ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 6, June 2015

Using Movie Scripts Assessing Box Office Performance: A Kernel Based Approach

K.R. Dabhade,

Asst. Prof,
P.E.S. College of Engineering,
Aurangabad, India.
Ms. S.S. Ponde
Asst. Prof,
Computer Science Department, D.I.E.M.S
Aurangabad, India

Abstract:--A methodology to predict performance of box office of movie at time of Green lightning, when only budget and script is available. The three level of extraction of textual features from screen writing domain knowledge (Genere & content), natural language processing techniques (bag of words) and human input (Semantic variables). This textual variable defines a distance metrics of scripts which are used as input for kernel based approach for assessing box office performance.

Keywords— Semantic variable, green lightning

I. Introduction

The script which is to be turned into movies i.e. at time of green lightning the filmmakers need to assess performance of box office which is based on movie script and then the estimated budget is allocated. This is done when factors like actor, director are unknown. Generally producers prefer five to ten past movie script which are similar to focal script so that this script can do well business at box office. But there was no methods to identify the similarity of scripts. So as to answer question like should we focus on theme of scripts the actual words Scenes and dialogue. Thus it is goal of proposed system to answer the above question and develop decision aids to studio makers so that they can decide an script which can gain good revenue on box office.

Finally, we develop methodology based on mining of text & kernel approach. The textual features which is extracted from scripts are ordered from Higher level to Lower level the higher level contents are developed by human readers (genere & content) and lower level features are (Semantic variable and actual words) extracted by machine. We then define a “distance metric” between scripts based on their textual features. We estimate the feature weights. Given the distance metric with estimated feature weights we use a kernel-based approach to assess the box office performance for new scripts. The research contribution is threefold. First is to the best of our knowledge this paper is the first that collects and analyzes actual movie scripts (about 120-pages each). Second show that the kernel approach outperforms both regression and tree-based methods in the context of assessing box office performance. Third the estimated “feature weights” provide some insights about which textual features require particular attention when identifying useful “comps” for a new script.

The Section 2 describes an overview of the script data set & how we extract textual information from script. Section 3 describes the kernel-based approach and how can we estimate the obtained feature weights. In next section we compare our method with other benchmark methods and present a hypothetical portfolios selection scenario this proposed method can gives 29 percent lower mean square error.

II. Textual Features from Movie Scripts

The data is comprised of more than 300 movies script which are available online we than record the U.S box office revenue and production budget from IMDB i.e Internet Movie Database.

2.1 Genre and Content Variable:

The textual information in movie script can summarize by the “content” variable and genere of scripts summarize by overall theme of movie.

The genre of script we considered eight genres and the content describes the variable which give detail about script like ending of story is happy or sad?

We considered eight genre based on category of movie Romance(ROM),Thriller(THR),Drama(DRA),Comedy(COM)Horror(HOR),Family(FAM),Action(ACT) and Sci-fic (SCI).

The set of 24 questions is provided about storyline of each script based on genre. These questions are simply yes & no type which have been identified by script writing experts.

2.2 Semantic Variables

This variable captures the textual information from the scripts of movie and it provides a preview that how the final movie will look like.

The script is organized into interior/exterior scenes whereas each scene is comprised characters dialogue.

The semantic variable is second layer of textual information where structure of an script is captured and final preview is provided about the script.

Here we define two level.

(i)At scene level- Here we can obtain total no of scenes in movie & the way how an character interact with co-actor.

(ii) Dialogue level- Here We can obtain the manner how character communicates all information is carried from script.

i) Number of scenes (NSCENE).
ii) Interior scenes percentage (INTPREC).
iii) Number of dialogues (NDIAG).
iv) Average of dialogues length (AVGDIAGLEN).
v)The "concentration index" of dialogues (DIAGCONC).

We use HH index i.e. Herfindahl-Hirschman index to compute the concentration index of dialogues. The value of HH index is between 0 & 1.The higher index indicates concentration of a few characters in a dialogues.

2.3 Bag-of-Words Variables

The bag of words is third layer of textual information by using natural language processing technique. The words used in scripts and frequencies of their usage are backbone of storyline.

We can extract bag of words through scripts using the following steps.

(i)We then eliminate all punctuations," stop words" and a Standard English names.

(ii) A stemming algorithm is used for reducing words to simplest form.

After eliminating stemming & stop words even though there are thousands of unique words appeared in one or more scripts. Hence we compute an "importance index" for each word.

$$I_i = \left(1 - \frac{d_i}{D}\right) \times N_i ,$$

No	Word	MAX	Featured Weight	Mean	SD	Min
1	1::ToyStory(1995): :Animation|Children's|Comedy	1	1::ToyStory(1995): :Animation|Children's|Comedy	0.0	0.0	0.0
11	11::American President	0	11::American President	3.090909090909	10.1980390271855	0.0

Figure2.1Table Summary statistic of variables

Above formula is used to measure importance index Where d_i denotes no of scripts which contains i_{th} Word. And N_i is total frequency occurrence of i_{th} word. We keep only few 100 words as important words and finally we perform LSA to further reduce dimensionality of the words document matrix. Based on singular-value decomposition (SVD) it provide us to index each script by a set of "scores". Because the singular values shows an "elbow" at the two singular-value solution we retain two latent semantic scores for each script labeled as LS1 and LS2 and use them as textual features for further analyses.

