The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1257
29(7), 1257–1270

Prediction of box office using data mining

Seonghyeon Jeon[a] · Young Sook Son[a,1]

[a]Department of Statistics, Chonnam National University

(Received July 22, 2016; Revised September 17, 2016; Accepted October 20, 2016)

Abstract

This study deals with the prediction of the total number of movie audiences as a measure for the box office. Prediction is performed by classification techniques of data mining such as decision tree, multilayer perceptron(MLP) neural network model, multinomial logit model, and support vector machine over time such as before movie release, release day, after release one week, and after release two weeks. Predictors used are: online word-of-mouth(OWOM) variables such as the portal movie rating, the number of the portal movie rater, and blog; in addition, other variables include showing the inherent properties of the film (such as nationality, grade, release month, release season, directors, actors, distributors, the number of audiences, and screens). When using 10-fold cross validation technique, the accuracy of the neural network model showed more than 90 % higher predictability before movie release. In addition, it can be seen that the accuracy of the prediction increases by adding estimates of the final OWOM variables as predictors.

Keywords: data mining, decision tree, multilayer perceptron(MLP) neural network, multinomial logit model, online word-of-mouth(OWOM), prediction of box office, support vector machine, 10-fold cross validation

1. 서론

영화진흥위원회(Korean Film Council; KOFIC, 2016)의 한국 영화산업 결산에 따르면 2014년부터 연속 2년 동안 영화산업 매출액은 2조원을 돌파하였고, 2013년부터 연속 3년 동안 극장 총 관객수는 2억명을 돌파하였다. 글로벌 산업정보조사기관인 IHS 자료에 따르면 2015년 우리나라 인구 1인당 연간 평균 영화 관람 횟수는 4.2회로 세계 최고 수준임을 보인다. KOFIC (2015)의 표본수 2006명에 대한 영화 소비자 조사 결과에 의하면 전국 만 15세 이상 59세 이하 소비자들의 2015년 1년간 극장 영화 관람률은 94.2%이며 극장 영화 관람 편수는 년 평균 8.6편으로 조사되었다. 한국영화 사상 첫 천만 관객 영화였던 2003년에 개봉된 '실미도' 이후 13년 동안 총 13편의 천만 관객 한국영화가 나왔다. 그 중 약 62%인 8편이 2012년부터 최근 4년 동안 매년 2편씩 천만 관객을 동원하였다. 특히 2014년에 개봉한 영화 '명량'은 우리나라 인구의 약 1/3이 넘는 1,761만 명이 관람하여 역대 1위의 관객수를 기록하였다. 이러한 통계수치들의 흐름은 최근 한국 영화산업의 환경을 낙관적으로 볼 수 있는 청신호임은 분명하다. 그러나 개별 영화의 수익 면에서 살펴보면 밝지만은 않다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0022864).
[1]Corresponding author: Department of Statistics, Chonnam National University, 77, Yongbong-ro, Buk-Gu, Gwangju 61186, Korea. E-mail: ysson@jnu.ac.kr

KOFIC (2016)의 한국 영화산업 결산에 따르면 2015년에 극장에서 개봉한 한국영화 232편 중에서 총 제작비 10억 원 이상이거나 전국 개봉 스크린 수가 100개 이상인 작품으로 분류되는 상업영화 73편의 평균 총 제작비는 52.3억 원이며 제작비의 구성 비율을 살펴보면 약 70%가 순수 제작비이며 약 30%는 마케팅비로 구성되어 있다. 73편의 상업영화 중 손익분기점을 상회하는 비율은 21.9%, 수익률 100%를 상회하는 비율은 9.6%, 수익률 −90%를 하회하는 비율은 23.3%에 달한다. 영화산업은 고수익, 고위험의 특성을 가지고 있기 때문에 가능한 초기 단계에서 영화 흥행을 정확히 예측하는 것은 영화산업 관련 업계 종사자들에겐 매우 중요한 과제일 것이다.

영화산업에서의 수익은 크게 극장 입장권 판매에 따른 극장 매출, IPTV, 케이블 TV, 인터넷 VOD 서비스 등의 매출에 따른 디지털 온라인 시장 매출, 그리고 완성작 및 서비스의 해외 수출에 따른 해외 매출의 3가지로 나눌 수 있다. KOFIC (2016)의 한국 영화산업 결산에 따르면 2015년 한국 영화산업의 총 매출액은 2조 1,131억 원으로 나타났고 그 중 극장 매출액이 1조 7,154억 원으로 전체 매출액의 약 81.2%를 차지하고 있으며 그 밖에 디지털 온라인 시장 매출액이 15.8% 그리고 해외 매출액이 3%를 차지하고 있다. 극장 매출 수익은 영화산업 전체 수익의 80% 이상을 차지하고 있을 뿐만 아니라 디지털 온라인 시장 매출 및 해외 매출은 극장 매출에 직접적인 영향을 받는 부가 매출이기 때문에 극장 매출 수익을 정확히 예측하는 것이 영화 흥행 예측의 핵심일 것이다.

그동안 영화 흥행을 예측하는 연구들에서는 주로 영화 속성을 나타내는 변수인 스크린 수, 등급, 국적, 장르, 감독, 배우, 배급사 등을 예측변수로 활용하여 영화 흥행을 예측하는 경우가 많았다. 최근 소셜미디어의 급격한 발달로 인해 영화 포털사이트의 커뮤니티가 활성화되고 블로그, 뉴스 등을 통한 온라인 구전의 영향력이 커짐에 따라 소셜미디어를 활용한 영화 흥행 예측의 연구들이 많아지고 있다. 실제 KOFIC (2015)의 영화 소비자 조사에 따르면 소비자들의 관람 영화 선정 시 주로 인터넷(67.1%)을 통해 정보를 획득하며 다음으로는 주변인의 의견(61.1%), 영화 광고(54.7%)의 순서였다. 인터넷을 통한 정보획득 중에서는 포털 사이트의 영화 섹션(29.2%), 포털 사이트 뉴스 섹션(19.2%), 그리고 블로그(9.7%)로 부터 영화 정보를 취득하며 이들 소셜미디어들이 인터넷 정보 획득의 약 58.1%를 점유한다. 따라서 포털 사이트의 영화 평점 및 평가자 수, 뉴스 수, 그리고 블로그 수와 같은 온라인 구전(online word-of-mouth; OWOM) 변수들이 온라인 입소문 효과로써 영화 흥행에 많은 영향을 미칠 것이라 판단된다.

본 연구에서는 영화 흥행 예측을 다룬다. 일반적으로 영화 흥행과 관련된 연구는 영화 흥행에 영향력을 미치는 예측변수들의 선택에 관한 연구와 이들 예측변수들로 부터 영화 흥행을 예측하는 연구의 두 가지 주제로 분류된다. 앞으로 소개할 이전 연구들은 영화 흥행 예측을 목적으로 적용한 예측모형에 대한 적합도 혹은 예측의 정확도를 나타낸 연구들이다. 이들 연구에서 모형의 적합도와 예측의 정확도는 구별되어야 한다.

