려고하는 시도는 예전부터 있어왔다. 우선 국외의 연구동향을 살펴보면 비교적 초기연구인 [1]에서는 정적 데이터만을 대상으로 예측 방법을 연구하였다. 이 당시는 SNS가 활성화되지 않아 구전효과를 정량적으로 평가할 방법이 없었다. 이 연구에서는 정적 데이터 중에서 영화배우, 작가, 감독 등 인적요인이 가장 중요하다고 판단하였다. [2]에서는 포털의 영화정보 사이트에서 관객들의 영화평을 중심으로 흥행여부를 예측하였다. 여기서는 영화평에서의 정량적 평점과 비정형 데이터를 이용한 감성분석(sentiment analysis) 결과도 활용하였지만 이들은 흥행과는 무관하다고 밝히고, 영화평의 규모, 즉 대중의 관심도가 흥행에 직접적으로 연관이 있다고 결론을 내렸다. [3,4,5]에서는 블로그나 트위터(twitter)에서의 영화에 대한 감성분석을 통해 흥행여부를 예측하였고, 감성분석 결과가 흥행과 연관성이 있음을 밝혔다. [2]와 [3,4,5]는 감성분석이 흥행여부를 예측하는 중요한 요소인가를 결정하는데 있어서 서로 반대되는 결과를 보이는데, 이는 대상 데이터, 시기, 분석 방법 등에 의한 차이로 보이며 어느 연구가 더 신뢰할만한지는 현 단계에서 명확히 밝히기 쉽지 않다. 이외에 소셜 미디어 분석 전문업체인 피지올로지(Fizziology)에서는 SNS를 분석하여 영화개봉 전후에 걸쳐 영화흥행 예측결과를 랭킹 순으로 제공하고 있다. 이 업체는 90% 이상의 높은 정확도로 예측이 가능하다고 주장하고 있으나 공식적인 검증은 이루어지지 않고 있다,

국내에서도 다양한 방법으로 영화흥행을 예측하려는 시도가 있어왔다. 우선 [6]에서는 영화개봉일과 개봉 1주일 후를 기준으로, 정적 데이터(배우, 감독, 스크린수, 제작사 등)와 동적 데이터(관련 트위터 수, 관객 평점, 영화평 수 등의 소셜 데이터) 중 어느 정보가 더 흥행에 영향을 미치는가를 분석하였고, 정적 데이터와 동적 데이터와 모두 포함된 결과가 영향력이 크다는 결론을 내렸다. 예측 실험은 특정 시점에 무작위로 영화를 선택하여 실시하였다. 따라서 학습이나 테스트로 활용되는 영화들의 흥행 성적이 서로 간에 격차가 매우 커서 예측 정확도가 비교적 높았으나 상식선에서 판단할 수 있는 정도의 예측력을 벗어나지는 않는다. [7]에서는 국내 빅데이터 분석 업체인 티버즈(TIBUZZ)와 펄스K(PulseK)를 이용하여 영화 흥행과 관련된 주요 단어와의 관련성을 분석하였다. 그 결과 배우, 스토리, 감독과 같은 단어들이 영화 흥행과 관련이 있음을 밝혀냈다. [8]에서는 영화평에 언급된 영화의 특징(배우, 감독, 스토리, 효과, 음악 등)들이 영화흥행과 어떠한 연관이 있는가를 분석하였고, 결과적으로 스토리, 효과, 배우에 대한 언급량과, 이들 단어들에 대한 긍정적인 언급여부가 흥행과 관련 있다고 결론을 내렸다. 다만 이 연구도 특정 연도의 흥행 성적이 상위인 영화부터 하위 영화까지 폭넓게 선택한 영화들을 대상으로 진행하였으므로 예측 정확도가 높을 수밖에 없었다. [9]에서는 영화 개봉전의 정적 데이터 보다 개봉후의 동적 데이터인 블로그 수, 뉴스 언급량, 영화평 평가자 수가 더 중요한 변수임을 실험적으로 증명하였다. 하지만 이러한 변수들이 개봉후 1~3달이 지난 시점에 수집된 데이터이다. 이는 흥행 여부가 거의 결정 난 다음의 데이터를 이용하는 것이므로 흥행과의 관련성이 있는 것인 당연할 뿐만 아니라 개봉이 한참 지난 후의 데이터를 이용한다는 것은 개봉전후에 판단할 예측인자로서의 가치가 없음을 의미한다. [10]에서는 구전효과를 흥행요소에 반영하기 위해 온라인 버즈량을 고려하였다. 그 이외에 다양한 정적, 동적 변수를 활용하였으나, 배우, 개봉 스크린수, 배급사영향력, 온라인 버즈량 등이 흥행 여부와 관련 있다고 결론을 내렸다. 그러나 여기서도 50만 이하의 영화와 500만 이상의 영화 등 영화 종류와 규모에 관계없이 실험데이터를 수집해서 예측에 활용하였다.

