2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing

Multi-model or Single-model?

A study of movie box-office revenue prediction

Guijia He
Soongsil University
School of Computer Science & Engineering
Seoul, Republic of Korea
twofirst@hotmail.com

Soowon Lee*
Soongsil University
School of Computer Science & Engineering
Seoul, Republic of Korea
swlee@ssu.ac.kr

***Abstract*—Although many studies tried to predict movie revenues in the last decade, the performance and conclusions are conflictive because different data is used. Some studies report that using social data like reviews can obtain the better prediction than using only metadata of movies, but we demonstrate metadata can beat social data in some cases. In this paper, we utilize EM (Expectation Maximization) algorithm to divide movies into several groups, and then for each group we learn one model to predict movie box-office revenue separately. Experimental results show that using multiple models (Multi-model) can obtain more accurate prediction than using a single model (Single-model).**

Keywords—movie box-office; movie revenue; movie gross; movie revenue prediction; Multi-model

I. INTRODUCTION

The movie industry is of high profit for the public because of its entertainment nature, but it is of high risk for economists due to its unpredictability. The budget of a movie generally covers millions, tens of millions of dollars and sometimes even more. However, the gross revenue of a movie is not only decided by its budget. Just like Jack Valenti, president and CEO of the Motion Picture Association of America, mentioned that, '... No one can tell you how a movie is going to do in the marketplace...' [1], a big-budget film may fail to recover its cost while a small-budget film may become a black horse. Hence, an interesting question is how to forecast the movie gross revenue as early as possible.

Previous studies to predict movie box-office revenue can be mainly classified into two types, the metadata-based prediction and the social-based prediction, according to the data used. The basic idea of the metadata-based method is that the movie itself can determine its destiny, that is gross revenue. The meta information of a movie includes genres, MPAA (Motion Picture Association of America) rating, budget, directors, actors, the number of screens on the first week and so on. This information can be crawled from Online movie databases like IMDB (www.imdb.com) [2]. The other one, the social-based method, considers that the gross revenue of a movie should be affected not only by the movie itself but also by public participation including comments, opinion and so on [3]. In the previous studies, however, there exist some disagreements, particularly for metadata-based prediction. A typical one is about star power. Some of the researchers think star power is one of the most important factors for a movie to be a blockbuster while some others do not think so. Due to the inconsistent conclusion of metadata-based prediction, external information from the internet like reviews is used to help predict gross revenue of movies. Although some studies have demonstrated that social data is helpful to predict movie revenue, we think the movie itself can tell more. This is because the metadata of a movie represents its value and the value should be reflected by its revenue. For the disagreements mentioned above, we think the reason is due to the unsuitable predictive models. Most of the previous studies use only one model (Single-model) to predict movie revenues, and we do not think it is enough. In this research, we build several models (Multi-model) to resolve the problem. The basic idea is that each movie has its distinguishing features, and it is not suitable to explain all of the movies by only one model. We hence consider partitioning the movies into different groups. In this way, the model learned from a group of movies can well predict a new movie if the movie is similar to the ones in the group.

The contents of this paper are organized as follows. The next section reviews some previous studies about movie box-office gross prediction. Then the dataset and the predictive task is illustrated. Our proposed model is described in the fourth section. In the experiment section, we explain how to build the muti-Model and compare our results with the previous research. Finally, we conclude our work in the last section.

II. RELATED WORK

A. Metadata-based Prediction

An earlier research to predict box-office revenues of movies using metadata is reported by Simonoff and Sparrow [2]. They conducted two experiments, prediction prior to release and after the first weekend of release. Meanwhile, they divided their dataset into two groups according to whether the number of screens is more than 10. They found that the accuracy of predictions obviously increases after the first weekend of release for the movies opening on more than 10 screens. Although the division according to 10 can explain their dataset to some extent, it is not a standard division for other datasets. Hence, an automatic partition method is needed to apply to any dataset. In our research, we utilize EM

* corresponding author

978-1-5090-0154-5/15 $31.00 © 2015 IEEE
DOI 10.1109/CIT/IUCC/DASC/PICOM.2015.46

(Expectation Maximization) algorithm to automatically divide dataset, and then we learn several models from the partitions to predict movie gross revenues.

Vany and Walls analyzed the relation between star power and movie box-office gross [4]. They illustrated that large budgets and star presence can create the biggest of flops, and much smaller budgets and lack of star presence, in contrast, do not prevent a film from becoming a box-office hit. In Sharda and Delen study [1], the forecasting problem is converted into a classification problem based on neural networks. Their experimental results show that the major contributors to the prediction are the number of screens, high technical effects, and the high star value. The effect of star power in their report is opposite to Vany's conclusion.