모형의 적합도는 가지고 있는 모든 데이터로 추정한 모형에 의한 목표변수의 추정치와 실제 목표변수 값과의 일치도를 의미한다. 즉, 모형의 훈련과 목표변수 예측에 같은 자료가 사용된다. 예측을 다루는 문제에서는 적합도가 높은 모형이 예측의 정확도도 높을 가능성은 있으나 과적합으로 인하여 새로운 자료에 대해서도 항상 예측의 정확도가 높다고 보장할 수는 없다. 따라서 예측을 다루는 문제에서는 교차 검증을 통한 예측모형의 예측력을 검증하는 것이 필수적이다. 교차 검증의 대표적인 방법으로 k-중 교차검증(k-fold cross validation)이 있다. 이 방법은 관측된 데이터들을 랜덤하게 k개의 집단으로 나눈 뒤, $k-1$개의 집단을 모형의 훈련에 사용하고 나머지 1개의 집단을 예측에 사용하여 모형의 정확도를 측정한다. 이러한 절차를 독립적으로 k회 수행하여 각 회차에서 얻은 정확도를 평균하여 모형의 전체 정확도로 표현한다. k-중 교차 검증은 모든 자료가 모형 훈련과 예측에 분리되어 독립적으로 사용된다는 장점이 있고 k회 반복 수행함으로써 예측 모형의 신뢰성을 높일 수 있다.

Sharda와 Delen (2006)은 ShowBIZ 사이트에 기록된 박스오피스 시장에서 1998년부터 2002년까지 개봉된 834편의 영화를 대상으로 순수익을 10개의 범주로 구분한 범주형 목표변수를 사용하여 순수익을 예측하였다. 등급, 장르, 스타 가치, 특수 효과, 속편 여부, 스크린 수와 같은 영화 속성변수들을 예측변수로 사용하여 다항로짓모형, 판별분석, CART, 그리고 multilayer peceptron(MLP) 신경망모형에 의해 순수익을 예측하였다. 교차 검증을 위한 10-중 교차 검증에 의해 예측력을 비교한 결과 신경망모형의 정확한 분류가 36.9%, ±1의 범주 내 분류가 75.2%로 가장 우수하였다.

Zhang 등 (2009)은 중국 박스오피스 시장에서 2005년부터 2006년까지 개봉된 241편의 영화를 대상으로 순수익을 6개의 범주로 구분한 범주형 목표변수를 사용하여 순수익을 예측하였다. 예측변수로는 영화 속성변수인 국적, 감독, 배우, 장르, 경쟁 영화 수, 개봉 일, 영화관 수, 스크린 수 등을 사용하였고, 교차 검증을 위한 6-중 교차 검증에 의해 예측력을 비교한 결과 기존의 MLP 신경망모형에 비해 back propaganda(BP) 신경망모형의 정확도가 더 높음을 보였고 이때 정확한 분류는 68.1%, ±1의 범주 내 분류는 97.1%의 정확도를 보였다.

Kim과 Hong (2011)은 한국 박스오피스 시장에서 2010년 1월부터 12월까지 개봉된 316편의 영화를 대상으로 총 관객수를 5개의 범주로 구분한 범주형 목표변수를 사용하여 총 관객수를 추정하였다. 감독, 배우, 국적, 장르, 등급, 스크린 수, 배급사와 같은 영화 속성변수, 그리고 OWOM 변수에 해당되는 네이버 포털 평점, 블로그, 트위터 등 소셜미디어로부터 생성되는 온라인 버즈(buzz)의 크기를 예측변수로 사용하여 모형을 적합시켰다. 판별분석에 비해 신경망모형과 다항로짓모형에서 적합도가 높게 나타났고 특히 다항로짓모형이 95.1%의 적합도를 보였다.

Kim 등 (2013)는 한국 박스오피스 시장에서 2011년 10월부터 2012년 8월까지 개봉된 영화 103편 중에서 상영기간, 총 관객수 등을 기준으로 선별된 47편의 영화를 대상으로 양적 목표변수로서 총 수익 및 개봉 t주($t = 1, 2, 3, 4$) 후 수익을 예측하였다. 이를 위해 예측변수로는 각 시점 별 영화의 누적 수익, 스크린 수, 좌석 수, 트위터와 페이스북의 SNS 데이터에 대한 긍정 혹은 부정 의견 수를 예측변수로 사용하였다. 각 시점에서 수익을 예측할 때 예측 시점에서 수집 가능한 데이터를 사용하였다는 점에서 실제 예측 시점의 현실적 환경을 제대로 반영한 분석 방법이라 평가할 수 있다. 교차 검증을 위해 47-중 교차 검증에 의해 예측력을 비교한 결과 mean absolute error(MAE)의 관점에서는 회귀모형이 BASS diffusion모형에 비해 우수하였으나 root mean square error(RMSE) 관점에서는 1주 후를 제외하고 BASS diffusion 모형이 회귀모형에 비해 우수하였다.

Song과 Han (2013)은 한국 박스오피스 시장에서 2008년부터 2011년까지 개봉된 대략 505개의 영화들 중 순수익이 5억 원 이상인 206편의 영화를 대상으로 순수익을 백분위수에 따라 10개 범주로 나눈 후 각 범주에 1점부터 10점까지 준 점수를 양적 목표변수로 사용하여 순수익을 예측하였다. 예측변수로서 영화장르, 등급, 속편 여부, 감독, 배우, 명절 개봉 일 여부, 방학 개봉 일 여부, 개봉 월, 개봉 월 평균 기온, 국내 영화의 참여비율 등과 같은 영화 속성변수 만을 사용하였으며 선형모형, random forests model, gradient boosting model을 예측모형으로 사용하였다. 자료 중 랜덤하게 70%는 모형 훈련에, 30%는 예측에 사용되었으며 이와 같은 교차 검증을 총 1,000번 반복하여 계산된 mean square error(MSE)의 평균 관점에서 gradient boosting model이 가장 우수하였다.

Yim과 Hwang (2014)은 한국 박스오피스 시장에서 2013년 4월부터 10월까지 개봉된 영화들 중 무작위로 60편의 영화를 선택하여 총 관객수를 5개의 범주로 구분한 범주형 목표변수를 사용하여 총 관객수를 추정하였다. 예측변수로서 개봉 일, 등급, 상영 시간, 감독, 배우, 국적 등과 영화 속성변수와 네이버 포털 평점, 평가자 수, 개봉 1주 전과 1주 후에 트위터 내에서 해당 영화가 언급된 비율과 같은 OWOM 변수를 사용하였다. 나이브 베이지안 분류 기법을 사용하였을 때 개봉 일에 78.3%, 개봉 1주일 후에 95%의 적합도를 보였다. Kim과 Hong (2011) 그리고 Yim과 Hwang (2014)에서 사용한 네이버 포털

평점 혹은 평가자 수는 모두 영화가 종영된 후에 수집된 자료이므로 영화 종영 전 시점에서 총 관객수를 예측하는 데 사용되는 것은 현실적으로 불가능하다.

