Predictive Visual Analytics – Approaches for Movie Ratings and Discussion of Open Research Challenges

Mennatallah El-Assady, Wolfgang Jentner, Manuel Stein,
Fabian Fischer, Tobias Schreck and Daniel Keim

Abstract— We present two original approaches for visual-interactive prediction of user movie ratings and box office gross after the opening weekend, as designed and awarded during VAST Challenge 2013. Our approaches are driven by machine learning models and interactive data exploration, respectively. They consider an array of different training data types, including categorical/discrete data, time series data, and sentiment data from social media. The two approaches are only first steps towards visual-interactive prediction, but have shown to deliver improved prediction results as compared to baseline non-interactive prediction, and may serve as starting points for other predictive applications. Furthermore, an abstract workflow for predictive visual analytics is derived. We also discuss promising challenges for future research in visual-interactive predictive analysis, including design space, evaluation, and model visualization.

Index Terms—Visual Analytics, Interactive Prediction, System Design, Evaluation.

1 Introduction

Predictive analysis is an important part of data analysis, dealing with the problem of quantitatively or qualitatively assessing the outcome of a given process. Statistics and machine learning provide an abundance of methods for prediction, for example, regression and classification analysis. While such methods have been successfully applied in many application domains as expert tools, they typically require the choice of appropriate models, parameters and training data. Therefore, they are not directly applicable by non-expert users. Also, the assessment of prediction uncertainties remains difficult, especially if the prediction models are applied in a black box manner. Recent research has started to consider visual-interactive approaches for predictive analysis, aiming to improve the prediction process and enable wider user groups to use predictive tools. The quality of such predictions heavily depends on the used model and the underlying training data. Furthermore, background knowledge needs to be taken into account during model generation and analysis. Visual Analytics makes use of novel visualizations to capture such knowledge and allow users to steer the analysis process to eventually improve the prediction. In this work, we propose two approaches to predict movie ratings using visual analytics. Both approaches were awarded during the IEEE VAST Challenge 2013[1], an international contest to solve complex visual analytics tasks. In 2013, datasets were given to predict movie ratings and box office sales.

The contributions of this paper are (i) two approaches to predict movie ratings. While the first one is based on machine learning and designed to effectively interact with computational models, the second approach is driven by interactive data exploration. Furthermore, we propose (ii) a generalized workflow for predictive visual analytics. In addition, we contribute (iii) a discussion of promising challenges for future research in visual-interactive predictive analysis.

2 Related Work

While various commercial systems support predictive analytics, their usage of visualization is often limited to present the results. Visual analytics approaches on the other hand, make use of interactive visual representations to incorporate expert knowledge directly into the analysis and prediction process, which is recognized to be effective and helps the analyst to understand the data [11]. Such a visual analytics system was proposed by Lu et al. [8] also awarded during the VAST Challenge 2013. Their system is highly related to ours, because they use various visualizations for feature selection and for controlling the models to predict box office gross and movie ratings. Mühlbacher et al. [9] propose a framework to build and validate regression models to combine a qualitative analysis of relationship structures by visualization and a quantification of relevance for ranking features. This supports the analyst in generating multidimensional models for prediction. Other work [4, 10] also focuses on visual regression analysis to provide interactive means to support model validation. The need for integration of expert knowledge in the prediction can also be seen in applications for time-series prediction. Hao et al. [5] propose a system to preserve important peaks and patterns in the prediction model for time-series to predict power consumption and server utilization in data centers. The effective use of visual analytics is also shown in visual classifier analysis, such as building and exploring decision trees [12]. Task-dependent filters and support vector machines can also be steered by interactive visualization systems to analyze social media data [2]. In contrast to the mentioned systems, we propose two interactive Visual Analytics approaches tailored to the task of predicting movie ratings: An interactive machine learning [3] and a visual interactive approach [1] to cover two possible ways to address predictive visual analytics challenges.

3 Approaches

Our approaches enable visual-interactive prediction of user movie ratings and box office gross for the opening weekend of the US cinema market, as proposed during the VAST Challenge 2013. The tool developments were supported by continuous weekly evaluation of prediction results of current movies. Analyst feedback was provided by the challenge organizers for every submission including the real ratings and box office grosses for the opening weekend of a movie. This allowed to monitor and improve our systems. The provided data consisted of a weekly copy of the Internet Movie Database (IMDb)[2] containing a large set of actor profiles, movie metadata, and user ratings. Furthermore, a predefined set of Twitter messages filtered for each movie was made available.

