한국컴퓨터정보학회 하계학술대회 논문집 제21권 제2호 (2013. 7)

다중회귀 분석을 이용한 영화 흥행 예측

정회윤°, 양형정*
°*전남대학교 전자 컴퓨터 공학과
e-mail:firetna809@naver.com° hyungjeong@gamil.com*

Predicting Financial Success of a Movie Using Multiple Regression Analysis

Hoe-Yun Jeong°, Hyung-Jeong Yang**
°*Chonnam National University

● 요 약 ●

영화의 흥행 요소를 파악하여 영화의 흥행 여부를 예측하는 것은 영화의 수익성 부분에서 아주 중요하다. 영화 시장이 과거와는 다르게 증가함에 따라, 다양한 영화 흥행에 관한 예측 연구들이 개발되었다. 본 논문에서는 영화 흥행 요소들을 수집하고 다중회귀 분석을 통해서 유의수준을 만족하는 흥행 요소들을 선택한다. 그 후, 이러한 요소들을 예측 방법들의 입력값으로 사용하여 영화 흥행을 예측한다. 성능을 비교하기 위해 본 논문에서 제안한 방법과 현재 개발된 영화 흥행 예측 방법(다중회귀, 의사결정트리, 인공신경망)들을 정확도와 평균제곱근오차를 통해 예측 모형의 성능을 비교한다. 그 결과, 다중 회귀 분석을 통해 유의한 흥행요소들만을 고려한 예측 방법의 정확도가 모든 흥행 요소들을 고려한 예측 방법보다 평균 8.2% 향상되었고, 현재까지 개발된 영화 흥행 예측 방법보다 더 높은 예측 성능을 보여준다.

키워드: 영화 흥행 예측(Predicting Financial Success of a Movie),다중회귀(Multiple regression), 의사결정트리(Decision Tree), 인공신경망(Artificial neural network)

I. 서 론

영화 시장의 성장은 국내 시장뿐만 아니라 모든 국가에서 성장 산업으로 주목 받고 있다. 특히, 국내 영화의 경우, 과거의 위상과는 다르게 현재 급성장하고 있는 산업이다. 국내 영화 시장은 매년 꾸준히 성장하고 있으며 영화의 수요와 소비가 크게 증가하고 있다.

영화 시장이 커짐에 따라, 영화 산업의 규모와 예측 가능한 발전전망, 영화의 흥행 요소와 같은 영화에 대한 많은 연구들이 진행되었다. 그 중에 어떤 영화 요소들이 영화 흥행에 영향을 미치는가는 영화의 수익성면에서 아주 중요한 연구라고 할 수 있다. 우수한 영화가 시장에서 참패하기도 하고 전혀 의외의 작품이 흥행에 성공을 하기도 한다. 이러한 현상을 분석하기 위해 많은 연구들이 진행되었고, 영화들의 흥행 여부를 예측하였다.

본 논문에서는 다중회귀 분석을 통해서 유의하다고 선정된 영화 흥행 요소들만을 고려하여 영화 흥행 예측 방법을 제안하고 기존에 연구된 예측 방법들보다 예측 정확도가 높음을 확인한다.

먼저, 영화 흥행 요소들을 수집하고 수집된 요소들은 다중회귀 분석을 통해서 유의수준을 만족하는 흥행 요소들을 선택한다. 그 후, 유의한 흥행 요소들만을 예측에 대한 입력값으로 고려한 영화 흥행에 대한 예측을 실시한다. 제안하는 영화 예측 방법과 연구된 영화 흥행 예측 방법들의 성능을 비교하기 위하여 정확도와 평균제곱근오차(RMSE, Root Mean Square Error)를 사용하여 예측 결과의 성능을 평가한다. 그 결과, 정확도는 이전의 흥행 예측 방법들보다 평균 8.2% 향상되었다.