Jeon과 Son (2016)은 한국 박스오피스 시장에서 2012년부터 2015년까지 4년 동안 개봉된 영화 276편을 대상으로 영화 예측 변수로써 OWOM 변수의 효과에 관한 연구를 하였다. 영화 속성변수 뿐만 아니라 네이버와 다음 포털의 평점 및 평가자 수, 네이버 포털 블로그 수, 네이버 포털 뉴스 수와 같은 OWOM 변수들과 극장 총 관객수와의 연관성분석 결과 포털 평가자 수, 뉴스 수, 블로그 수와 같은 OWOM 변수들이 관객수에 유의한 영향을 주는 변수로 나타났다. 후속 연구로서 본 연구에서는 이들 예측변수와 총 관객수를 5개의 범주로 나눈 범주형 목표변수를 사용하여 총 관객수를 예측하였다. 예측 시점을 개봉 전, 개봉 일, 개봉 1주 후, 개봉 2주 후의 시점으로 구분하여 각 시점에 수집 가능한 자료만 사용하여 영화 흥행을 예측하였다. 투자, 제작, 배급, 상영의 4단계로 이루어진 영화 관련 산업에서는 가능한 초기 시점인 영화 개봉 전 시점에 흥행을 정확히 예측할 수 있다면 향후 투자와 같은 의사 결정에 신속히 대처할 수 있을 것이다. 영화 흥행 예측을 위해 극장 매출 수익 즉, 박스오피스(box office) 대신 극장 관객수를 사용하였는데 그 이유는 두 변수 간의 상관관계가 약 0.99로 매우 높게 나타났고 매출 수익 보다는 관객수가 영화 흥행을 평가하는 데 있어서 보다 직관적이고 익숙한 지표이기 때문이다. 예측모형으로는 데이터마이닝에서 분류의 목적으로 많이 사용되는 의사결정나무, MLP 신경망모형, 다항로짓모형, support vector machine(SVM)을 사용하였으며 10-중 교차 검증에 의해 예측의 정확도를 평가한 결과 신경망모형이 다른 모형에 비해 우수하였으며 특히 개봉 전 예측의 정확도는 92.39%이었으며 개봉 일, 개봉 1주 후, 개봉 2주 후 정확도는 각각 93.48%, 97.46%, 97.83%이었다.

2. 데이터 설명

본 연구에서 사용한 영화 자료는 한국에서 2012년 1월 1일 이후 개봉하여 2015년 12월 31일 이내에 상영종료가 된 영화들 중 총 관객수가 50만 명 이상인 276개 영화이다. 총 관객수가 50만 명 미만인 영화들의 경우, 독립자본영화 혹은 상영기간이 짧은 영화들이 많아 OWOM 효과를 나타내는 변수들의 자료수집에 어려움이 있어서 분석에서 제외시켰다.

Table 2.1은 자료 분석에 사용한 변수들을 정의한 표이며 주요 변수들에 대한 기술통계분석 결과는 Jeon과 Son (2016)을 참조한다. 영화 속성을 나타내는 변수들에 대한 자료 수집은 영화진흥위원회의 통합전산망 KOBIS 사이트(www.kobis.or.kr)를 활용하였다. 총 관객수(Audience)는 최종 영화의 흥행 지표를 나타내는 양적변수이나 Table 2.2와 같이 5개 그룹으로 나눈 범주형 변수를 목표변수로 사용하였다. 이러한 범주화는 영화산업계에서 보통 1,000만 영화, 대박 영화, 중박 영화로 부르는 기준이 총 관객수가 각각 1,000만 이상, 500만 이상, 300만 이상인데 기인한다. 개봉 일 관객수, 개봉 1주 후까지의 누적 관객수, 개봉 2주 후까지의 누적 관객수를 나타내는 변수는 각각 Daudience, A1audience, A2audience이다. 이들 관객수 변수들은 각 해당 시점 별 영화의 흥행 지표를 나타낸다. 배급사 및 상영사에서는 개봉하기 전 영화의 기본적인 특성들을 바탕으로 예상되는 관객수만큼 개봉 일 스크린 수(Dscreen)를 확보하며 개봉 후 관객수의 변화 추이에 따라 스크린 수를 가감한다. 이를 반영한 변수가 개봉 1주 후까지의 누적 스크린 수(A1screen)와 개봉 2주 후까지의 누적 스크린 수(A2screen)이다.

국적(Nationality)은 외국 영화들 중에서 빈도수가 낮은 국적이 많았기 때문에 크게 국내 영화, 국외 영화 2가지로 구분하였다. 등급(Grade)의 경우 전체 관람가, 12세 이상 관람가, 15세 이상 관람가, 18세 이상 관람가, 즉 청소년 관람불가로 구분하였다. 영화의 개봉 시기를 나타내기 위해 개봉 일의 월 단위를 기준으로 1월부터 12월까지의 개봉 월(Month)을 구하였다. 또한 개봉 계절(Season)은 3월에서 5월을 봄, 6월에서 8월을 여름, 9월에서 11월을 가을, 12월에서 2월을 겨울과 같이 4개 계절로 구분하였다.

Table 2.1. Definition of variables

Variable type	Variable name	Variable description
Properties of the film	Daudience	Number of audiences on release day
	A1audience	Number of audiences after release 1 week
	A2audience	Number of audiences after release 2 weeks
	Dscreen	Number of screens on release day
	A1screen	Number of screens after release 1 week
	A2screen	Number of screens after release 2 weeks
	Nationality	Nationality(Domestic films, Foreign films)
	Grade	Film rating(General, 12+, 15+, 18+)
	Month	Release month(January-December)
	Season	Release season(Spring, Summer, Autumn, Winter)
	Dirscore	Director score
	Actscore	Film star score
	Distscore	Distributor score
Online word-of-mouth	Nbscore	Naver portal rating before release
	Nbnum	Number of Naver portal raters before release
	Nascore	Naver portal rating after release
	Nanum	Number of Naver portal raters after release
	Dbscore	Daum portal rating before release
	Dbnum	Number of Daum portal raters before release
	Dascore	Daum portal rating after release
	Danum	Number of Daum portal raters after release
	Bblog	Number of blogs before release
	A1blog	Number of blogs after release 1 week
	A2blog	Number of blogs after release 2 weeks
	Ablog	Number of blogs after release
	Bnews	Number of news before release
	A1news	Number of news after release 1 week
	A2news	Number of news after release 2 weeks
	Anews	Number of news after release

Table 2.2. Categorical target variable: audience

Category	1	2	3	4	5
Total number of audiences(unit: million)	> 10	5–10	3–5	1–3	< 1

영화의 제작 과정에서 흥행에 영향을 줄 수 있는 요인으로 감독과 배우가 있다. 감독이 최근 3년 간 제작했던 영화들에 대한 관객수의 평균 값을 감독 점수(Dirscore)로 정의하였다. 배우 점수(Actscore)는 주연 배우 2명이 최근 3년 간 출연했던 영화들에 대한 관객수의 평균 값을 주었다. 영화의 배급 과정에서 흥행에 영향을 줄 수 있는 요인으로 배급사가 있다. 배급사가 최근 3년 간 배급했던 영화들에 대한 관객수의 평균 값을 배급사 점수(Distscore)로 나타내었다.