B. Social-based Prediction

Recently more researchers tend to forecast movie box-office revenues by combining social data. Social data influences movie revenues mainly in two ways, volume and valence. The volume means the total number of times that a movie occurs on the internet (e.g. news). The valence stands for the polarity of the sentiment of the public, positive or negative for example. Liu [5] and Duan et al. [6] named social data as word-of-mouth (WOM) and examined how it helps explain box-office revenue of movies. Besides them, Zhang and Skiena predicted movie gross through news analysis [7]. Liu et al. built a model named ARSA using blogs to predict sales performance in movie domain [8]. Asur and Huberman used the chatter from Twitter to forecast box-office revenues for movies [3]. In contrast, Joshi [9] and Yu [10] paid their attention to analyzing movie reviews.

By text mining and sentiment analysis, above reports demonstrated that social data can help forecast movie revenues. However can metadata provide a better prediction than social data? In this article, we improve the accuracy of metadata-based prediction by building a few models, and the comparison results show that in some case metadata can predict better.

III. Data and Task

A. Dataset

In this study, we use the data provided by Joshi et. al. in their research [9]. The dataset contains 1718 movies from 2005 to 2009, and each movie includes metadata and reviews, crawled from two websites, Meta-Critic (www.metacritic.com) and The-Numbers (www.the-numbers.com). In this study, we only use metadata to demonstrate the Multi-model can provide comparable prediction performance. Furthermore, the dataset is partitioned into training, development and test sets. Concretely, the training set contains 1147 movies from 2005 to 2007, the development set contains 317 movies in 2008, and the test set contains 254 movies during 2009. The three sets are exactly the same with Joshi's configuration so that the performance comparison between theirs and ours is possible and meaningful.

B. Features

1) First weekend gross: The feature stands for the box-office gross of movies during the first weekend in U.S. Since the gross is long right-tailed, that ranges from $95 to roughly $200 million, the logarithm of the gross is used in our work.

2) Genre: The genres of a movie can reflect its type, story, and scenario. In our dataset, there are twenty-one genres in total, such as Action, Comedy, Drama, Horror, Science Fiction and so on. In some sense, Genre is a prior factor for audiences to choose a movie. Therefore, many movie studios tend to attract more potential audiences by fusing a few genres in their movies. In this work, we describe each genre as a binary number. For the genre "Drama", for example, if a movie contains the genre, the value of Drama is one and zero if not. So the genres of a movie can be seen as a vector with twenty-one dimensions, in which 1 means a certain genre is present and 0 stands for its absence.

3) MPAA: The Motion Picture Association of America (MPAA) rating in the dataset includes six types. They are G (general audiences), PG (parental guidance suggested), PG-13 (parents strongly cautioned for children under 13), R (children requires accompanying parent or adult guardian), NC-17 (no one 17 and under admitted), and U (unrated). Since MPAA rating can influence the number of audiences, we consider that it can further influence movie gross revenues. Hence, we convert MPAA rating to numeric data based on movie gross revenues. Concretely, for each type of rating, we calculate the logarithm of its average gross in the training set. This means that the MPAA values in the development and test sets are calculated based on the training sets. If two movies possess the same MPAA rating, their MPAA values are the same.

4) Star: There are three types of star information: Highest_Actors (top 50 highest-grossing actors and actresses), Oscar_Actors and Oscar_Directors. This information is gathered from a website named "boxofficemojo" and Wikipedia. For each star type in a movie, we set a value by counting the number of persons in the type. For example, in the movie "The Da Vinci Code", there are two highest actors, one Oscar actor, and one Oscar director.

5) Holiday: This information records whether a movie is released on holidays, including summer, Christmas, Memorial Day, Independence Day and Labor Day. Similarly, we use a vector with binary value to feature the holidays of the movies.

6) Screens: The number of screens on the first week has been verified as an important feature with a strong relation to movie gross revenue [1, 2]. In our dataset, the number of screens ranges from 1 to 4366, that means these numbers may come from different distributions so that they should be divided and modeled separately.

7) Others: There are some other features like running time and budget. For running time, there are many missing values as 0, we hence do not use the feature in our work. Meanwhile, we do not think it is strongly related to movie box-office gross. For budget, a general viewpoint is big-budget productions can lead high gross revenues. However, most of the budget numbers are unknown. Moreover, as described in Simonoff's report [2], the production budget only represents a part of the cost, and some other fees like advertising are ignored.