이와 같이 기존이 연구들은 국내외를 막론하고 정적 데이터와 동적 데이터를 이용하여 흥행의 인자가 무엇인지를 탐색해왔다. 본 논문에서도 이와 유사한 인자들을 사용하여 흥행예측 가능여부를 실험하였다. 다만 기존 연구와는 다르게 특정 기간의 흥행 상위 영화만으로 대상으로 하여, 비교적 흥행에 성공한 영화들 사이에 흥행을 판별하는 요인이 무엇인지를 세밀하게 분석하였다.

III. 데이터 수집 및 전처리

예측 모델의 생성하기 위한 첫 단계로 기계학습을 위한 데이터들을 수집하였다. 서론에서 언급한 바와 같이 영화와 관련된 데이터는 정적 데이터와 동적 데이터로 나눌 수 있다. 정적 데이터는 영화사, 배우, 감독, 초기 스크린 수 등 영화 개봉 전에 변동 없이 결정된 사항들이다. 본 논문에서는 개봉 1주일후의 스크린 수 와 관객 수도

그림 1. 영화진흥위원회에서의 영화정보 검색
Fig. 1. Movie Information Retrieval in Korean Film Council

정적 데이터로 간주하였다. 동적 데이터는 영화 개봉 직전 또는 이후에 흥행과 관련되어 영화 관계자이외의 주체(관객, 소비자, 기자 등)에 의해 생성되는 관련 데이터를 말한다. 예를 들어 블로그 수, 영화평 관련 데이터, 언론 노출빈도 등을 들 수 있다.

본 논문에서는 대상 영화 2009년도부터 6년간 국내 박스오피스 최상위 100개의 영화를 대상으로 하였다. 즉 6년 동안 비교적 흥행에 성공한 100개의 영화를 대상으로 예측모델을 생성하여 평가하였다. 우선 국내 영화들에 대한 정적 데이터의 대부분은 영화진흥 위원회 홈페이지로 부터 수집하였다. 영화진흥위원회에서는 그림 1과 같이 개봉영화에 대한 다양한 검색이 가능하며, 각종 정보를 다운받을 수 있다.

이와 같이 수집된 각 영화별 정적 데이터 리스트는 표 1과 같다. 이 표의 정적 데이터 중에서 감독과 배우는 인물명이 아닌 해당 인물이 과거 출연했던 영화에서의 평균관객수로 표현하였다. 감독과 배우의 관객동원 능력을 정량적으로 표현하기 위해서다. 최종 관객수는 예측대상이 되므로 모델 생성을 위한 기계학습과정에서는 종속변수로 활용된다. 본 연구에서는 제작비도 흥행예측에 중요한 변수로 판단하였으나 영화진흥위원회 홈페이지에서는 확보가 불가능해서 배제하였다.

