Genre Classification for Movies

Anonymous

1 Introduction

Movies have been a staple of modern entertainment for more than one hundred years, arguably since the first feature film in 1906. With the growth of the Internet Age at the turn of the 21st century, companies such as Netflix and Hulu have made movie-watching an instantaneous experience. Attempts to organise the ever-expanding library of movies by genre have typically been carried out by hand, therefore there is a growing need for a method which can be efficient while maintaining high accuracy.

The task is to classify, and with the wealth of information about movies available online, machine learning is well placed to succeed in this endeavour. Given enough raw data about movies, with their correct genre labels, supervised learning methods such as logistic regression, decision trees and neural networks can learn the relationship between the raw data and the genre, and use that relationship to make predictions about movies for which only the raw data is available.

Neural networks have received much attention with regards to genre classification. For example, Wehrmann & Barros (2017) introduced a novel deep neural network architecture to analyse raw trailer footage which outperformed state of the art benchmarks in genre classification. In addition, Bahuleyan (2018) uses only a convolutional neural network to compare the effectiveness of raw spectrogram data and hand-picked audio features for music genre classification. However, comparatively little attention has been given to methods such as logistic regression and decision trees, perhaps due to their simplicity and theoretical drawbacks. This paper investigates the performance of these two methods using a dataset from Deldjoo et al. (2018) and Harper et al. (2015) and performs error analysis to better understand their behaviour with respect to their theoretical properties.

2 Data Preparation

The dataset used is a merger of audio-visual features extracted from each movie's trailer (Deldjoo et al. 2018) and text information including the title, year and tags (Harper et al. 2015). The movie's tags and audio-visual features were deemed to be the most useful features given their repetitive nature within a genre, and so were the focus for the classification tasks. The training dataset was unbalanced, the exact distribution of genres is shown in Table 1.

Genre	Frequency
Romance	791
Drama	713
Thriller	598
Comedy	583
Sci_Fi	417
Fantasy	298
Mystery	270
Horror	244
War	241
Crime	237
Documentary	207
Musical	154
Children	106
Adventure	104
Action	86
Western	83
Film_Noir	78
Animation	30

Table 1: Genres and their frequency in the training set.

The text data in the tags was transformed using a common method known as one-hot encoding. Each unique word within the training set is assigned as a feature, with a value of one if it appears in the movie's tags, and zero if it does not. Thus, the tag data is transformed into a sparse matrix. This is useful for classifiers such as logistic regression which require numerical features that can be multiplied with the feature weights. Common words such as 'a' and 'the' are generally removed at this step as they do not assist in discriminating between classes, however these were not present in

the tag data.

3 Baseline Measure

A multinomial Naive Bayes classifier was selected as a baseline of performance for evaluating decision trees and logistic regression. Naive Bayes is a very efficient algorithm, and in practice it performs well on the task of text classification, despite its restrictive assumption of feature independence (Raschka 2017).

Applied to only the tag data in the dataset, the Naive Bayes classifier achieved an accuracy of 39.5%, and an f-score of 37.5%. As this was the baseline, further error analysis was not performed.

4 Decision Tree

A decision tree is a simple machine learning method for classification which uses 'rules' to partition the data and reduce its entropy. The lower the entropy, the more likely a 'random' guess will be the correct class.

The first decision tree model DT1 trained was unrestricted with respect to the number of features to utilise. This enables the model to fit the training data very closely, and can create a model that is highly complex and prone to overfitting. This issue is demonstrated in Table 1, as the model has learned the training data so proficiently that it can predict with 100% accuracy, but only when given the training data. The validation data performs comparatively poorly at 20.4%, worse than the Naive Bayes baseline.

Model	Training accuracy	Validation accuracy
DT1	1.0	0.204
DT2	0.273	0.241

Table 2: Accuracies on training data and validation data for two decision tree models.

To reduce overfitting in a decision tree model, the depth of the tree was limited for DT2. This puts an upper bound on the complexity of the model. Given that the model adds each decision rule 'greedily' based on entropy reduction, this should limit the model to generalisable rules. The parameter for the maximum depth was iterated to maximise performance, with the results shown in Table 1. The training data accuracy fell to 27.3%, and the validation data accuracy increased to 24.1%.

The reduced difference between the two accuracies indicates that the issue of overfitting has been rectified, and the reward is a modest improvement in performance. However, the performance is still underwhelming, and the reason may lie in the 'greedy' nature of the model's algorithm. Greedy algorithms do not backtrack after a step is made, so they can become trapped in local optima. When the model first begins training, it has a choice of 334 features to create the first decision rule, and so the likelihood of entering a 'dead end' with regards to performance is much higher than for models with only a few features to select from. To counteract this, Principal Component Analysis can reduce the high dimensionality of the data, and ensemble methods can mitigate the issue by combining several decision trees together, both of which would be suitable for future work.

5 Logistic Regression

Logistic Regression is a more powerful algorithm for classification which uses a linear combination of numerical features to generate a probability that an instance belongs to a particular class. In the case of this multi-class problem, a probability is generated for each class, and the most probable class is assigned at the output. The key assumption for logistic regression to work is that there exists a linear separation between classes. If the assumption holds, the algorithm performs well, and given that the audio and visual data has already had features extracted from it, any nonlinear relationships may have already been taken care of.

However, the curse of dimensionality can make finding a linear separation trivial, at the cost of overfitting the data. With 334 features, this is likely to impact the model's performance, and so lasso regression is used to reduce this overfitting while also mitigating the impact of irrelevant features, in effect decreasing the dimensionality of the data.