OWOM 변수로서 개봉 전 평점(Nbscore, Dbscore), 개봉 전 평가자 수(Nbnum, Dbnum), 개봉 후 평점(Nascore, Dascore), 개봉 후 평가자 수(Nanum, Danum)는 네이버 및 다음 포털 영화 사이트에서 각각 수집하였다. 포털 영화 평점 변수에 해당하는 Nbscore, Nascore, Dbscore, Dascore 등은 영화에 대한 긍정 및 부정의 평가를 반영하고 있다. 포털 네티즌 즉, 포털 영화 평가자는 각 영화에 대해 $0, 1, 2, \ldots, 10$점으로 평가할 수 있다. 각 평가자가 부여한 평점의 총합을 총 평가자 수로 나눈 값이 포털

영화 평점이다. 이렇게 계산된 포털 영화 평점은 10점에 가까울수록 영화에 대한 긍정적 평가로, 0점에 가까울수록 부정적 평가로 간주할 수 있다. 본 연구에서 사용되는 모든 영화 자료는 상영 종료된 영화에 대해서 사후에 자료 수집을 하였다. 개봉 1주 후 및 개봉 2주 후 포털 영화 평점 및 평가자 수를 구하기 위해서는 개별 영화에 대해 개봉 1주 후 및 개봉 2주 후 시기까지에 해당하는 영화 섹션 페이지를 넘겨가며 평점 및 평가자 수를 일일이 더하거나 혹은 평균을 내야 하는 절차를 따라야 하는데 이러한 종류의 데이터 수집은 현실적인 어려움이 있어 제외하였다.

개인 홈페이지와 같은 블로그는 인터넷을 통해 대형 미디어에 못하지 않은 힘을 발휘할 수 있는 1인 미디어의 성격을 갖는다. 뉴스는 공신력 있는 언론사 등에 의해 대중들에게 전달되어진다. 블로그 수, 뉴스 수의 수집은 국내 포털 점유율이 가장 높고 회원을 가장 많이 보유하고 있는 네이버의 검색 엔진을 사용하였다. 개봉 전 블로그 수(Bblog) 및 개봉 전 뉴스 수(Bnews)는 영화 개봉 1달 전으로 부터 개봉 전날까지 영화제목의 검색 건수를 사용하였고 개봉 후 블로그 수(Ablog) 및 개봉 후 뉴스 수(Anews)는 영화의 개봉일로부터 3달 후까지 영화제목의 검색 건수를 사용하였다. 이는 대부분 영화의 상영기간이 1달 내외이고 블로그나 뉴스의 게시물들이 상영종료 후 뒤늦게 올라오는 경우를 감안하였다. 개봉 1주 후, 개봉 2주 후 블로그 수를 나타내는 변수인 A1blog, A2blog와 개봉 1주 후, 개봉 2주 후 뉴스 수를 나타내는 변수인 A1news, A2news는 개봉 일로부터 각각 1주 후, 2주 후까지의 누적 건수를 사용하였다.

3. 예측 모형

본 연구에서는 총 관객수의 분류에 의한 영화 흥행 예측을 위해 데이터마이닝에서의 주요 분류 기법에 해당하는 의사결정나무, 신경망모형, 다항로짓모형, 그리고 support vector machine(SVM) 기법을 사용하였다. 의사결정나무, 신경망모형, 다항로짓모형에 의한 예측을 위해 SAS Institute Inc. (2012)의 SAS Enterprise Miner 12.1을 사용하였고, SVM의 경우 SAS는 기본적으로 이항 목표변수 만을 지원하기 때문에 R Project Package인 'e1071'의 SVM 함수를 사용하여 분석하였다.

3.1. 의사결정나무

의사결정나무의 분류나무는 의사결정규칙을 나무 구조의 형태로 도표화하여 목표변수의 범주를 분류하는 기법이다. 나무 구조로 표현이 되기 때문에 분류 결과의 해석이 용이하며 주요한 예측변수에 관한 정보를 얻을 수 있는 장점이 있다. 분류나무는 목표변수의 각 범주에 속하는 빈도수를 기초로 하여 마디의 분리가 일어난다. 이 때의 분리 기준으로 CHAID 알고리즘을 통한 Pearson의 카이제곱 통계량을 사용하였다.

3.2. 신경망모형

신경망모형은 매우 유연한 비선형모형으로서 예측변수들을 결합하여 각 은닉마디에 전달하고 은닉마디들의 결합을 출력마디에 전달함으로서 목표변수의 범주를 분류하는 분류모형이다. multilayer perceptron(MLP) 신경망모형의 구조는 예측변수들로 구성되는 입력층, 은닉마디들로 구성되는 은닉층, 그리고 목표변수의 범주들로 구성되는 출력층으로 이루어진다. SAS Enterprise Miner 12.1에 의한 신경망모형의 분석에서는 은닉층의 수가 오직 1개로 고정되어있고 은닉마디의 수는 1부터 64까지 선택이 가능하다.