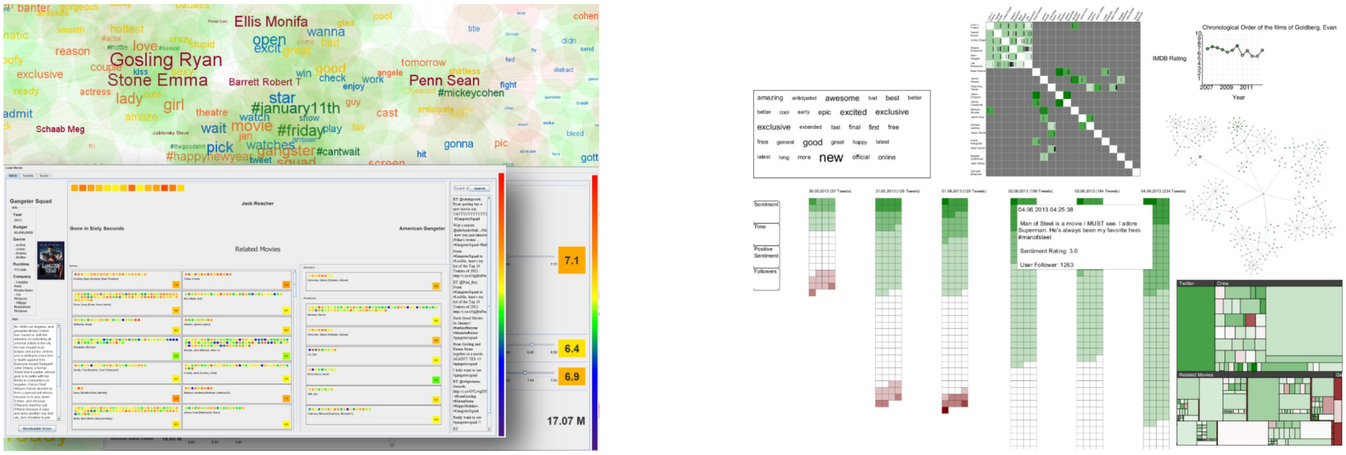
Our goal is allowing non-experts to create a semi-automatic prediction by using the opinions reflected in the social media, background

- *The authors are with Data Analysis and Visualization Group, University of Konstanz. E-mail: firstname.lastname@uni-konstanz.de.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

[1] `http://vacommunity.org/VAST+Challenge+2013`

[2] `http://www.imdb.com/`



(a) Interactive Machine Learning Approach

(b) Visual-Interactive Approach

Fig. 1. Our approaches based on machine learning (a) and visual navigation (b) to predict ratings and box office gross of upcoming movies.

knowledge of the user, and known facts from previous movies or actors. Our two approaches are detailed in the following subsections.

3.1 Interactive Machine Learning Approach

The first approach [3] is based on machine learning and is designed to effectively interact with the automated models (see Figure 1(a) for an illustration). Among different machine learning models for the given task, in our experiments, neural networks performed best, as evaluated with a 10 cross-fold validation on historic data. Because of variances in the error we train different neural network models with varying input vectors and refine the models for each movie genre separately as well as one for all genres together. A number of actors, the director, and two producers are considered as input vectors. By default, an average over all past performances for each cast and crew member is calculated.

In a pre-learning interaction phase the user may discard ratings of old performances which will modify the average. This interactive filtering allows the system to include background knowledge into the final prediction. The user is able to weight the cast and crew members by ranking them without any restrictions. The neural network models take the filtered and weighted input vectors and calculate their predictions on-the-fly. Also, the user is supported in the prediction task by a so-called Tweet graph showing the general presence, relations, and sentiments towards the members of the cast and crew, as reflected in social media.

In a post-learning interaction phase a weighted average is calculated to present a preliminary prediction score. The user may interact with the system and adjust this score within a certain limit (uncertainty range) defined by the standard deviation of the predictions of the different neural networks.
Since the predictions of the neural networks cannot be intuitively interpreted, two simple models assist the user in monitoring the automatic prediction. The first model considers related movies and calculates an average of their ratings. The second model calculates the average of the cast and crew members. Both predicted ratings can also be adjusted by the user within the models' uncertainty ranges.

The whole system is designed to provide fast feedback loops. Each interaction with the system results in an immediate change in the predictions. Metadata and general information for the movie as well as moods and opinions reflected in the social media are presented to assist the user. The user may interact with the system in different ways and at different points in the prediction process to integrate her expert and background knowledge. A demo of the system is also available in a video[3].