Model	Accuracy
LR1	0.428

LR2	0.348
LR3	0.615

Table 3: Accuracies for three logistic regression models.

Table 3 shows that for the first model trained, LR1, an overall accuracy of 42.8% was achieved for the validation data, an improvement on the Naive bayes baseline. Analysis of the precision and recall scores for each class show 'sci-fi' and 'documentary' as the most accurately predicted classes, likely due to uniquely identifying tags such as 'aliens', 'future' for sci-fi, and 'true story', 'biography' for documentary. 'Film noir' and 'adventure' were not predicted correctly at all, however with only two and four instances respectively in the validation dataset, it is difficult to analyse the cause.

Genre	Precision	Recall
Adventure	0.0	0.0
Crime	0.5	0.2
Musical	0.17	0.1
Comedy	0.48	0.58
Drama	0.33	0.63
War	0.77	0.33
Thriller	0.37	0.54
Sci Fi	0.77	0.63
Documentary	0.65	0.61
Western	0.0	0.0
Animation	0.5	0.33
Romance	0.39	0.35
Film Noir	0.0	0.0
Mystery	0.75	0.17
Horror	0.38	0.63
Children	0.5	0.33
Action	0.0	0.0
Fantasy	0.5	0.33

Table 4: Precision and recall scores for each genre for model LR1.

There is evidence that the unbalanced nature of the dataset as shown in Table 1 may have a deleterious effect on the model's predictions. The classes 'comedy', 'drama', and 'thriller' were vastly over-predicted, indicated by their low precision but high recall scores, and these make up three of the top four classes in terms of frequency in the training data. A weighting was applied for model LR2 to each instance during the training phase, inversely proportional to the instance's class frequency. While the precision and recall have become balanced for the aforementioned classes, overall the accuracy has dropped (Table 3). As the validation data is also unbalanced in a similar way to the training data, it is advantageous to over-predict the most popular classes in order to maximise the total accuracy.

True genre	Predicted genre	Frequency
Drama	Romance	17
Mystery	Thriller	10
Comedy	Romance	7

Table 5: Top 3 mis-predictions for model LR1.

Finally, a closer look at the confusion matrix in the first logistic regression model reveals an interesting trend. The most common mis-predictions (Table 5) are predicting romance instead of drama (17 times), thriller instead of mystery (10 times), and romance instead of comedy (7 times). While this is a classification task that requires a single output label, many movies are cross-genre, or purposefully sit on the boundary between genres. Movies which rightly deserve two labels would then be pared down to one in the dataset, causing confusion for the machine learning classifiers. The top fourteen genres according to box office revenue for the past twenty years illustrate this problem (Table 6). Romantic-comedy is highly popular, romance itself does not exist (it is absorbed into drama), and mystery does not exist either (it is absorbed into thriller/suspense).

Genre	Market share
Adventure	27%
Action	20%
Drama	16%
Comedy	15%
Thriller/Suspense	8%
Horror	5%
Rom-com	4%
Musical	2%
Documentary	1%
Black Comedy	1%

Table 6: Top 10 genres by market share 1995-2020 (The Numbers, 2020).

This lack of clear definition between classes violates the linearity assumption for logistic

regression and may be the reason it cannot achieve higher results. Model LR3 attempts to highlight this cross-genre issue by predicting two genres for an instance, and outputting true if one of these is the correct genre. This results in an accuracy spike to 61.5% (Table 1). Of course, this method will cause an expected increase in accuracy even if the instances were not cross-genre, as it gives the classifier a 'second chance', however the size of the jump is worth considering, and may indicate that 100% accuracy on such a dataset is not feasible.

6 Conclusion

The paper implemented and analysed the use of decision tree and logistic regression models on a dataset of textual 'tags' and audio-visual features from each movie's trailer.

Decision tree models performed poorly even after restricting the depth of the tree, and it is put forth that this is primarily due to the large number of features, limiting the effectiveness of the model's greedy algorithm. Feature reduction techniques may prove useful for future work.

The logistic regression model outperformed the baseline, though the unbalanced nature of the dataset caused significant over-prediction in high-frequency classes. Weighting the classes during training did not improve overall accuracy. More instances of the low-frequency classes may have made a greater difference. It was also noted that a significant portion of movies share two genres, and that some genres may be a subset of others and therefore a multi-label output may allow the classifier to match reality more closely.

References

Bahuleyan, H. (2018). Music Genre Classification using Machine Learning Techniques. *Computer Science, Engineering. Arxiv.* https://arxiv.org/abs/1804.01149

Deldjoo, Y. and Constantin, M. G. and Schedl, M. and Ionescu, B and Cremonesi, P. (2018). MMTF-14K: A Multifaceted Movie Trailer Feature Dataset for Recommendation and Retrieval. *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys, 2018.*

Harper, F. and Konstan, J. A. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19.*

Raschka, S. (2017). Naive Bayes and Text Classification I - Introduction and Theory. *Computer Science, Machine Learning. Arxiv.* https://arxiv.org/abs/1410.5329

The Numbers (2020). Genre Movie Breakdown 1995-2020. *The Numbers: Where Data and Movie Business Meet.* https://www.the-numbers.com/market/genres

Wehrmann, J. and Barros, R. C. (2017). Movie genre classification: A multi-label approach based on convolutions through time. *Applied Soft Computing, 61, 973-982.* https://doi.org/10.1016/j.asoc.2017.08.029