Furthermore, some directors and stars intend to get gross receipts instead of their salaries. Therefore, the budget feature is not used in our work.

C. Task

The task in our work is to predict the box-office gross of a movie during its first weekend. The first-weekend gross is a very important indicator, and in many studies it is seen as the first step to predicting the total revenues of movies. A business development executive formerly at New Line Cinema describes the first-weekend gross as "predictive of what the movie will do overall.." [2]. In addition, in order to compare our prediction results with the previous research, the task in our work focuses on the U.S. domestic gross on the first weekend, that is the same with Joshi's task [9].

IV. Proposed Model

As mentioned above, we notice that there exist some opposite conclusions in some studies using metadata. One example is about star power analyzed in [1] and [3]. We think it is caused by two reasons. One is due to different datasets are used in the two studies, and the other is that there may exist a few patterns in a dataset. This means star power is indeed useful in some certain types of movies but is not helpful for some other ones. So it is not appropriate to build only one model(Single-model) to explain the entire datasets. In order to solve the problem and to improve prediction performance, we hence attempt to build a series of models (Multi-model) instead of Single-model.

A. Partition Algorithm

Using Multi-model to predict movie gross revenues has been proposed in Simonoff's report [2]. They found it is harder to predict revenues of the movies whose the number of screens on the first week is less than 10. They hence divided their dataset into two groups, according to whether the screen number is more or less than 10, and then they built two models separately to fit the two groups of data. Our work is inspired by their report. However, a problem is why screens should be divided by 10? Is it meaningful and appropriate for other datasets? Can we make an automatic partition for different datasets? In order to solve these problems, in this work we utilize a machine learning algorithm named EM.

EM (Expectation Maximization) algorithm is an iterative method for finding the maximum likelihood of parameters in statistical models. The EM iteration includes two steps, expectation (E) step and maximization (M) step. Here we consider that the number of screens on the first week comes from various distributions with latent means and variances.

Let X be the screen numbers, Z be the number of parts to be divided, and Θ be the unknown parameters (means and variances in this work). In E step, the expectation that a screen number belongs to a partition is evaluated using the current means and variances. In M-step, the parameters are computed to maximize the expected log-likelihood found on the E step. The EM algorithm starts with an initial set of parameters and iterates until the partitions cannot be improved.

E step:

$$Q(\theta|\theta^{(t)}) = E_{Z|X,\theta^{(t)}}[log\ L(\theta; X, Z)] \quad (1)$$

M step:

$$\theta^{(t+1)} = \underset{\theta}{arg\,max}\ Q(\theta|\theta^{(t)}) \quad (2)$$

B. Modeling

In this work, our goal is to learn a Multi-model to fit different datasets. Utilizing EM algorithm, each number of screens is allocated into a group with maximum likelihood. In this way, the dataset is partitioned into a few groups based on the number of screens. Concretely, our process is implemented as follows.

- Let N be the number of groups divided by EM algorithm. Considering the sufficient training data and the simplicity principle, in our experiment we change the number of N from 1 to 10.
- For the ith partition in the training set, we learn a linear regression model, noted as Model_i, based on the AIC method with backward-stepwise in R. AIC (Akaike Information Criterion) is a measure of the relative quality of statistical models for a given dataset. Given a collection of models, AIC can estimate the quality of each model, relative to each of the other models. Our regression model starts with all candidate variables. And then deleting one variable who can improve the model the most by being deleted. The process is repeated until no further improvement is possible.
- In order to determine what number of N is the best, we implement EM algorithm and evaluate performance on the development set. For example, we use the Model_i to predict gross revenues of the movies in the ith partition of the development set. And the best number of N means the best performance can be obtained when the training and the development set is divided into N groups by EM algorithm.
- According to the best number of N, we partition the training set and the test set into N groups separately. Again, we learn N models by training set and predict gross revenues of the movies in the test set.

V. Experiments

Most of the previous studies try to improve prediction performance by combining external information like social data. In this work, however, we attempt to enhance the performance of metadata by building multiple models. Concretely, we partition the training set and test set into a few parts based on EM algorithm, and for each part in training set we learn an independent model to predict box-office revenues of the movies in the corresponding part of the test set. In order to compare with Joshi's work, we use the same evaluation metrics mentioned in their report, (1) mean absolute error (MAE) and (2) Pearson's correlation between the predicted and the actual revenues.