II. 관련 연구

[1]은 추정 모수의 사전 분포를 모호사전분포로 정의함으로써 변수들의 불확실성을 반영할 수 있고, 영화들의 이질성을 고려할 수 있는 베이지안 선택 모형을 제안하였다. 베이지안 선택 모형과 인공신경망의 비교 결과, 베이지안 선택 방법이 상업적으로 성공한 영화를 예측하는데 있어 더 우수함을 나타내었다. 하지만 각 영화를 그룹별로 분류하여 그룹간의 이질성을 고려했을 뿐 영화의 이질성을 고려하여 분류하지는 못했다. 본 논문에서는 각 영화간의 흥행요인들을 분석하고 분석된 요인들을 이용하여 다양한 예측 방법으로 실험을 진행하여 각 영화간의 이질성을 고려한다.

[2]는 할리우드의 영화가 개봉하기 전에 영화의 흥행을 예측하기 위하여 데이터 마이닝 방법(인공신경망, 의사결정나무, 지원벡터머신)을 사용하였다. 예측된 결과를 통해 9가지의 카테고리로 분류하여 예측 결과를 제시하였다. 하지만 수집된 모든 영화 흥행 요소들 중에 예측 정확도를 저해하는 요소가 있을 것이라는 점을

고려를 하지 못했다. 본 논문에서는 다중 회귀 분석을 통해서 수집된 영화 흥행 요소들 중 유의한 요소들만을 고려하여 예측 방법에 적용한다.

[3]은 영화 흥행성과 예측 모형을 예술 영화와 상업 영화 각각에 적용하여 영화 유형별 상이한 예측 요인을 비교 분석하였다. 회귀식의 회귀계수의 차이를 통해, 스크린 수, 관객 평가, 장르는 예술 영화와 상업 영화 모두에 영향을 미치는 흥행요인으로 나타냈으며, 감독 명성, 상영등급, 전문가 평가, 배급사 영향력, 영화 제작국, 개봉 시기는 예술 영화의 흥행성과에 대해서만 유의한 예측 변수로 검증되었다. 하지만, 검증된 예측 변수를 예측 방법에 적용시키지는 않았다. 본 논문에서는 검증된 유의한 예측 변수를 이용하여 다양한 예측 방법에 적용한다.

III. 본 론

본 연구에서는 흥행 영화를 예측하기 위해서 영화 흥행 요인들을 수집하였다. 수집된 흥행 요소를 다중회귀분석을 통해서 중요 요인들을 결정하고, 결정된 요인들을 통해서 다중회귀 분석(Multiple Regression Analysis), 의사결정트리(Decision Tree), 인공신경망(ANN, Artificial Neural Network)을 이용하여 예측을 실시하였다. 예측 된 결과에 대한 정확도와 평균제곱오차를 이용하여 예측 모형을 비교 분석하였다.

1. 데이터 수집

본 연구에서 사용된 데이터는 2004년 1월부터 2012년 12월까지 국내에서 상영된 영화 1,074편을 대상으로 하였다. 웹사이트와 영화진흥위원회을 통해서 얻은 영화감독, 배우, 배급사, 제작사, 개봉날짜, 첫날 스크린 수, 영화 상영시간, 장르, 영화 등급, 국적을 흥행 요소들을 직접 수집하였다.

수집된 데이터에서 흥행 요인들은 이산적인 값을 갖는 이산변수와 측정대상의 특성을 분류하기 위해 숫자를 분류하여 구분기호로 사용하기 위해 더미변수로 각 흥행요인들의 특징에 맞게 각각의 변수들을 정한다.

영화감독 : 이산 변수(제작비 보다 높은 수익률을 올린 영화의 수, 0 ~ 5)
배우 : 이산 변수(제작비 보다 높은 수익률을 올린 영화의 수, 0 ~ 5)
배급사 : 이산 변수(시장 점유율을 기준, 0~5)
제작사 : 이산 변수(시장 점유율을 기준, 0~5)
개봉일 : 이산 변수(2004에서 2012년까지 월별 관객 평균 값)
개봉 첫날 스크린 수 : 이산변수(개봉 첫날 국내에서 상용되는 횟수)
상영시간 : 더미변수(영화가 상영되는 시간,
장르 : 더미변수(액션, 사극, 범죄, 드라마, 코미디, 공포, 판타지, 어드벤처, 전쟁, 애니메이션, 미스터리, SF, 멜로, 스릴러, 가족, 다큐)
등급 : 더미변수('영상물등급위원회'에서 심사를 통해 정해지는 영화 상영 등급)
국적 : 더미변수(한국, 미국, 그 외 국가)

각 영화에 대한 흥행 성적을 종속 변수로 선정하고 표 1과 같이 나타내었다. 흥행 성적에 대한 가장 이상적인 변수는 투자 대비 수익이지만, 수익에 대한 자료 획득이 어렵기 때문에 '영화진흥위원회'에서 제공하는 영화 관객 수를 사용하여 표1과 같이 단계별로 나타내었다.