표 1의 동적 데이터는 대부분 네이버 포털을 이용하여 수집하였다. 본 연구에서는 예측 시점을 개봉 1주일 이후로 설정하였다. 따라서 동적 데이터는 수집 시점의 데이터가 아닌 각 영화의 개봉 1주일 후를 기점으로 그 이전에 게시된 데이터만을 사용하였다. 개봉일이나 그 이전에는 사실상 동적 데이터가 거의 없을 뿐만 아니라 어느 정도 존재한다 하더라도 구전효과로 생성된 데이터라기보다는 제작사의 홍보에 의해 생성된 데이터일 가능성이 매우 높다. 따라서 개봉일 이전의 흥행 예측은 정적 데이터만을 사용한 예측만이 가능하므로 정적/동적 데이터를 모두 활용하기 위해서는 개봉 1주일이 지난 시점에 흥행을 예측하는 것이 가장 적절할 것으로 판단하였다.

마지막으로 예측 대상이 되는 최종 관객수는 등급으로 처리되었다. 회귀분석과 같은 예측 기법을 사용한다면 예상 관객 수를 정확한 수치로 추정할 수 있으나 상대적 예측 정확도를 비교할 근거가 부족하여 관객 수에 따라 여러 단계로 등급을 나누어 흥행 등급을 예측하는 방식으로 실험하였다. 수집된 100개의 영화에서 최저 관객수는 약 184만 명이고 최고는 1,760만 명이다. 따라서 흥행 등급은 표 2와 같이 5개의 등급으로 분류하여 해당 등급을 예측하는 방식으로 실험을 실시하였다.

표 1. 수집된 정적/동적 데이터 리스트
Table 1. Collected Static/Dynamic Data List

정적 데이터	동적 데이터
개봉일 배급사 감독(평균관객수) 배우(평균관객수) 제작사 최초 스크린수 1주일 후 스크린수 1주일 후 누적관객수 장르 상영시간 관람등급 최종관객수(예측대상)	관련 뉴스기사 수 블로그 수 영화평 평점 분포 영화평 수 단, 개봉 1주일 이전에 생선된 데이터만 수집

표 2. 종속 변수 정의
Table 2. Dependent Variable Definition

흥행등급	단위(만)
A	1,000 ~
B	750 ~ 1,000
B	500 ~ 750
D	200 ~ 500
E	~ 200

Ⅳ. 예측 모델 생성 및 실험 결과

예측 기법은 대부분 데이터마이닝의 분류기법을 활용하였다. 대표적인 분류기법에는 결정트리, KNN(K-Nearest Neighbor), SVM(Support Vector Machine), 나이브베이즈 분류(Naïve Bayes Classification), 신경망(Neural Network) 등이 있으나, 본 논문에서는 대표적인 확률모델인 나이브베이즈 분류와 요즘 주목받고 있는 신경망을 이용하여 평가하였다. 독립변수로는 정적 데이터의 조합, 동적 데이터의 조합, 정적/동적 데이터의 조합으로 분류하였다. 각 조합들은 독립변수인 최종 관객 수와의 상관분석을 통하여 비교적 상관관계를 보이는 변수들 위주로 조합하였다. 우선 표 3은 정적 데이터만으로 조합한 독립 변수들이다. 여기서 S1은 대부분의 정적 데이터를 포함이고 S6로 갈수록 상관관계가 분명한 변수만으로 조합을 생성하였다. 표 4는 동적 데이터로 구성된 변수 조합을 보여준다. 동적 데이터는 종류가 많지 않아 다양한 조합으로 실험을 실시하였다. 마지막으로 표 5는 정적과 동적 데이터를 모두 포함한 조합을 보여준다. 여기서 정적과 동적 데이터는 표3 과 표 4를 이용한 실험에서 비교적 정확도가 높았던 변수 조합을 이용하여 구성하였다. 이와 같이 표 3부터 표 5까지 총 15가지의 변수 조합을 실험을 실시하였으며, 앞서 언급한 바와 같이 분류기법은 나이브베이즈 분류, 신경망을 이용하였다. 또한 실험 평가를 위해 100개의 영화에 대해서 무작위로 70%의 영화를 기계학습을 위한 데이터로 활용하고 나머지 30%를 테스트용으로 사용하였으며, 정확한 실험을 위해 수차례 실험을 반복하여 평균값으로 분류 정확도를 측정하였다.