TABLE I. RANGE OF SCREEN NUMBERS IN EACH PARTITION OF TRAINING SET

Partition Index / Total Partitions	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
1	1-4366									
2	1-17	18-4366								
3	1-8	9-72	82-4366							
4	1-8	9-65	71-1921	1945-4366						
5	1-8	9-65	71-1912	1921-3690	3707-4366					
6	1-8	9-71	72-2040	2051-2972	2981-3241	3261-4366				
7	1-7	8-58	61-458	489-2332	2350-2822	2828-3267	3274-4366			
8	1-2	3-11	12-93	100-1829	1845-2362	2381-2725	2755-3267	3274-4366		
9	1-2	3-10	11-72	82-1265	1289-1945	1984-2310	2332-2725	2755-3267	3274-4366	
10	1-2	3-10	11-72	82-985	1000-1984	1995-2394	2411-2652	2655-3424	3434-3634	3645-4366

TABLE II. EVALUATION RESULTS ON THE DEVELOPMENT SET

N	Correlation	MAE($M)
1	0.815	11.228
2	0.922	4.240
3	0.918	3.473
4	0.831	4.835
5	0.578	6.002
6	0.907	3.546
7	0.893	3.513
8	0.895	3.546
9	0.894	3.553
10	0.846	4.256

TABLE III. PREDICTION MODELS WITH COEFFICIENTS

Factors	Coefficients		
	Model_1	Model_2	Model_3
	1<=Screen<=8	9<=Screen<=72	82<=Screen<=4366
Intercept	8.947***	15.430***	1.890e+01***
Screen	0.468***	0.032***	1.859e-03***
Log_MPAA	0.159***	-	-
Christmas	1.992*	-	9.904e-01
Memorial Day	0.925*	-	-
Highest Actors	0.771**	0.597	-
Oscar Actors	0.412	-	-
Oscar Directors	0.905	-	3.593e-01
Action	-	-0.971	-
Adventure	-	1.610**	-3.256e-01*
Comedy	-		-2.488e-01*
Family Kids	-	-2.077*	-4.697e-01**
Fantasy	-1.173*	1.458*	-
Horror	-	-1.927**	-4.028e-01**
Musical	-	-	7.407e-01**
SuspenseThriller	-	0.598	-
War	-	-	5.308e-01

*p<0.05, **p<0.01, ***p<0.001

- absence of the feature

A. Best Partition Number

As described in the last section, we separate the dataset into *N* parts and the number *N* varies from 1 to 10. Table I lists the range of screen number for each partition of the training set. In the table, the rows means how many subsets are divided by EM algorithm and the columns stand for each partition number. Notice that the range in the development and the test set may be slightly different.

In order to determine what number of *N* is the best, we evaluate performance on the development set, and the results are shown in Table II. When only one model is learned, the Pearson's correlation is very low and the MAE is rather high. In contrast, when more than one models are learned, the prediction performance on the development set increases obviously except for the case when *N* is five. The results demonstrate that Multi-model can significantly improve the prediction performance compared to Single-model.

Furthermore, we try to analyze the reason and find that when *N* is five, the number of movies in a training partition (screen number ranges from 3707 to 4366) is very small, only three percents of the entire training set. The lack of data, therefore, leads to the prediction bias.

Through the table, learning two models can get the highest correlation but the MAE is very large. By contrast, learning three models can significantly reduce errors and hold acceptable correlation. Considering the importance of both correlation and MAE, here we choose three as the best *N* number and analyze the models in the next part.

B. Model Analysis

In the last part, we evaluate the development set and choose the best partition number as three. The learned three models along with their coefficients are shown in Table III. Notice that only significant features are listed. The symbol "-" in Table III stands for the absence of a feature in a model.

Through the models and their coefficients, the number of screens on the first week seems to be the most important feature in all of the models. This is consistent with the previous reports [1, 2, 6]. Except for screens, significant features in the models are obviously different. MPAA shows strong significance in Model_1 compared with that in the others. We explain this reason as when a very small number of screens are scheduled, the movie with G or PG type would attract more audiences than the one with R type. Moreover the Model_1 also shows that for the movies (with only a few screens), it is not a bad idea to release them on holidays, particularly on Christmas and Memorial Day. Maybe because some audiences would choose an unplanned movie when their planned one is sold out on holidays. However, Genre does not show significant effects on Model_1. In contrast, it can influence other models more or less.