표 1. 관객수를 이용한 영화 흥행 성적

	1	2	3	4	5
관객수	~100만	200만	300만	400만	500만
	6	7	8	9	10
관객수	600만	700만	800만	900만	1000만~

2. 흥행요인 선정

영화 흥행 성적에 대한 수집된 영화 흥행 요인이 미치는 영향을 분석하기 위하여 식(1)을 통해 다중회귀 분석을 실시하였으며, 구체적인 분석 결과는 표 2에 나타내었다[4].

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i, i = 1,2,\ldots,n \quad (1)$$

다중회귀분석은 2개 이상의 독립변수가 종속변수에 미치는 영향을 분석한다. 수집된 독립변수를 단계별로 투입하는 방법을 통해서 유의수준이 0.05를 만족하지 못할 경우 탈락시켜 중요 변수를 선정한다.

표 2. 영화 흥행에 대한 회귀분석

	비표준화된 계수	표준화된 계수	표준오차	t-value	p-value
영화감독	0.010	0.034	0.008	1.799	0.072
배우	0.020	0.068	0.007	3.579	0.000
배급사	0.020	0.068	0.007	0.022	0.002
제작사	-0.040	-0.137	0.013	1.422	0.000
개봉일	0.000	0.000	0.003	0.056	0.955
스크린수	0.340	1.172	0.008	0.459	0.000
상영시간	0.010	0.034	0.004	3.170	0.001
장르	-0.040	-0.137	0.003	1.336	0.181
등급	-0.020	-0.069	0.001	0.654	0.513
국적	-0.010	-0.034	0.009	1.513	0.030

표2 는 다중회귀분석을 실시한 결과이다. 이 결과를 통해 종속변수에 대한 유의한 독립변수를 고려하기 위해서 유의수준 p-value < 0.05를 만족하는 모형을 선택하여 다중회귀 모형을 구성한다. 그 결과 '개봉 첫날 스크린 수', '국적', '제작사', '배급사', '배우', '상영시간'이 영화 흥행에 대해 유의한 영향을 미치는 변수로 나타났다.

3. 인공 신경망 영화 흥행 예측 방법

영화 흥행 예측 방법을 비교하기 위한 방법으로 인공 신경망(ANN, Artificial Neural Network)을 사용하여 예측한다. 인공신경망은 입력층(Input Layer), 출력층(Output Layer) 그리고 sigmoid 함수를 전달함수로 사용하는 은닉층(Hidden Layer)으로 이루어진 다계층 구조로 구성된다.

본 논문에서는 1개의 입력층과 1개의 은닉층, 1개의 출력층으로 구성된 3계층 퍼셉트론학습 알고리즘을 사용하였으며 학습방법은 역전파 학습방법을 이용하였다.

입력층은 표 2의 속성들 중에서 다중회귀 분석을 통해 검출된 유의한 변수 6개('개봉 첫날 스크린 수', '국적', '제작사', '배급사', '배우', '상영시간')를 이용하여 입력값으로 사용한다. 출력층은 흥행 성적에 따른 분류를 위해 1부터 10까지의 값들로 분류하여 10개의 출력값으로 나타내었다.

표 3. 은닉층의 노드 수에 따른 정확도(단위 : %)

은닉 노드수	1	2	3	4	5	6	7	8
예측 정확도	72.7	78.1	81.2	82.1	83.6	76.2	84.8	84.4
은닉 노드수	9	10	11	12	13	14	15	16
예측 정확도	85.6	87.4	87.1	89.6	87.5	88.1	87.1	82.2

표 3은 은닉층의 노드 수에 따른 정확도를 나타낸 결과이다. 은닉층의 최적의 노드를 찾기 위해 은닉층의 노드의 수를 1부터 16개까지 조정하면서 정확도가 높은 은닉층의 노드 수를 선택하였다. 그 결과, 은닉층 노드의 수가 12개일 때, 가장 높은 정확도를 나타내었다.