3.2 Visual Interactive Approach

Our second approach [1] relies mainly on the background knowledge and expertise of the user supported by interactive visualizations (See Figure 1(b) for an illustration). Effectively, this approach is a nearest-classifier, where the prediction is a weighted average of training data. The key task of the system is to support the user to search, compare, and weight relevant training data for the prediction. The user is presented with data overviews and detailed views on demand. The data is collected in a tree structure, where the tree root represents the prediction. Every leaf node in this data structure holds information about one entity. For example, a specific crew member, genre, or a related movie. Additionally, every node has an associated weight that determines its impact on the prediction. The weights can be modified by the user to influence the outcome of the prediction. Every node is colored to indicate its score. The change in the weight is immediately presented to the user as the ascending nodes change their scores as well.

A treemap visualization assists the user to detect outliers and provides a general overview over the node-weights. The weight of each node is represented by the size of its corresponding node in the treemap. In addition, detailed visualizations are provided, including detail views per actor, previous performances, and won awards. The user is also able to personalize the visualizations to view the data from different perspectives and increase the efficiency of the prediction.

To prevent an *early fixation* problem as known from decision psychology, we hide the prediction value while the user conducts the training data search. Only once the selection and weighting process is completed, will the final prediction be shown. A demo of the system is also available in video[4].

4 Results from VAST Challenge and Workflow

We here discuss results and experiences made during evaluation of our two approaches during June and July 2013.

Regarding the machine learning based approach (cf. Section 3.1), we trained the neural networks on historic movie rating data as starting points for our predictions. Therefore, the minimal mean squared error (MSE) is achieved when predicting long-term values. The ratings range from 1 to 10, out of which the MSE is 0.46 for long-term ratings. The MSE for the predicted ratings after the opening weekend is slightly higher with 0.61. Without user interaction, the MSE is higher (0.74). Therefore, it can be concluded that the background knowledge of the user considerably improves the accuracy of the prediction. This background knowledge includes facts such as important events that occur during the opening weekend or simply other movies that people tend to watch rather than the current one. An important

[3] http://youtu.be/XhJDPa9FNck

[4] http://youtu.be/M2gLhe6cQQA

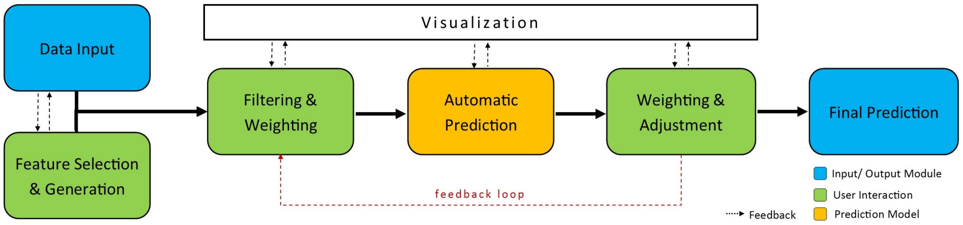


Fig. 2. Generalized Workflow Diagram: This workflow extends the general predictive analysis workflow and can be embedded into a KDD process. By adding a visualization component, the user can interact with the system in feedback loops.

observation is that the ratings at the opening weekend might tend to be rather extreme and average out over time. The automatic models are, therefore, unable to predict extreme ratings accurately.

Interestingly, the results of the visual interactive approach (cf. Section 3.2) are comparable. Specifically, a team of four students, all with interest and background knowledge in movies, worked for about 30 minutes with the tool for each prediction. Overall, the MSE for the opening weekend is 0.47. We notice that non-animated comedies are most difficult to predict. When animated movies are left out, the MSE improves to 0.25. As observed in the other approach, user interaction yields better results. This approach relies on the expertise of the user. Therefore, a group of users improves the prediction result.

We can summarize our practice in predicting with our tools in a workflow (cf. Figure 2). The usable data is limited by the available sources and also by the application. The user can interactively generate and select features according to her analysis goals. To integrate her background knowledge and opinion with the automatic prediction, the user may apply filtering and weighting on the selected features. On demand, visual feedback may be applied to show the user how the interaction influences the prediction. The adjustment of the predicted score provides a secondary opportunity to include the user in the prediction process. In case of using multiple automated models for prediction, weighting the models can be applied as post-learning interaction phase. Then, the adjustment should be limited to not let the user fully control the prediction. Visualizing and defining considerable limitations improves the usability and also the final prediction. As for the other user interactions, direct feedback is important.

5 Research Questions and Challenges

Inspired by the results of the VAST Challenge 2013 and the lessons learned during the development of both tools, this section discusses multiple elements which form interesting and promising future research challenges. First, practical system design choices are described. Second, methodological implications we see are discussed.

