Movie Rating Prediction using Multimedia Content and Modeling as a Classification Problem

Fatemeh Nazary[1], Yashar Deldjoo[2]
[1] University of Pavia, Italy, [2]University of Milano-Bicocca, Italy
fatemeh.nazary01@universitadipavia.it,deldjooy@acm.org

ABSTRACT

This paper presents the method proposed for the recommender system task in Mediaeval 2018 on predicting user global ratings given to movies and their standard deviation through the audiovisual content and the associated metadata. In the proposed work, we model the rating prediction problem as a classification problem and employ different classifiers for the prediction task. Furthermore, in order to obtain a video-level representation of features from clip-level features, we employ statistical summarization functions. Results are promising and show the potential of leveraging the audiovisual content for improving the quality of existing movie recommendation systems in service.

1 INTRODUCTION AND CONTEXT

Video recordings are complex audiovisual signals. When we watch a movie, a large amount of information is communicated to us through different multimedia channels, in particular, the audio and the visual channel. As a result, the video content can be described in different manners since its consumption is not limited to one type of perception. These multiple facets can be manifested by descriptors of visual and audio content, but also in terms of metadata, including information about the movie's genre, actors, or plot of a movie. The goal of movie recommendation systems (MRS) is to provide personalized suggestions about movies that users would likely find interesting. Collaborative filtering (CF) models lie at the core of most MRS in service today and generate recommendation by exploiting the items favored by other *like-minded* users [2, 9, 10]. Content-based filtering (CBF) methods on the other hand, base their recommendations on the similarities between the target user's preferred or consumed items and other items in the catalog, where this similarity is defined by computing a content-centric similarity using descriptors (features) inferred or extracted from the item content, typically by leveraging *textual metadata*, either editorial, *e.g.*, genre, cast, director, or user generated, *e.g.*, tags, reviews [1, 8]. For instance, the authors in [12] developed a heterogeneous social-aware MRS that uses movie-poster images and textual description, as well as user ratings and social relationships in order to generate recommendations. Another example is [11], in which a hybrid MRS using tags and ratings is proposed, where user profiles are formed based on users' interaction in a social movie network.

Regardless of the approach, metadata are prone to errors and expensive to collect. Moreover, user feedback and user-generated metadata are rare or absent for new movies, making it difficult or even impossible to provide good quality recommendations a scenario known as the *cold-start* problem. The goal of the current MediaEval task [3] is to bridge the gap between advances and perspective in multimedia and recommender systems communities [7]. In particular, participants are required to use the audiovisual content and metadata in order predict global ratings of users provided to movies (representing their appreciation/dis-appreciation) and the corresponding standard deviation (characterizing users agreement and disagreement). This task is novel in two regards. First, the provided dataset uses movie clips instead of trailers [4–6], thereby providing a wider variety of the movie's aspects by showing different kinds of scenes. Second, including information about the ratings' variance makes it possible to assess users' agreement and to uncover polarizing movies [3].

2 PROPOSED APPROACH

The proposed framework can be divided into three phases:

(1) **Multimodal feature fusion:** This step is carried out in the multimodal phase for hybridization of the features. It aims to fuse two descriptors of different nature (e.g., audio and visual) into a fixed-length descriptor. In this work, we chose *concatenation* of features as a simple early fusion approach toward multimodal fusion.

(2) **Video-level representation building:** The novelty of this task is that it uses *movie clips* instead of movie trailers [3] in which each movie has several associated clips. This step aims to aggregate clip-level representation of features in order to build a video-level representation so it can be used in the classification stage. In this work, we adopted aggregation methods based on *statistical summarization* including *mean*(), *min*() and *max*() to obtain video-level representations of the features.

(3) **Classification:** The provided scores (global ratings and their stds) are continuous values. Our approach to the prediction problem consisted of treating it as a classification problem. This means prior to classification, the target scores are quantized to predefined values. We chose 2-level uniform quantization for global rating meaning the ratings were mapped to one of the values in the set $\{0.5, 1, 1.5, ..., 4.5, 5\}$. As for std, we chose 10-level plus 16-level uniform quantization where in the latter case, the higher number of levels were chosen to provide a larger resolution to the narrow distribution of std scores (std values are quite compact around [0.5-1.5] whereas global ratings are spread in the range [0-5]). Finally for classification, we investigated three classification approaches: logistic regression (LR), k-nearest neighbor (KNN) and random forest (RF) classifier.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

Table 1: Results of Classification in terms of RMSE, SoA: state of the art. The results are calculated on the development set. The 4 submitted runs are highlighted in bold selected from best unimodal and hybrid model.