IV. 실 험

본 연구에서는 3개의 예측 방법의 성능을 비교하고자 한다. 1,074개의 영화와 7개의 흥행 요소를 토대로 실험을 진행하였다. 수집된 모든 흥행요소를 고려한 결과와 회귀분석을 통해 유의한 흥행요소만을 고려하여 예측한 방법으로 나뉘어 실험하였다. 예측에 사용될 방법은 다중회귀 모형, 의사 결정 트리, 신경망을 통해서 비교하였다.

표 4. 영화 흥행 예측 모형간의 예측 정확도 및 RMSE

	다중 회귀 분석	의사결정 트리	인공 신경망	
모든 흥행 요소	정확도	74.1%	44.3%	82.0%
	RMSE	1.7141	1.7435	1.6820
유의한 흥행 요소	정확도	77.8%	57.7%	89.6%
	RMSE	1.5411	1.5196	1.4074

표 4는 영화 각 흥행요소를 고려한 흥행 예측 방법들의 정확도와 RMSE를 나타낸 결과이다. 식 (2)는 평가를 위해 사용한 RMSE(Root Mean Squared Error)로서 실제로 흥행하는 영화와 시스템이 예측한 영화 흥행 정도의 차이를 비교하는 방법이다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i,j}(P_{i,j} - r_{i,j})^2} \qquad (2)$$

본 실험을 통해서 다중회귀를 사용하여 유의한 흥행 요소만을 입력값으로 사용한 인공신경망이 흥행 예측 방법이 정확도 89.6%로 가장 우수한 성능을 보여주었다. 또한, 다중회귀를 사용하여 유의한 요인만을 고려한 방법이 평균적으로 정확도를 8.2% 향상시켰다.

V. 결 론

본 연구에서는 다중회귀 분석, 인공 신경망, 의사결정 트리 모형을 이용하여 영화의 흥행을 예측하였다. 다중회귀 분석을 통해서 유의한 흥행요인들을 선택하였다. 이 흥행요인들을 이용하여 3개의 예측 방법들을 이용하여 영화 흥행을 예측하였다. 그 결과, 유의한 흥행요인들만을 고려한 영화 흥행 예측 방법에 대한 정확도가 평균 8.2% 향상되었으며, 인공신경망을 이용한 영화 흥행 예측 방법이 89.6%로 가장 성능이 우수하였다.

감사의 글

"본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 IT 연구센터 지원사업의 연구결과로 수행되었음"
(NIPA-2013-H0301-13-3005), "이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2012-047759)"

참고문헌

[1] Gyeongjae Lee, Ujin Jang, "Predicting Financial Success of a Movie Using Bayesian Choice Model", Korean Institute of Industrial Engineers, 2006

[2] Soojin Lee, Taeryong Jeon, Gyeongdong Back and Sungshin Kim, "A Movie Rating Prediction System Based on Personal Propensity Analysis", Preceedings of kIIS Fall Conference 2008 Vol. 18, No2 pp. 203 - 206, 2008

[3] So-Young Kim, Seunghee Im and Yeseul Jung, "A Comparison Study of the Determinants of Performance of Motion Pictures: Art Film vs. Commercial Film", Journal of Korea Contents Association, Vol. 10 No. 2 pp. 381-393, 2010

[4] Rackin Choi, "A Multiple Regression Analysis on

Developing the Profitability Model of Local Cultural Festival“, Korea Socety of Computer Information, Vol. 16, Issue 10, 10. 2011

[5] Deniz Demir, Olga Kapralova, Hongze Lai, “Predicting IMDB movie ratings using Google Trends”, December 15, 2012

[6] S. Kabinsingha, S. Chindasorn, C. Chantrapornchai , “A Movie Rating Approach and Application Based on Data Mining”, International Journal of Engineering and Innovative Technology (IJEIT) Vol. 2, Issue. 1, July 2012

[7] Dursun Delen and Ramesh Sharda, “Predicting the Financial Success of Hollywood, Moveis using an Information Fusion Approach“ pp. 30-37, Endüstri Mühendisliði Dergisi, December 2009, 298-300, Jan. 2012.