본 연구에서 사용된 신경망모형은 다음과 같이 구성된다. $X_1, X_2, \ldots, X_p$를 예측변수라 놓자. 그러면 $i(i = 1, 2, \ldots, 276)$번째 영화에 대한 $j(j = 1, 2, \ldots, J)$번째 은닉마디 H_{ij}는 식 (3.1)과 같은 쌍곡탄젠

트(hyperbolic tangent) 함수 tanh(·)에 의해서 계산된다.

$$H_{ij} = \tanh(\zeta_{ij}) = \frac{\exp(\zeta_{ij}) - \exp(-\zeta_{ij})}{\exp(\zeta_{ij}) + \exp(-\zeta_{ij})}, \tag{3.1}$$

여기서

$$\zeta_{ij} = u_{0j} + u_{1j}X_{i1} + u_{2j}X_{i2} + \cdots + u_{pj}X_{ip}.$$

최종적으로 i번째 영화가 범주 $k(k = 1, 2, 3, 4, 5)$일 확률 $P(Y_i = k)$을 식 (3.2)와 같이 계산하여 가장 높은 확률 값을 주는 범주로서 각 영화의 총 관객수 범주 k를 결정한다.

$$P(Y_i = k) = \frac{\exp(\eta_{ik})}{\sum_{k=1}^{5} \exp(\eta_{ik})}, \tag{3.2}$$

여기서

$$\eta_{ik} = v_{0k} + v_{1k}H_{i1} + v_{2k}H_{i2} + \cdots + v_{Jk}H_{iJ}.$$

3.3. 다항로짓모형

다항로짓모형은 범주형 목표변수가 갖는 범주가 3개 이상일 때 목표변수의 분류에 사용하는 로지스틱모형이다. 다항로짓모형에 의해 i번째 영화가 범주 $k(k = 1, 2, 3, 4, 5)$일 확률 $P(Y_i = k)$을 식 (3.3)과 같이 계산하여 가장 높은 확률값을 주는 범주로 각 영화의 총 관객수 범주를 결정한다.

$$P(Y_i = k) = \frac{\exp(\eta_{ik})}{\sum_{k=1}^{5} \exp(\eta_{ik})}, \tag{3.3}$$

여기서

$$\eta_{ik} = \begin{cases} \beta_{0k} + \beta_{1k}X_{i1} + \beta_{2k}X_{i2} + \cdots + \beta_{pk}X_{ip}, & k = 1, 2, 3, 4, \\ 0, & k = 5. \end{cases}$$

3.4. Support vector machine

Support vector machine(SVM)은 두 범주 사이의 거리(margin)를 최대로 해주는 초평면(hyperplane)을 분류 함수로 사용하여 목표변수 값을 분류하는 기계학습법으로 신경망모형과 함께 많은 응용문제에서 우수한 성능을 보여주는 분류기법이다. SVM은 기본적으로 이항 분류 문제를 푸는 알고리즘이며 다 범주의 분류는 이항 분류 규칙을 따른다. 즉, 모든 다 범주에 대해서 1 대 1 이항 분류를 대응시킨 후 투표에 의해 각 관측치가 속할 최종 범주를 찾는다.

4. 박스오피스 예측

이제 앞서 설명했던 의사결정나무, MLP 신경망모형, 다항로짓모형, 그리고 SVM을 사용하여 총 관객수 예측을 수행해보기로 한다.

총 관객수 예측은 개봉 전(Before), 개봉 일(Release), 개봉 1주 후(After 1 week), 그리고 개봉 2주 후(After 2 weeks) 시점의 총 4가지 시점으로 나누어 예측하였다. Table 4.1은 이러한 예측 시점에 따

Table 4.1. The variables usable in accordance with the time

Time	Before			Release			After 1 week			After 2 weeks		
	Input	Tree	Logit	Input	Tree	Logit	Input	Tree	Logit	Input	Tree	Logit
Nationality	O	X	O	O	X	X	O	X	O	O	X	O
Grade	O	X	X	O	X	X	O	X	O	O	X	O
Month	O	X	X	O	X	X	O	X	O	O	X	O
Season	O	X	X	O	X	X	O	X	X	O	X	X
DirScore	O	O	O	O	O	X	O	O	X	O	O	O
ActScore	O	X	X	O	X	X	O	X	X	O	X	O
DistScore	O	X	X	O	X	X	O	X	O	O	X	O
Nbscore	O	O	O	O	X	X	O	X	O	O	X	O
Nbnum	O	X	X	O	X	X	O	X	O	O	X	O
Dbscore	O	X	X	O	X	X	O	X	O	O	X	O
Dbnum	O	O	O	O	X	O	O	X	O	O	X	O
Bblog	O	O	O	O	O	O	X	X	X	X	X	X
Bnews	O	X	O	O	X	X	X	X	X	X	X	X
Daudience	X	X	X	O	O	O	X	X	X	X	X	X
Dscreen	X	X	X	O	X	X	X	X	X	X	X	X
A1audience	X	X	X	X	X	X	O	O	O	X	X	X
A1screen	X	X	X	X	X	X	O	X	O	X	X	X
A1blog	X	X	X	X	X	X	O	O	O	X	X	X
A1news	X	X	X	X	X	X	O	X	O	X	X	X
A2audience	X	X	X	X	X	X	X	X	X	O	O	O
A2screen	X	X	X	X	X	X	X	X	X	O	X	O
A2blog	X	X	X	X	X	X	X	X	X	O	O	O
A2news	X	X	X	X	X	X	X	X	X	O	X	O
Nascore	X	X	X	X	X	X	X	X	X	X	X	X
Nanum	X	X	X	X	X	X	X	X	X	X	X	X
Dascore	X	X	X	X	X	X	X	X	X	X	X	X
Danum	X	X	X	X	X	X	X	X	X	X	X	X
Ablog	X	X	X	X	X	X	X	X	X	X	X	X
Anews	X	X	X	X	X	X	X	X	X	X	X	X

라 사용 가능한 변수들을 보여준다. 연속형 변수들 중에서 포털 평점을 제외한 변수들은 이상치가 존재하며 비대칭 분포의 형태를 보이므로 로그변환을 한 후 분석을 실시하였다. Input 열의 표시 O는 각 시점별로 사용가능한 예측변수들을 의미하며 표시 X는 사용되지 않은 예측변수들이다. 관객수와 스크린 수를 제외한 영화의 속성 변수들은 모든 시점에서 공통으로 사용되었다. 개봉 일, 개봉 1주 후 및 개봉 2주 후 포털 영화 평점 및 평가자 수를 구하는 현실적인 어려움은 이미 2장에서 언급하였다. 따라서 포털 평점 및 평가자 수는 개봉 전 자료만 이용 가능하므로 이들 변수들은 4개의 시점 모두에서 공통으로 사용되었다. 관객수, 스크린 수, 블로그 수, 그리고 뉴스 수는 시점 별로 다르게 사용된다.

의사결정나무를 사용할 때 해당 마디가 더 이상 분리가 일어나지 않고 끝 부분의 마디가 되도록 하는 정지규칙으로서 카이제곱 검정의 유의수준은 0.2, 최소 관측치의 수는 10, 최대 가지의 수는 2, 그리고 최대 나무의 깊이는 10으로 지정하였다. 신경망모형의 분석에서는 은닉마디의 수를 1부터 64까지의 시뮬레이션 결과로부터 결정하였다. 다항로짓모형에 의한 예측에서는 단계별(stepwise) 선택에 의해 선택된 예측변수를 사용하였는데 이때 변수 추가 기준의 유의수준을 0.2, 변수 제거 기준의 유의수준을

Table 4.2. Goodness of fit

Time	Tree	NN	Logit	SVM
Before	48.75%	99.64%	57.61%	60.14%
Release	65.58%	100.00%	71.01%	68.48%
After 1 week	77.54%	100.00%	84.06%	79.71%
After 2 weeks	89.13%	100.00%	90.58%	84.42%

0.1로 하여 분석하였다. SVM에 의한 분류에서 커널 함수는 선형함수와 방사형 기준함수(radial basis function)를 사용하여 두 함수 중에서 보다 높은 정확도를 보여주는 결과를 제시하였다.