5.1 System Design

To create a visual-interactive prediction system, several design choices have to be made which directly influence the quality of the created predictions. Depending on the target user group and the available resources, the system requirements and components differ. This subsection will detail on the structure of the overall design space from different perspectives.

5.1.1 Requirements

Before developing a system, a requirement analysis is inevitable. Besides the important components for predictive systems, visual-interactive prediction systems require certain specifications:

Data To compensate the lack of background or expert knowledge in the input data, the additional integration of more users in a collaborative way should be considered to improve the prediction results. Semi-automatic systems benefit from users knowledge, but also require data for learning automatic models that assist or even perform the prediction. This training and test data should be consistent with data required by the model. As user input data may include changing factors over time, e.g., periodic/seasonal dependencies (weather, holidays) or changes with no periodic patterns (inflation), the models should not be trained on inconsistent data sets. Ideally, the system would be able to assess the quality of user input data and accept, reject, or request for update regarding user data.

User Expertise In both approaches presented in Section 3, the user interaction improved the prediction results significantly. However, a broad spectrum of user interactions also requires more domain knowledge. A visual-interactive approach might acquire the user to be more familiar with the domain, but generates, therefore, more transparent and intuitively interpretable results. On the other hand, a machine-learning driven approach learns patterns from training data and does not rely as much on user expertise, but generates mostly non-intuitive models and predictions.

Usability The required time and effort to generate a prediction is one of the main factors that determine the usage and acceptance of a tool. It is safe to assume, that if the target group of a designed system are domain experts, who would use the system to generate a professional prediction, the tradeoff between accuracy and effort is in favor of the accuracy. However, if the application is designed for laypersons, an adaptive system which offers more or less options depending on how much knowledge a user has, would be a good design choice. The tradeoff between the interaction and configuration freedom of a tool and the usability should always be kept in mind. Especially, when designing a predictive system for non-experts, which should – in the optimal case – limit the range of possible final results to reduce false predictions.

5.1.2 Components

As proposed by the workflow in Figure 2, a predictive visual analytics system consists of multiple modules. In the following, we describe several components we deem important:

Prediction Model Since the prediction model is the heart of the system, it needs to be carefully chosen according to the system requirements. Depending on the chosen model, the visualization and interaction components can be designed. In our approaches we experiment with machine-learning driven and statistically based prediction models.

Visualization Beside the visualization of the results and parameters of a prediction model, some models can be visualized, e.g., Decision Trees, which facilitates the understanding of the prediction results and uncertainties. However, other models, e.g., neural networks are typically, hard to visualize, hence, it remains challenging to monitor and steer certain models with visual interfaces.

Interactions and Feedback Loops To account for the missing external factors and to allow the user to steer the prediction by incorporating her background knowledge, the final result should be adjustable within a certain range. However, if the independent results of various prediction models fall within the same range, the uncertainty of the final prediction is low and the allowed interaction can be limited. By this approach, the user can incorporate her intuitive and subjective opinions in the prediction but is limited within the uncertainty range. This yields good results for trends with no extremes.

5.2 Methodology

We discuss also a number of methodological questions we encountered during our experimentation with visual-interactive prediction systems during the VAST Challenge 2013.

5.2.1 Transferability of Solutions Between Prediction Domains

Applying visual-interactive prediction from one domain to the other is seen a challenge. It is not clear how domain-specific solutions need to be, or whether, a given system design may apply to various prediction domains. As our introduced approaches are based on movie predictions, it would be interesting to conduct further research on applying and adapting these systems to other domains. Predictive analytics has numerous applications in many fields. Examples include Customer Relationship Management (where customer behavior is analyzed for marketing, sales, and customer service goals); in Risk Management (where fraud detection and credibility analysis are done); or in Health Care (where, e.g., risks of disease spreads and the correct medical condition of patients according to their symptoms are considered).

Given the abundance of prediction models, data types, and tasks, we expect that no single system design will be able to support all combinations in this space. We see experimental and challenge-based approaches as exemplified by the VAST Challenge series as promising to identify candidate designs and then, reason about transferability.

Tasks and Tools Beyond Prediction We may enhance predictive system to tackle new challenges beyond prediction. In the following, we list some of the tasks that could be solved with an adapted predictive system:

- **Discovery**: Extending a predictive system to be used as an exploratory tool to find data patterns and trends would be an extension, that is not only interesting for expert user, but also for people interested in discovering more information and revealing some underlying structures of certain datasets. Such a tool could also be applied on old datasets to reveal discrepancies between the actual data and predicted values.
- **Reassurance**: Showing the certainty of a predicted result is an important aspect that influences the confidence of the analyst. In addition, improving the current prediction accuracy by learning from old results could build a more trustworthy system.
- **Reporting**: Analysis provenance is an important challenge when prediction results need to be justified. Hence, it is important to integrate logging capabilities to collect usage data based on the user interactions with the system. This data helps to get an idea how and why a user, or group of users, arrives at a certain prediction. Communicating and reporting this provenance data strengthens and justifies the decisions, makes them comprehensible, and reveals the certainty of the final prediction.