		movie clips					
		Avg			**Std**		
feature	**modality**	**LR**	**KNN**	**RF**	**LR**	**KNN**	**RF**
i-vector	audio (SoA)	0.56	0.68	0.57	0.15	0.17	0.15
BLF	audio (traditional)	0.55	0.65	0.58	0.15	0.16	0.14
Deep	visual (SoA)	0.57	0.63	0.56	0.14	0.15	0.14
AVF	visual (traditional)	0.58	0.65	0.57	0.14	0.16	0.14
Tag	metadata (user generated)	0.48	0.55	**0.39**	0.14	0.15	**0.14**
Genre	metadata (editorial)	0.52	0.60	0.53	0.14	0.23	0.18
i-vector + BLF	audio+audio	0.55	0.65	0.55	0.15	0.19	0.15
i-vector + Deep	audio+visual	0.57	0.63	0.57	0.14	0.16	0.14
i-vector + AVF	audio+visual	0.58	0.65	0.54	0.14	0.19	0.15
i-vector + Tag	audio+metadata	0.48	0.53	0.39	0.15	0.19	0.15
i-vector + Genre	audio+metadata	0.52	0.58	0.56	0.14	0.13	0.13
BLF + Deep	audio+visual	0.55	0.64	0.54	0.15	0.19	0.15
BLF + AVF	audio+visual	0.55	0.64	0.57	0.14	0.19	0.14
BLF + Tag	audio+metadata	0.55	0.54	0.49	0.16	0.18	0.14
BLF + Genre	audio+metadata	0.55	0.65	0.56	0.15	0.19	0.15
Deep + AVF	visual+visual	0.59	0.69	0.57	0.14	0.18	0.14
Deep + Tag	visual +metadata	0.40	0.64	0.38	0.14	0.25	0.15
Deep + Genre	visual+metadata	0.57	0.63	0.58	0.14	0.16	0.14
AVF + Tag	visual+metadata	0.44	0.78	0.38	0.15	0.42	0.15
AVF + Genre	visual+metadata	0.58	0.67	0.56	0.14	0.20	**0.12**
Tag + Genre	metadata+metadata	**0.36**	0.54	0.46	0.15	0.18	0.14

3 RESULTS AND ANALYSIS

The results of classification using the proposed approach are presented in Table 1. Regarding the comparison of classifiers, we can note that RF is the best classifier among others usually generating the best performance for each feature or feature combination while KNN is the worst (note that KNN classifier is a lazy classifier). Thus, in reporting the results, we mostly base our judgment on results obtained from RF and in some cases on LR. The final submitted runs are selected based on the ones performing the best on the development set, which are highlighted in bold in Table 1.

Predicting average ratings: From the result obtained it can be seen that the performance of all audio and visual features, regardless of their type i.e., traditional or state of the art, are closely similar to each other. These results with a close margin look similar to the performance of the genre descriptor. In fact the difference between the best audio or visual feature and genre is 6-7% while this difference with tag can reach up to 45%. These results are interesting and confirm that user-generated tags assigned to movies contain semantics that are *well correlated* with ratings given to movies by user, even though the users of tags and ratings are not necessary the same. For multimodal case, one can note that simple concatenation of the features can not improve the final performance substantially compared with unimodal audiovisual features. The best results are obtained in cross modal fusion for i-vector + AVF (compare 0.54 v.s Genre: 0.53) and BLF + Deep (0.54). However for metadata-based multimodal fusion, the general observation is that audiovisual features can slightly improve the performance of genre and tag (e.g., compare AVF+Tag: 0.44 v.s. Tag: 0.48 for LR and 0.38 v.s. 0.39 for RF), hinting that they have a complementary nature which can be better leveraged if right a fusion strategy is adopted.

Predicting standard deviation of ratings: As for predicting standard deviation of ratings, for unimodal case, it can be seen that except genre feature with the worst performance, the rest of audiovisual features and tag metadata have very similar results. This indicates that genre is the weakest descriptor and compared to others less capable of distinguishing difference in users' opinions. Note that under LR, genre descriptor performs similar to other audiovisual features. For multimodal case, the results for majority of combinations are pretty similar regardless of the classifier type. The best performing combinations are AVF + Genre and i-vector + Genre with the RMSE equal to 0.12 and 0.13.

4 CONCLUSION

This paper reports the description of the method for the "Recommending movie Using Content: Which content is key" MediaEval 2018 task [3]. The proposed approach consists of three main steps: (i) multimiodal fusion, (ii) video-level video representation building and (iii) classification. Results of experiments using three classification approaches are promising and show the efficacy of audiovisual content in predicting user global ratings and to a lesser extent for predicting rating variance.

REFERENCES

[1] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender systems*. Springer, 139–166.

[2] Charu C Aggarwal. 2016. Neighborhood-based collaborative filtering. In *Recommender Systems*. Springer, 29–70.

[3] Yashar Deldjoo, Mihai Gabriel Constantin, Thanasis Dritsas, Markus Schedl, and Bogdan Ionescu. 2018. The MediaEval 2018 Movie Recommendation Task: Recommending Movies Using Content. In *MediaEval 2018 Workshop*.

[4] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2018. Audio-Visual Encoding of Multimedia Content to Enhance Movie Recommendations. In *Proceedings of the Twelfth ACM Conference on Recommender Systems*. ACM. https://doi.org/10.1145/3240323.3240407

[5] Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. MMTF-14K: A Multifaceted Movie Trailer Dataset for Recommendation and Retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys 2018)*. Amsterdam, the Netherlands.

[6] Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. 2018. Using Visual Features based on MPEG-7 and Deep Learning for Movie Recommendation. *International Journal of Multimedia Information Retrieval* (2018), 1–13.

[7] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2018. Content-Based Multimedia Recommendation Systems: Definition and Application Domains. In *Proceedings of the 9th Italian Information Retrieval Workshop (IIR 2018)*. Rome, Italy.

[8] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*. Springer, 73–105.

[9] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer, 1–34.

[10] Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 47, 1 (2014), 3.

[11] Shouxian Wei, Xiaolin Zheng, Deren Chen, and Chaochao Chen. 2016. A hybrid approach for movie recommendation via tags and ratings. *Electronic Commerce Research and Applications* 18 (2016), 83–94.

[12] Zhou Zhao, Qifan Yang, Hanqing Lu, Tim Weninger, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Social-Aware Movie Recommendation via Multimodal Network Learning. *IEEE Transactions on Multimedia* (2017).