Compared with Model_1 and Model_2, the factors that influence Model_3 are various, including screens, holidays, star power and genres. This means a blockbuster movie needs to do the best in every field. An interesting find is that star power and holiday release do not show very important effects on Model_3. That is because, in its training partition set, most of the movies are scheduled with a large number of screens on the first week. In general, a movie with a large number of screens always contains some famous stars or is released on some holidays so as to ensure its seat occupancy rate. If only one model is learned based on the entire dataset, stars and holidays may be the two powerful variables. In our work, however, the movies with fewer screens are separated from the ones with more screens, so that each model is learned independently. Therefore, in the Model_3, star power and holidays seem not to be as important as we thought, but they are still very helpful in the other models.

C. Comparison

In this work, our goal is to demonstrate Multi-model can provide the better predictive performance than Single-model. We choose two baselines based on the results in Joshi's work [9]. The first baseline is the prediction result using only metadata, and the other is the result predicted by combining metadata and reviews. In order to prove whether our automatic partition method based on EM algorithm is comparable with the empirical division mentioned in [2], we make a third baseline based on Simonoff's method. The third baseline is to divide the dataset into two groups, according to whether the number of screens is more than 10. The compared results are shown in Table IV.

The comparison results show Multi-model can provide much better prediction performance than Single-model, both for MAE and Pearson's correlation. Although Joshi's work has proved that social data like reviews is useful for prediction, obviously only one model is learned for the entire dataset is not enough. By dividing the dataset into a few parts and learning models respectively, each model can fit a part of data very well so as to reduce the prediction errors. In addition, compared to the empirical division, our automatic division based on EM algorithm can significantly reduce MAE in spite of a slight decrease of correlation.

TABLE IV. COMPARISON WITH OTHER MODELS

Metrics	Single-model		Multi-model	
	Meta only	Meta + reviews	Empirical division	EM division (N=3)
MAE($M)	5.983	5.738	5.175	4.380
Correlation	0.722	0.819	0.876	0.867

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a Multi-model based on EM algorithm to predict the first weekend gross of movies only using metadata. Experimental results demonstrated that Multi-model can provide more accurate forecast results than Single-model. Meanwhile, by learning multiple models, metadata may obtain the better prediction even if without social data like reviews. Hence, our method can be used to increase prediction performance. In addition, compared to the empirical division, EM algorithm can provide reasonable partition and retain acceptable performance. For the future work, we plan to analyze budget information by dividing the dataset according to whether the budget is known.

ACKNOWLEDGMENT

This work was partly supported by the ICT R&D program of MSIP/IITP [13-912-03-003, Development of Event Extraction and Prediction Techniques on Social Problems by Domains] and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2013R1A2A2A04016948) and Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2014 (Grants No. C0221419).

REFERENCES

[1] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," Expert Systems with Applications, vol 30, pp.243-254, 2006.

[2] J. S. Simonoff and I. R. Sparrow, "Predicting movie grosses: winners and losers, blockbusters and sleepers," Chance, vol. 13, no. 3, pp.15–24, 2000.

[3] S. Asur, B. A. Huberman, "Predicting the Future with Social Media," 2010, http://www.arxiv.org arXiv:1003.5699v1..

[4] A. D. Vany and W. D. Walls, "Uncertainty in the movie industry: does star power reduce the terror of the box office," J. Cultural Economics, vol. 23, no. 4, pp. 285–318, 1999.

[5] Y. Liu, "Word of mouth for movies: its dynamics and impact on box office revenue," Journal of Marketing, vol. 70, pp. 74–89, July 2006.

[6] W. Duan, B. Gu, and A. B. Whinston, "The dynamics of online word-of-mouth and product sales: an empirical investigation of the movie industry," J. Retailing, vol. 84, no. 2, pp. 233-242, 2008.

[7] W. Zhang and S. Skiena, "Improving movie gross prediction through news analysis," Conf. Web Intelligence, pp. 301-304, 2009

[8] Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: A sentiment-aware model for predicting sales performance using blogs," the 30th Annual international ACM SIGIR conference on research and development in information retrieval , pp. 607–614, 2007,

[9] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: an experiment in text regression," Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 293–296, 2010.

[10] X. Yu, Y. Liu, J. X. Huang, and A. An, "Mining online reviews for predicting sales performance: a case study in the movie domain," IEEE Trans. on Knowl. and Data Eng., vol. 24, no. 4, pp. 720-734, April 2012.