신경망모형과 SVM에 의한 분류시, 의사결정나무 혹은 다항로짓모형에 의해 선택된 예측변수들만 사용하여 분류하였을 때의 정확도는 각 예측 시점에서 사용가능한 예측변수 모두를 사용했을 때보다 일관성 있게 정확도가 더 낮았다. 따라서 신경망모형과 SVM은 각 예측시점에서 사용가능한 예측변수를 모두 사용하였다.

다음은 총 관객수 예측에 사용된 각 모형의 성능 비교를 위하여 전체 데이터에 대한 모형의 적합도, 10-중 교차 검증을 통한 정확도, 그리고 2015년 1년에 대한 예측의 정확도를 제시하였다.

4.1. 전체 데이터에 대한 적합도

Table 4.2는 전체 데이터를 의사결정나무(Tree), 신경망모형(NN), 다항로짓모형(Logit), SVM에 의해 적합시켰을 때 각 모형의 적합도를 목표변수의 정확한 분류의 비율인 정확도로 나타내었다. Figure 4.1(a)는 신경망모형에서 은닉마디의 수의 변화에 따른 정확도를 나타낸 그림이다. 은닉마디의 수가 증가할 수록 정확도도 증가 추세이며 은닉마디의 수가 15에서 정확도는 거의 100%에 이른다. Table 4.2에서 신경망모형은 은닉마디의 수가 15인 경우의 정확도를 제시하였다. 모든 모형들은 예측 시점이 늦어질 수록 적합도는 더 높아진다. 특히 신경망모형은 다른 모형에 비해 적합도가 매우 높으며 거의 100%의 적합도를 보인다.

의사결정나무 혹은 다항로짓모형은 목표변수에 유의한 예측변수들을 선별해주는 기능이 있다. Table 4.1의 Tree 열과 Reg 열의 표시 O은 의사결정나무와 다항로짓모형의 각각의 선택 기준에 의해 선택된 유의한 변수들이며 표시 X는 선택되지 못한 변수들이다. 관객수 및 블로그 수는 모든 시점에서 두 모형에 의해 선택된 가장 중요한 변수라고 볼 수 있다. 그 외에도 의사결정나무는 모든 시점에서 감독 점수를, 다항로짓모형은 모든 시점에서 개봉 전 다음 포털의 평가자 수를 선택하였다.

4.2. 10-중 교차 검증에 의한 예측

Table 4.3은 10-중 교차 검증에 의한 예측을 하였을 때 의사결정나무(Tree), 신경망모형(NN), 다항로짓모형(Logit), SVM의 정확도를 나타낸다. Figure 4.1(b)는 신경망모형에서 은닉마디의 수의 변화에 따른 정확도를 나타낸 그림이다. 은닉마디의 수가 증가할 수록 정확도는 증가 및 감소가 반복되지만 전체적으로는 증가 추세이며 은닉마디의 수가 25개 이상이 되면 정확도에 크게 차이가 없는 것을 볼 수 있다. Table 4.3에서 신경망모형은 은닉마디의 수가 15인 경우의 정확도를 제시하였다. Table 4.2의 적합도와 비교하면 과적합의 징후로서 의사결정나무, 신경망모형, 그리고 다항로짓모형의 정확도는 더 낮아졌다. 그러나 SVM은 개봉 전 시점을 제외하면 오히려 10-중 교차 검증에서 정확도가 더 높아졌다. 모든 모형들은 예측 시점이 늦어질 수록 적합도는 더 높아지며 신경망모형, SVM, 다항로짓모형, 의사결정나무 순으로 정확도가 높다. 특히 신경망모형의 개봉 전 정확도 92.39%는 다른 방법에 비해 월등하게

Table 4.3. Accuracy: 10-fold cross validation

Time	Tree	NN	Logit	SVM
Before	48.19%	92.39%	56.16%	57.61%
Release	63.41%	93.48%	69.57%	71.38%
After 1 week	75.00%	97.46%	80.80%	84.42%
After 2 weeks	85.51%	97.83%	88.04%	92.03%

Table 4.4. Accuracy: prediction of 2015

Time	Use of final OWOM estimates	Tree	NN	Logit	SVM
Before	None	46.43%	89.29%	78.57%	83.93%
	Reg	50.00%	85.71%	83.93%	82.14%
	NN	50.00%	89.29%	85.72%	85.71%
Release	None	66.07%	98.21%	85.71%	92.86%
	Reg	66.07%	100.00%	96.43%	92.86%
	NN	71.43%	94.64%	58.93%	94.64%
After 1 week	None	76.79%	96.43%	94.64%	96.43%
	Reg	73.21%	98.21%	94.64%	94.64%
	NN	73.21%	94.64%	82.14%	98.21%
After 2 weeks	None	85.71%	100.00%	91.07%	98.21%
	Reg	85.71%	98.21%	87.50%	100.00%
	NN	85.71%	98.21%	92.86%	100.00%

더 높은 정확도이다.