표 3. 정적 데이터 조합
Table 3. Combination of Static Data

모델명	정적 데이터만으로 구성한 독립변수 리스트
S1	배급사, 감독, 배우, 제작사, 초기 스크린 수, 1주일 후 스크린 수, 1주일 후 누적 관객 수, 장르, 상영시간, 관람등급
S2	배급사, 감독, 배우, 제작사, 1주일 후 스크린 수, 1주일 후 누적관객수, 상영시간, 관람등급
S3	배급사, 감독, 배우, 1주일 후 스크린 수, 1 주일 후 누적 관객 수
S4	배급사, 감독, 배우, 1 주일 후 누적 관객 수
S5	배급사, 감독, 배우
S6	감독, 배우

표 4. 동적 데이터 조합
Table 4. Combination of Dynamic Data

모델명	동적 데이터만으로 구성한 독립변수 리스트
D1	영화평 수, 영화평 평점, 블로그 수, 뉴스 기사 수
D2	영화평 수, 영화평 평점, 뉴스 기사 수
D3	평점, 기사, 블로그 수
D4	영화평 수, 영화평 평점
D5	영화평 수, 블로그 수, 뉴스 기사 수

표 5. 정적/동적 데이터 조합
Table 5. Combination of Static/Dynamic Data

모델명	정적/동적 데이터로 구성한 독립변수 리스트
SD1	뉴스 기사 수, 블로그 수, 영화평 평점, 영화평 수 + 배급사, 감독, 배우, 1 주일 후 누적 관객 수
SD2	리뷰 개수, 영화평 평점, 뉴스 기사 수 + 감독, 배우
SD3	뉴스 기사 수, 영화평 평점, 영화평 수 + 배급사, 감독, 배우, 1 주일 후 누적 관객 수
SD4	뉴스 기사 수, 영화평 평점, 영화평 수 + 감독, 배우

실험 결과는 그림 2~4와 같다. 그림 2, 3, 4는 각각 표 3, 4, 5의 조합에 의한 실험 결과이다. 우선 전체적으로는 나이브베이즈 분류 방법에 비해서 신경망을 이용한 예측 기법이 평균적으로 약 10%정도 정확도가 높게 측정되었다. 이는 신경망이 나이브베이즈 분류에 비해 더 좋은 기법이라는 것을 의미하는 것은 아니고 독립변수들의 데이터 분포가 신경망에 더 적합하게 형성되어 있기 때문인

것으로 판단된다. 구체적으로 그림 2를 보면 나이브베이즈 분류가 39%~46%의 정확도를 보인반면 신경망은 45%~61%의 정확도를 보였다. 특히 신경망의 경우에는 두개의 독립변수인 감독과 배우로 예측하는 것이 61%로 가장 높게 예측되었다. 이는 기존의 연구와 결과가 거의 일치하는 것으로 정적 데이터 중에서는 인적 구성이 흥행에 가장 큰 영향을 미친다는 것을 의미한다. 하지만 영화산업의 특성상 유명배우나 감독이 참여하는 영화의 경우에는 다른 요인(제작사, 배급사, 스크린 수 등)도 상대적으로 비례하는 경우가 많으므로 실험결과만으로 단정적으로 판단하기에는 무리가 있다. 다음으로 그림 3에서는 나이브베이즈 분류와 신경망이 각각 45%~52%, 42%~55%의 정확도를 보여 유의한 차이를 보이지 않고 있다. 다만 D1, D2, D4와 같이 영화평 평점과 기타 변수들이 조합된 경우가 상대적으로 좋은 정확도를 보였다. 또한 그림 3의 전반적인 예측 정확도를 보면 정적 데이터만으로 예측한 그림 2와 비교하여도 뚜렷한 차이를 보이지 않고 있다. 즉, 동적 데이터를 이용한 예측과 정적 데이터를 이용한 예측이 정확도 측면에서 서로 유사하다고 볼 수 있다. 마지막으로 그림 5의 경우에는 나이브베이즈 분류와 신경망이 각각 45%~53%, 46%~68%의 정확도를 보였다. 특히 신경망의 경우 SD4가 68%의 정확도를 보여 실험결과 중에서 가장 좋은 정확도를 보였다. SD4는 감독, 배우, 영화평 평점 등 그림 2와 3에서 중요한 변수로 평가된 것들을 모두 포함한 경우였다. 결론적으로 본 연구에서 실험한 결과로는 주요한 정적 데이터와 동적 데이터를 모두 포함하는 것이 영화흥행을 예측하는 데 가장 좋은 방법이라고 결론을 내릴 수 있다.