5.2.2 Process Pipeline

The general visual analytics pipeline as introduced by Keim et al. [6, 7] can be specified to hold a predictive visual analytics pipeline as proposed in Figure 2. Research in the area of prediction analysis rarely includes visual interactive components with multiple user feedback loops and rather focuses on optimizing the prediction models and the automated data processing. The creation of a generalized process model for predictive visual analytics is, therefore, a desirable goal for future research in order to establish a foundation for this interdisciplinary field and target the added values of this symbiosis.

5.2.3 Evaluation

To fairly evaluate the effectiveness of a predictive system, benchmark data sets and adequate evaluation methods have to be developed. If we assume that user interaction is able to improve the prediction quality, we also need to assess the tradeoff between improvement in prediction quality, and user interaction investment. Ideally, we may find an optimal point where the best work allocation between algorithmic and interactive efforts are found. Evaluation is also challenging for collaborative predictive tasks. How can we attribute which system functionality fosters the collaboration, and supports optimal group work?

5.2.4 From Visual Data Mining to Visual Analytics

So far, visual-interactive prediction is in its beginning and a more fundamental understanding of ways to tightly integrate visualization, interaction, and prediction for improved results is needed. Research is needed to determine if and how traditional, non-interactive predictive models can be extended to semi-automatic prediction systems to improve the prediction quality, the usability of the system, or the understanding of the prediction results. State-of-the-art prediction tools mainly use visualizations either during data preprocessing stage (at the beginning of the prediction pipeline), or for result visualization (at the end of the prediction pipeline). In that end, they follow the Visual Data Mining paradigm, namely, utilizing visualization to support certain steps in the KDD pipeline. However, how to tightly integrate algorithms, model visualization, and visual-interactive steering is a challenging problem. We argue that by experimentation we may be able to deduct promising combinations of components in this large design space.

6 Conclusions

We presented two approaches to predict movie ratings and box office gross. One approach focuses on machine learning while the other is driven by interactive data exploration. Based on our work and evaluation of these approaches, we also developed a workflow model for visual-interactive prediction. Furthermore, we gave a discussion of promising challenges for future research in the area.

References

[1] F. Al-Masoudi, D. Seebacher, M. Schreiner, M. Stein, C. Rohrdantz, F. Fischer, S. Simon, T. Schreck, and D. A. Keim. Similarity-Driven Visual-Interactive Prediction of Movie Ratings and Box Ofce Results. In *IEEE VAST Challenge USB Proceedings*, 2013.

[2] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2022–2031, 2013.

[3] M. El Assady, D. Hafner, M. Hund, A. Jger, W. Jentner, C. Rohrdantz, F. Fischer, S. Simon, T. Schreck, and D. A. Keim. Visual Analytics for the Prediction of Movie Rating and Box Ofce Performance. In *IEEE VAST Challenge USB Proceedings*, 2013.

[4] Z. Guo, M. O. Ward, and E. A. Rundensteiner. Model space visualization for multivariate linear trend discovery. In *IEEE VAST*, pages 75–82, 2009.

[5] M. C. Hao, H. Janetzko, S. Mittelstdt, W. Hill, U. Dayal, D. A. Keim, M. Marwah, and R. K. Sharma. A Visual Analytics Approach for Peak-Preserving Prediction of Large Seasonal Time Series. *Computer Graphics Forum*, 30(3):691–700, 2011.

[6] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008.

[7] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010.

[8] Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski. Integrating predictive analytics and social media. In *Proc. IEEE Conference on Visual Analytics Science and Technology*, volume 2014, 2014.

[9] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013.

[10] H. Piringer, W. Berger, and J. Krasser. Hypermoval: Interactive visual validation of regression models for real-time simulation. *Comput. Graph. Forum*, 29(3):983–992, 2010.

[11] S. J. Simoff, M. H. Böhlen, and A. Mazeika, editors. *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, volume 4404 of *Lecture Notes in Computer Science*. Springer, 2008.

[12] S. van den Elzen and J. van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *Proc. IEEE Visual Analytics Science and Technology*, pages 151–160. IEEE, 2011.