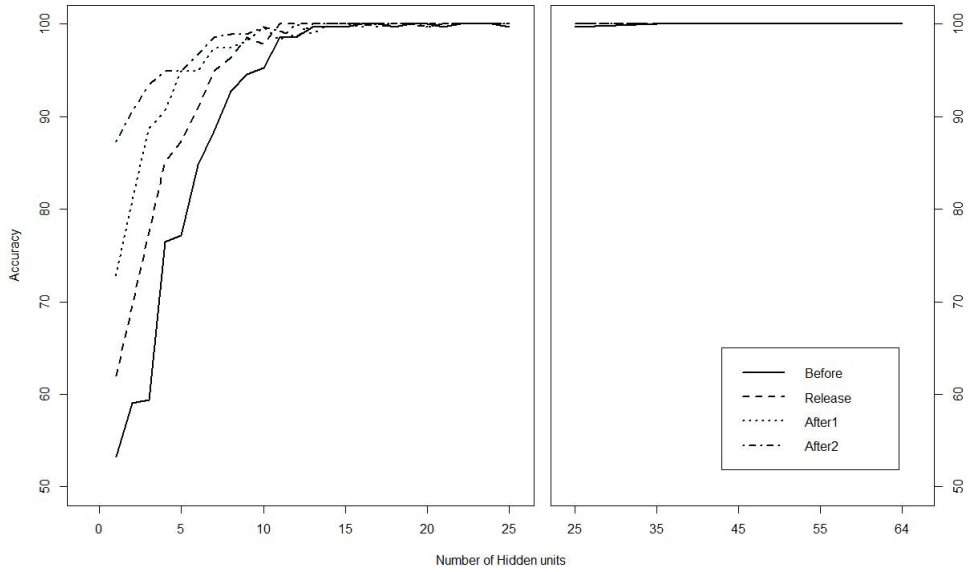
4.3. 추정된 OWOM 변수를 사용한 2015년 박스오피스 예측

Jeon과 Son (2016)에서 총 관객수에 영향력 있는 변수로 밝혀졌던 OWOM 변수들인 네이버 및 다음 포털의 평점(Nascore, Dascore) 및 평가자 수(Nanum, Danum)는 영화 상영 종료 후에 관측되며, 블로그 수(Ablog) 및 뉴스 수(Anews)도 개봉 후 3달 후, 즉 거의 영화 상영 종료 후에 관측되므로 4개의 예측 시점에서는 사용할 수 없다. 그러나 이러한 최종 OWOM 변수들은 총 관객수의 예측변수로 매우 중요한 변수이므로 추정하여 예측에 사용해보기로 한다.

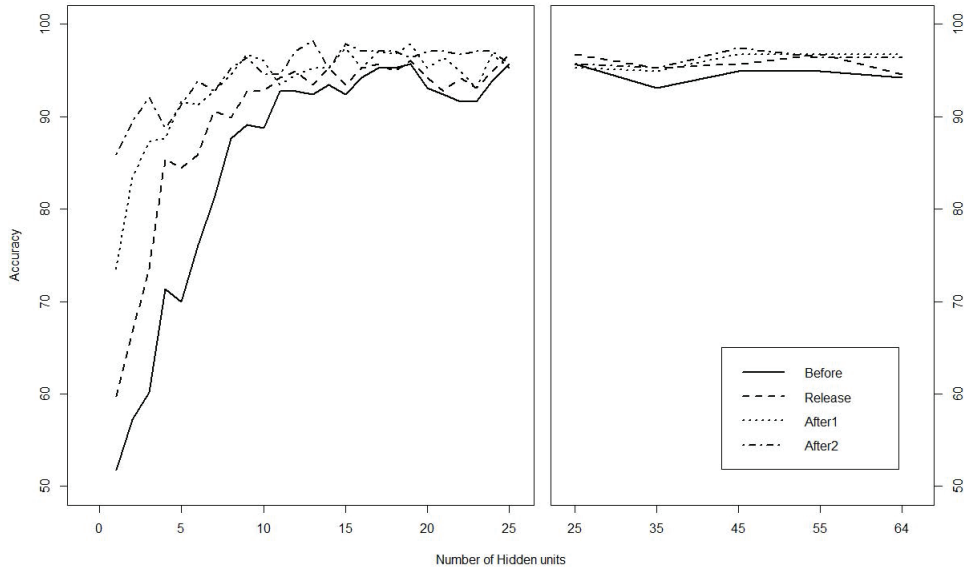
미래인 2015년의 총 관객수 예측을 위하여 2012년부터 2014년까지 3년 동안의 과거 자료를 훈련용 자료(trainig data)로 사용하고 2015년 1년 자료를 예측의 검증용 자료(test data)에 사용한다. 훈련용 자료의 각 시점에서 사용 가능한 변수들을 예측변수로 하고 최종 OWOM 변수들을 목표변수로 하는 다중회귀모형과 MLP 신경망모형를 추정한다. 추정된 모형으로 최종 OWOM 변수를 추정하여 예측변수로 추가하였다.

회귀모형 추정의 경우 단계적 변수선택법에 의해 변수 선택을 하였고 신경망모형에 의한 추정의 경우 모형의 적합도를 나타내는 Akaike Information Criterion(AIC)와 모형의 정밀도를 나타내는 RMSE를 최소로 하는 은닉마디 수가 1부터 5사이로 적절하게 나타나 1부터 5사이의 최적 은닉마디 수를 선택하였다.

Table 4.4는 2015년 총 관객수를 예측하였을 때 의사결정나무(Tree), 신경망모형(NN), 다항로짓모형(Logit), SVM의 정확도를 나타낸다. 이때 최종 OWOM 변수의 추정치를 추가하지 않았을 때(None), 회귀모형에 의해 추정된 최종 OWOM 변수의 추정치를 추가하였을 때(Reg), 그리고 신경망모형에 의해 추정된 최종 OWOM 변수의 추정치를 추가하였을 때(NN)의 3가지 정확도 결과를 보



(a) Using all data



(b) Using 10-fold cross validation

Figure 4.1. Accuracy plot of neural network.

여준다.

최종 OWOM 변수의 추정치를 추가하지 않았을 때(None)는 신경망모형, SVM, 다항로짓모형, 의사결정나무의 순으로 예측의 정확도가 높으며 Table 4.3의 10-중 교차 검증과 비교하였을 때 신경망모형은 크게 차이가 없으나 다항로짓모형 및 SVM의 정확도는 매우 높아졌다. 개봉 전 시점에 다항로

짓모형에 의한 예측결과의 예를 들어보면 Table 4.2, Table 4.3, 그리고 Table 4.4에서 정확도는 각각 57.61%, 56.16%, 78.57% 이다. Table 4.2의 결과는 2012년도부터 2015년도까지의 전체 데이터인 총 영화 276편을 모형훈련 및 적합에 사용하여 얻어진 결과이며, Table 4.3의 결과는 2012년부터 2015년까지의 총 영화 276편중에서 랜덤하게 248편을 모형훈련에 사용하였고 28편을 예측의 정확도 판정에 사용한 결과이다. Table 4.4의 결과는 2012년부터 2014년까지 3년 동안의 영화 220편을 모형훈련에 사용하였고 2015년의 영화 56편을 예측의 정확도 판정에 사용하여 얻어진 결과이다. 특히 이 경우의 데이터는 시간의 연속성을 보존하는 특징을 갖는다. 따라서 Table 4.2와 Table 4.3의 표본은 비슷한 구성을 가지므로 Table 4.2는 Table 4.3에 비해 일반적인 특성대로 과적합 징후를 보이지만 Table 4.4에 대해서 그렇지 않은 것은 영화 데이터의 연도별 특성 차이 즉, 표본 구성의 차이로 인한 결과라고 유추한다.

Table 4.4의 16개 정확도 결과 중에서 최종 OWOM 변수의 추정치를 추가하지 않았을 때(None)에 비해서 최종 OWOM 변수의 추정치를 추가하였을 때 정확도가 더 높아진 결과는 11개 결과이며 최종 OWOM 변수의 추정치를 추가하지 않았을 때(None) 오히려 정확도가 더 높은 경우는 2개 결과이다. 신경망모형에 의해 추정된 최종 OWOM 변수의 추정치를 추가하였을 때(NN) SVM의 정확도는 모두 더 높아졌다. 특히 개봉 일에 회귀모형에 의해 추정된 최종 OWOM 변수의 추정치를 추가하였을 때 신경망모형에 의한 예측과 개봉 2주 후 최종 OWOM 변수의 추정치를 추가하였을 때 SVM에 의한 예측은 100%의 정확도를 보였다.