본 논문의 실험결과를 이전의 관련연구들과 비교해보면 본 논문의 예측 정확도가 상대적으로 낮게 나타나는 경향을 보이고 있다. 이는 앞서 언급한 바와 같이 본 논문의 예측기법에 대한 성능이 떨어진다기보다는 데이터간의 편차가 크지 않기 때문인 것으로 분석된다. 즉 실험 데이터간의 흥행성공 여부가 다른 연구들의 데이터처럼 뚜렷하지 않아 보다 정확도가 정량적으로는 떨어지나 이는 등급 분류가 세밀하고, 영화간의 특징 차이가 크지 않아 발생된 것으로 해석할 수 있다. 다만 본 논문의 결과가 상업적으로 활용될 수 있으려면 서로 규모가 서로 유사한 영화라 하더라도 이를 예측할 수 있는 독립변수를 최대한 발굴하고, 최신 예측 기술인 딥러닝(deep learning) 기법을 도입하여 예측 정확도를 높이는 것이 필요하다고 판단된다.

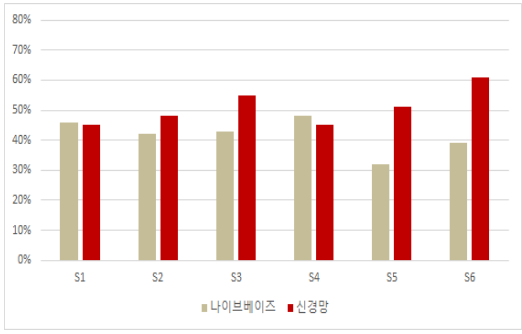

그림 2. 정적 데이터를 이용한 예측결과
Fig. 2. Prediction Results Using Static Data

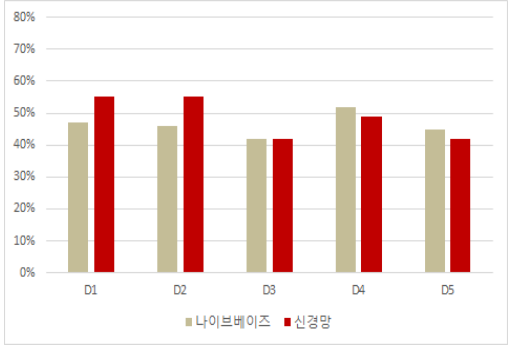

그림 3. 동적 데이터를 이용한 예측결과
Fig. 3. Prediction Results Using Dynamic Data

V. 결론

본 논문에서는 머신러닝 기법을 이용하여 영화의 흥행성적을 예측하는 기법을 제안하였고 그 결과를 실험적으로 평가하였다. 예측 모델을 생성하기 위해서 영화와 관련된 정적 데이터와 동적 데이터를 수집하였다. 예측 모델은 수집된 데이터의 다양한 조합으로 생성하였으며 나이브베이즈 분류와 신경망을 이용하여 예측 정확도를 평가하였다. 실험 결과 본 논문이 수집한 데이터에 대해서는 신경망이 나이브베이즈 분류보다 더 좋은 정확도를 보였으며, 배우, 감독과 같은 정적 데이터와 영화평 평점, 뉴스기사 수, 블로그 수 등 동적 데이터를 조합한 모델이 가장 좋은 성능을 보였다. 또한 실험 데이터 구성에 있어서 규모면에서 유사한 영화들임에도 불구하고 최대 예측

