Project Proposal: Movie Genre Classification

Matt McCreesh Kaitlynn Prescott

1 INTRODUCTION

For our final project, we will be implementing machine learning algorithms for predicting the genre of a movie based on different features. Features include important aspects of the movie and its production. This will be a supervised multi-class classification problem. This is an important problem for use in marketing and movie recommendations on streaming platforms. If a streaming service can accurately predict the genre of new movies that are entering their service, they can categorize them for viewers who want to see movies of a specific genre. This allows them to predict and provide more fine-grained genres than the movie might come labeled as. It also can give the benefit from a marketing perspective of ensuring that the released plot overview matches the genre it is aimed for, as a drama that has a description that comes of comedic may need to have changes to the released plot overview to better market the movie. Some potential applications of this include learning the common genres of movies in which certain actors or actresses appear, or which type of movies certain directors choose to direct. With movies becoming such a dominating force in the entertainment industry, a tool like this can be extremely helpful to moviegoers and potential filmmakers alike. Our project is in scope for this class as we will use several supervised classification techniques including logistic regression, a neural network, a bag of words approach for text classification, and more techniques we cover in class.

2 BACKGROUND

Classifying movie genres is not a new problem. One approach to movie genre classification was in a paper titled "Characterization of Movie Genre Based On Music Score" [3] which also appeared in the IEEE database. This research involved classifying movie genres based on non-vocal music from films. They used Support Vector Machines to do both pairwise and multiclass genre classification considering genres of romance, horror, drama, and action. One drawback of this is that it requires access to music scores from movies which is not always available and is not available in our dataset for this study. In other previous work titled "Movie Genre Classification with Convolutional Neural Networks" [2], video of movie trailers was used to classify the genre of the movie using a convolutional neural networks. A limitation of this approach is that it requires access to many movie trailers. Storage can become an issue if downloading trailer video of movies in bulk. As our data is different, we will take a different approach than the approaches described in these papers.

One paper more applicable for our project is "Movie Genre Classification from Plot Summaries using Bidirectional LSTM" [1]. This paper that appeared in the 2018 12th IEEE International Conference on Semantic Computing which discussed using Long Short Term Memory, a recurrent neural network, to classify movie genres based on the movie's plot summary. They describe considering the genre information represented by each sentence using Bi-LSTM. They also took a document level approach analyzing summaries as a whole with the LSTM approach. They compared this to using standard RNNs (recurrent neural networks) at a sentence and document level. Lastly, they compared this to a logistic regression model using bag of words and TD-IDF. In this research, movies were classified by genres of thriller, horror, comedy, and drama. Using the sentence level Bi-LSTM approach, precision, accuracy, macro f1, and micro f1 score were all between 67 and 68 percent. One limitation is that the data limited itself to only text and not take into account other features. It may be possible to get even better analysis using other features that can easily be found about movies. A drawback of the paper is that it only classified movies with coarse grained genres, while in our research we will try to predict more fine-grained genres and possibly do multilabel classification.

3 DATA

The data will contain important information about the movie. The field we will be using for predictions is the genre category. The remaining fields, including the title, director, cast, release year, revenue, and plot description, will be used for training. We will primarily focus on the plot overview for testing our algorithms. The dataset we will be using comes from the Kaggle Box Office Prediction competition, but instead of predicting box office revenue we will predict genre. The dataset can be found at https://www.kaggle.com/c/tmdb-box-office-prediction/data.

There are many fields available in this dataset. The first column is id which is not useful and will not be considered during classification. Another feature is belongs to collection, which tells if a movie is a part of a series or collection of movies and which collections it belongs to. Budget is another category which tells how much money was spent on production of a movie, while revenue tells how much money it grossed. Homepage gives a link to the homepage of the movie, though many movies are missing this feature and sometimes it points to invalid URLs so it might not be a great feature to consider. Another feature is imdb id, which can be used to reference data on IMDB. Features like original title and original language are self explanatory as they give the language and the title of the movie. Overview gives a brief overview of a movie. The popularity feature is a real valued feature determining how popular a movie is. Poster path provides a path provides a path to an image of the poster of the movie. Production country and production company are fields that tell where a movie was produced and in what country it was produced. The release date feature tells when a movie was released, while runtime tells how many minutes a movie is in duration. Spoken languages tell which languages are spoken in the movie. Status tells if a movie is released or not, but it appears that all of them are released so that is not necessarily useful. Tagline is a one sentence promotional line about the movie. Keywords are keywords given about the movie that are likely to be very useful for classifying movies by genre. Title is the title of the movie, cast is the actors in the movie, and crew is the movie's crew. Many of these features might not be useful and may be removed. Other features, like tagline and keywords, can probably be combined with the movie description to form a larger and stronger feature.

4 METHOD

Our objective is to be able to predict, with reasonable accuracy, the genre of a movie. This is within the scope of this class as we will use several methods we have discussed in lectures and other techniques that will be discussed in future lectures according to the course schedule.

We will approach this problem by comparing existing approaches, such as logistic regression, nearest neighbors, and neural networks, and analyzing them based on accuracy and efficiency. We will start by recognizing that one of our features, the plot overview, is purely text based and can be represented as a "bag of words". There is also a feature that provides references to movie posters, so this feature would be image based if we choose to use it. Revenue is a real value feature that we will do feature scaling in before using. Once we have done feature extraction, feature scaling, and feature selection, we will build multiple models. One will be using logistic regression to attempt to build a linear classifier for genres. We will also attempt to do k nearest neighbors for classification of genres. Lastly, we will attempt to use a neural network to classify genres. Specifically, we will use LSTM, a Recurrent Neural Network on the description of the movie to classify genre based on description.

We will attempt to do multilabel classification, and there have been some papers about multilabel classification problems. Sklearn includes modules for multilabel classification. In case this problem becomes too complicated, a fallback would to modify the output variable to make features more coarse grained and not multilabeled.

5 EVALUATION

We will evaluate each algorithm in terms of accuracy, precision, and efficiency, and we will compare the results of each. We will evaluate model parameters with cross validation, and then choose the best model parameters to train the entire training data set with. We will then report our results of accuracy and precision on the testing dataset. By comparing metrics for each model on the testing data, we will be able to determine the best algorithm we implement with respect to each evaluation technique. Analyzing the algorithms in this way will allow us to compare the benefits and disadvantages of each algorithm, and give us a stronger understanding of the pros and cons of common machine learning techniques.

6 REFERENCES

[1] A. M. Ertugrul and P. Karagoz, "Movie Genre Classification from Plot Summaries Using Bidirectional LSTM," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, 2018, pp. 248-251.
[2] G. S. SimÃţes, J. Wehrmann, R. C. Barros and D. D. Ruiz, "Movie genre classification with Convolutional Neural Networks," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 259-266.
[3] A. Austin, E. Moore, U. Gupta and P. Chordia, "Characterization of movie genre based on music score," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 421-424.