가장 예측력이 뛰어난 신경망모형에 의한 예측결과를 보면 개봉 전 시점은 최종 OWOM 변수의 추정치를 추가하지 않았을 때(None)에 비해 OWOM 변수의 추정치를 추가하였을 때 예측력의 향상이 없으며, 개봉 일 혹은 개봉 1주 후 시점은 회귀모형에 의한 최종 OWOM 변수 추정치를 추가하였을 때가 최종 OWOM 변수의 추정치를 추가하지 않았을 때(None)에 비해 예측력의 향상이 있으며, 개봉 2주 후 시점의 경우는 오히려 최종 OWOM 변수의 추정치를 추가하지 않았을 때(None)가 예측력이 더 뛰어나다. 따라서 Table 4.2, Table 4.3, 그리고 Table 4.4를 종합하여 판단해보면 MLP 신경망모형을 예측모형으로 사용하고, 개봉 일 혹은 개봉 1주 후 시점의 경우는 회귀모형에 의해 OWOM 변수를 추정하여 입력변수로 사용하면 예측력을 향상시킬 수 있을 것으로 기대한다.

5. 결론

본 연구에서는 국적, 등급, 개봉 월, 개봉 계절, 감독, 배우, 배급사, 관객수, 스크린 수와 같은 영화의 내재적인 속성을 나타내는 변수와 네이버 및 다음 포털의 평점과 평가자 수, 블로그 수, 뉴스 수와 같은 OWOM 변수들을 활용하여 2012년부터 2015년까지의 관객수 50만 이상인 국내 영화 276편을 대상으로 영화 흥행 척도인 총 관객수의 예측을 하였다. 예측은 개봉 전, 개봉 일, 개봉 1주 후, 개봉 2주 후의 4가지 시점에서 예측하였고 각 시점에서 관측 가능한 변수들만을 예측변수로 사용하였다.

예측 방법으로는 데이터마이닝의 주요 분류 기법인 의사결정나무, MLP 신경망모형, 다항로짓모형, 그리고 SVM을 사용하였다. 모든 자료를 대상으로 적합시켰을 때 신경망모형의 적합도는 거의 100%의 정확도를 보였다. 10-중 교차 검증에서는 신경망모형, SVM, 다항로짓모형, 의사결정나무 순으로 정확도가 높다. 특히 신경망모형의 개봉 전 정확도는 92.39%로서 다른 방법에 비해 매우 더 높았다.

Jeon과 Son (2016)에서 총 관객수에 영향력 있는 변수로 밝혀졌던 OWOM 변수들인 각 포털의 평점, 평가자 수, 블로그 수 및 뉴스 수는 거의 영화 상영 종료 후에 관측되므로 4개의 예측 시점에서는 사용할 수 없다. 그러나 이러한 최종 OWOM 변수들을 다중회귀모형 혹은 MLP 신경망모형에 의해 추정하여 예측변수로 사용하였을 때 2015년 상영 영화에 대한 총 관객수 예측의 정확도는 보다 개선되었다.

본 연구에서는 각 시점에서 얻을 수 있는 자료만을 활용하여 예측을 수행하였기 때문에 현실성(reality)

이 있으며 기존의 영화 속성 변수에 더해 OWOM 변수들을 추가적으로 사용함으로서 과거 연구 결과들에 비해 상대적으로 높은 예측력을 보였다.

References

Jeon, S. and Son, Y.S. (2016). Effect of online word-of-mouth variables as predictors of box office, *The Korean Journal of Applied Statistics*, **29**, 657–678.

Kim, T., Hong, J., and Koo, H. (2013). Forecasting box-office revenue by considering social network services in the Korean market, *Journal Teknologi (Social Sciences)*, **64**, 97–101.

Kim, Y.H. and Hong, J.H. (2011). A study for the development of motion picture box-office prediction model, *Communications for Statistical Applications and Methods*, **18**, 859–869.

Korean Film Council (2015). 2015 Korean film consumer survey, *Korean Film*.

Korean Film Council (2016). 2015 Korean film industry settlement, *Korean Film*, **71**.

SAS Institute Inc (2012). *Getting started with SAS Enterprise Miner 12.1*, SAS Institute Inc., Cary.

Sharda, R. and Delen, D. (2006). Predicting box-office success of motion pictures with neural networks, *Expert Systems with Applications*, **30**, 243–254.

Song, J. and Han, S. (2013). Predicting gross box office revenue for domestic films, *Communications for Statistical Applications and Methods*, **20**, 301–309.

Yim, J. and Hwang, B. (2014). Predicting movie success based on machine learning using twitter, *KIPS Transactions on Software and Data Engineering*, **3**, 263–270.

Zhang, L., Luo, J., and Yang, S. (2009). Forecasting box office revenue of movies with BP neural network, *Expert Systems with Applications*, **36**, 6580–6587.

데이터마이닝을 이용한 박스오피스 예측

전성현[a] · 손영숙[a,1]

[a]전남대학교 통계학과

(2016년 7월 22일 접수, 2016년 9월 17일 수정, 2016년 10월 20일 채택)

요 약

본 연구는 영화 흥행의 척도로서 총 관객수의 예측을 다루었다. 의사결정나무, MLP 신경망모형, 다항로짓모형, support vector machine과 같은 데이터마이닝 분류 기법들을 사용하여 개봉 전, 개봉 일, 개봉 1주 후, 그리고 개봉 2주 후 시점 별로 예측이 이루어진다. 국적, 등급, 개봉 월, 개봉 계절, 감독, 배우, 배급사, 관객수, 그리고 스크린 수와 같은 영화의 내재적인 속성을 나타내는 변수 뿐만 아니라 포털의 평점과 평가자 수, 블로그 수, 뉴스 수와 같은 온라인 구전 변수들이 예측변수로 사용되었다. 10-중 교차 검증에서 신경망모형의 정확도는 개봉 전 시점에서도 90% 이상의 높은 예측력을 보였다. 또한 최종 온라인 구전 변수의 추정치를 예측변수로 추가함으로서 예측의 정확도가 더 높아짐을 볼 수 있다.

주요용어: 다항로짓모형, 데이터마이닝, 영화 흥행 예측, 온라인 구전, 의사결정나무, 10-중 교차 검증, MLP 신경망모형, support vector machine

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2011-0022864).

[1]교신저자: (61186) 광주광역시 북구 용봉동 300번지, 전남대학교 통계학과. E-mail: ysson@jnu.ac.kr