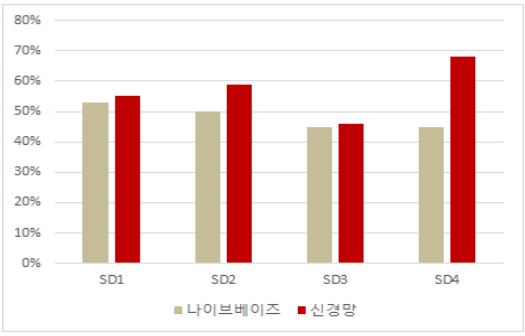

그림 4. 정적/동적 데이터를 이용한 예측결과
Fig 4. Prediction Results Using Static/Dynamic Data

정확도가 68%를 보여 실용적 적용 가능성도 보였다. 하지만 아직까지는 개봉 초기에 흥행정도를 정확히 판단하는 것은 제한적이라고 판단된다. 이를 극복하기 위해서는 정교한 예측을 위한 새로운 변수들에 대한 발굴이 필요하며 딥러닝과 같은 최신 예측 기술 적용가능성도 검토해야할 것으로 판단된다.

References

[1] S. Albert, “Movie Stars and the Distribution of Financially Successful Films in the Motion Picture Industry,” Journal of Cultural Economics, Vol.22, No.4, pp.249-270, 1998.

[2] Y. Liu, “Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue,” Journal of Marketing, Vol.70, No.3, pp.74-89, 2006.

[3] G. Mishne and N. S. Glance, “Predicting Movie Sales from Blogger Sentiment,” In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp.155-158, 2006.

[4] L. Lica and M. Tuta, “Predicting Product Performance with Social Media,” Informatica Economica , Vol.15, No.2, pp.46-56, 2011.

[5] S. Asur and B. A. Huberman, “Predicting the future with social media,” Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010. p. 492-499.

[6] J. Yim and B. Hwang, Predicting Movie Success based on Machine Learning Using Twitter, KIPS transactions on Software and Data Engineering, Vol. 3, No. 7, pp.263-270. 2014

[7] O. Lee et al. Movie Box office Analysis using Social Big Data, J. of The Korea Contents Association, Vol. 14, No. 10, 2014

[8] S. Cho et al. Predicting Movie Sales through Online Review Mining, Proceedings of the Korea Society of Management Information Systems Conference, 2014

[9] S. Jeon and Y. Son, Effect of Online Word-of-Mouth variables as Predictors of Box Office, The Korea Journal of Applied Statistics, Vol. 29, No. 4, pp. 657-678, 2016

[10] Y. Kim and J. Hong, A study for the Development of Motion Picture Box-Office Prediction Model, J. of The Korean Statistical Society, Vol. 18, No. 6, pp. 859-869, 2011.

[11] S. Lee, J. Cho, C. Kang, and S. Choi, Study on Prediction for a Film Success Using Data Mining, J. of the Korean Data and Information Science Society, Vol. 26, No. 6, pp.1259-1269, 2015.

저자 소개

장 재 영(정회원)

- 1992년: 서울대학교 계산통계학과 (이학사)
- 1994년: 서울대학교 계산통계학과 (이학석사)
- 1999년: 서울대학교 계산통계학과 (이학박사)
- 2000년~현재: 한성대학교 컴퓨터공학과 교수

<주관심분야 : 데이터베이스, 정보검색, 데이터마이닝>

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.