2.4 Summary and Potential Data Limitations

The summary statistics for each variable in data set is taken. All textual variables and the (log-) production budget is considered and used as predictors in a kernel-based approach which forecast box office performance.

III.A Kernel based approach to forecast box office performance

The kernel-based method utilizes a distance metric to identify the "similarity" between a new observation and each observation in the training database. The kernel- based approach is free of functional form this allow flexibility to capture complex relationship between features in textual script and box office performance. Hence we feel that kernel based approach is appropriate, So that correct relationship between textual variable of scripts & box office. Another approach of kernel based is it is business friendly as we can directly communicate to studio manager.

3.1 Kernel based approach

The Notation based y_i is response variable we define it for each movie y_i is much closer to normality than box office revenue the features we consider here are the textual variables extracted from each script along with its production budget. The distance metric between two observations is defined, based on (weighted) Euclidean distance as follows:

$$d(\overrightarrow{x_i}, \overrightarrow{x_l}) = \sqrt{\sum_{i=1}^{j} v_j^2 (x_{ij} - x_{lj})^2} \text{ ----------------(1)}$$

is a vector of "feature weights".

As shown that the conceptual argument above we set the value of θ by appealing to studios' domain knowledge. The studio managers typically look at no more than 10 "comps" when making a green-lighting decision. Therefore, we select θ such that any "comp" beyond the 10th will receive minimal weight this is achieved by setting θ so that on average the 10th comp receives a weight that is proportional to the density of a standard normal distribution at two standard deviations from the mode, Hence the 11th or further comps have weights that are negligible.

3.2 Calibration of Featured Weight ($\vec{v}$)

We calibrate the featured weight $\vec{v}$ as a starting point a reasonable "default choice" is to put equal weight on every variable i.e $V_j = 1$. We refer it as Kernel-I approach. We will evaluate its predictive performance verses kernel-II Approach that involves features weight. The proposed approach is based on cross validation to calibrated features weight $\vec{v}$ for kernel – II approach. We define Leave one out mean squared error, LOOMSE a key component of our objective function. We let i=1…n(n=265)index the scripts in training sample & let $\hat{z}_i(\theta,\kappa,\vec{v})$ be the predicted value of the log box office revenue of i^{th} script when all except the i^{th} script are used as training data. Z_i denotes actual log box office revenue for i^{th} script.

$$\text{LOOMSE}(\theta,\kappa,\vec{v}) = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i(\theta,\kappa,\vec{v}))^2 \text{------------------(2)}$$

3.3 Portfolio Selection

Now we demonstrate the potential economic significance of our proposed method & we conduct a hypothetical portfolio selection exercise so that we can compares the performance of the comps-based approach with our proposed Kernel-I/II methods. We consider the following portfolio selection setting. Suppose we would like to pick r scripts to form a movie portfolio.

First, based on the predicted box office revenue and the given production budget, we compute the predicted ROI of each of the 35 scripts in the holdout sample. Then, scripts in the oldout sample are ranked based on predicted ROI, and the r scripts that have the highest predicted ROI are selected. We vary from 5 to 20, and compare the ROIs of the overall portfolios which are selected by the comps based method Kernel-I and the Kernel-II method, respectively. The results are shown in Fig. 2. While there is a lot of variability in portfolio ROIs ((total box office – budget)/budget) across all methods, portfolios selected by Kernel-I and Kernel-II approaches consistently provide higher portfolio returns compared to those selected by the comps-based method. when r = 10 movies scripts are selected to form a portfolio, the selections by Kernel-I and Kernel-II method yield portfolio ROIs of 130.3 percent (Box office = $1184.7M; Budget = $514.5M) and 134.6 percent (Box office = $1236.3M; Budget = $527.0M), respectively, while the selection by the comps based method yields a ROI of 76.4 percent (Box office =$307.8M; Budget = $174.5M).8 Across different values of r(from 5 to 20), the median ROI of portfolios selected by Kernel-I and Kernel-II is around 134.0 and 134.1 percent (respectively), while the median ROI of portfolios selected by comps-based method is only around 83.9 percent. Thus, it is clear that the improvement in prediction accuracy afforded by the Kernel-I/II methods is also economically significant.

IV. CONCLUSION

The paper consist a methodology which is depend on the kernel-based approach to predict the box office potential of movie scripts at the point of green-lighting.

REFERENCE

[1]. G.J. Alexander and M.B. Alexandre, "Economic Implications of Using a Mean-VaR Model for Portfolio Selection: A Comparison with Mean-Variance Analysis," J. Economic Dynamics and Control, vol. 26, no. 7/8, pp. 1159-1193, 2001.

[2]. D.W.K. Andrews, "Asymptotic Optimality of Generalized CL,Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," J. Econometrics, vol. 47, no. 1991,pp. 359-377, 1991. [3] I.R. Blacker, The Elements of Screenwriting. Macmilan Publishing, 1998.

First Author: K. R. Dabhade, Asst.Prof, P.E.S College of Engineering, Aurangabad, India.

Second Author: Ms. S.S. Ponde, Asst. Prof, Computer Science Department, D.I.E.M.S,Aurangabad, India.