COMP90049 Introduction to Machine Learning
Project 2 Report

Anonymous

1 The task

Nowadays, movies have become one of the most affordable and enjoyable entertainments. Different types of movies pursue the same optimal goal, which is to deliver its own story, ideas to the audiences. However, the movie genres can be the main considerable reason when it comes to choosing a movie to watch. In this project, Dataset are trained through models in order to classify the movie genres of unseen movie data instances. The hypothesis of the project is Multilayer Perceptron would outperform both Naive Bayes and Logistic Regression models given such datasets with diverse features, because of its remarkable ability of automatically learning on features. To test the hypothesis, multiple methods have been implemented and compared throughout the process.

2 A short summary of some related literature

From previous works, there are multiple ways shown to classify movie genres given different features. Soria et al. (2008) attempted to train and predict by analysing the visual feature through complicated learners such as a multi-entry neural network. On the other side, training on movie subtitles has also been tried with both supervised learners Multinomial Naive Bayes, and Multilayer Perceptron algorithm (Pieters & Wiering, 2018). The experiment result showed that neural networks with optimized parameters often outperform simpler models. However, simple classifiers can also achieve reasonable results with appropriate feature engineering. In this project, the given dataset contains both textural features and numerical (float vector) features. Therefore, different feature engineering methods can be very effective in learners' performance.

3 A conceptual description

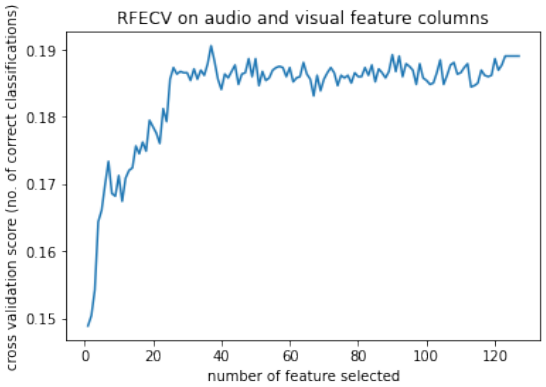
3.1 Missing values

By observing the training dataset, 3 instances are missing "title" feature value, which is negligible compared to its total amount of 5840 training instances. Thus, the 3 instances with missing value were removed directly from the datasets.

3.2 Feature Selection

Given a large dataset, feature selection by inspection was first performed. "movieId", "YTId" and "year" feature columns were dropped as they do not seem to contain meaningful information to the movie genres classification.

There are 127 numerical feature columns related to "audio" and "visual", such many features can have overmuch weight in training. Therefore, Recursive Feature Elimination with cross-validation method was used and selected the best 37 feature columns (Plot 1).



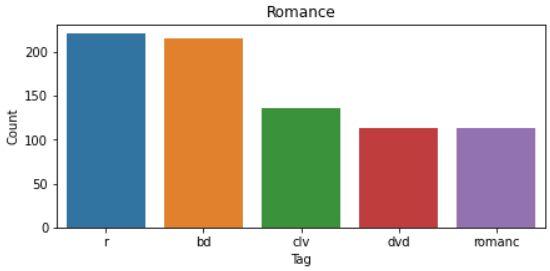
Plot 1- cross-validation score VS. number of features selected.

3.3 Feature Engineering

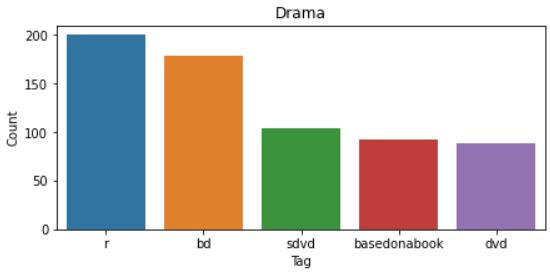
3.3.1 "tag" and "title" features

Since tag and title feature consist of English words, they are first converted to lowercase and all punctuations were removed. Removing English stopwords was then applied as these

words do not relate to the true meaning of the feature values. It also helps to lower the dimension space in "bag of words" based feature representation methods later in the process. Moreover, the feature values have words with similar meanings but in different forms, such as “vampire” and “vampires”, “beautiful” and “beauti”. To group similar meaning words, stemming was performed which reduces inflection in words to their root form. After visualizing on tag feature values, the result shows words such as “r” and “bd” rank in the top 3 appearing words most time regardless of the true class labels (Plot 2 & Plot 3). Consequently, such words were removed from tag feature values.



Plot 2- Top 5 appearing tags for instance with genres “Romance”.



Plot 3- Top 5 appearing tags for instance with genres “Drama”.

3.3.2 Audio and visual feature columns

It is found from the observation that "audio" and "visual" feature columns consisting of both positive and negative float numbers. Also, "visual" feature values lie over a broader range comparing the "audio" feature values. Min-Max Normalization was applied to standardise feature values fits in a range [0, 1], which ensures to weigh each feature column equally in their representation in training.

3.4 Text vectorization

To feed our dataset into different machine learning models, text feature values have to be encoded into an integer or floating-point representation. 2 different vectorization methods were attempted, word count and TF-IDF vectorization. Thus, features can be contacted to form different combinations as the input dataset.

3.5 Learners

Zero R was initially performed as the baseline method. To improve the accuracy based on the baseline performance and testing on the proposed hypothesis, Naive Bayes (2 different types), logistic regression, and Multilayer Perceptron were implemented. These classifiers were chosen because of 2 reasons. First, Naive Bayes learns a model the joint probability *P(x, y)*, and using the Bayes rule to obtain the conditional distribution. On the other side, logistic regression optimizes *P(y|x)* “directly”. It can be interesting to see how the performance differs between generative and discriminative models. Second, to demonstrate the notable learning ability of Multilayer Perceptron, it is intriguing to compare it with simpler algorithms, Naive Bayes and Logistic Regression.

4 Evaluation of classifier(s) over the validation dataset

The logistic regression model was performed based on various combination of features over the validation dataset (Table 1). This is done because the same training data input should be used when comparing the performance of different models. Testing on datasets with different feature combinations shows which one is more effective on classifying the right genres.

Feature combination	Accuracy
“tag” (TF-IDF) feature only	0.375
“tag” (TF-IDF), “audio” and “visual” vector form	0.413
“tag” and “title” (TF-IDF)	0.381

Table 1- Performance of applying LR on various feature combinations.

Logistic regression is also performed on the merged dataset of "tag" (count vector), "audio" and "visual" features (Table 2).

Feature combination	Accuracy
"tag" (count vector), "audio" and "visual" vector form	0.395

Table 2- Performance of applying LR on various feature combinations continued.

As the result shows, the "tag" (TF-IDF), "audio" and "visual" feature combination achieves better accuracy. Therefore, this combination of features is mainly used for testing on other classifiers. Multinomial Naive Bayes, Complement Naive Bayes and Multilayer Perceptron were applied to the dataset with the same feature combination (Table 3).

Classifier	Feature Combination	Accuracy
Zero R (baseline)	"tag" (TF-IDF), "audio" and "visual" vector	0.171
Multinomial Naive Bayes	"tag" (TF-IDF), "audio" and "visual" vector	0.388
Complement Naive Bayes	"tag" (TF-IDF), "audio" and "visual" vector	0.395
Logistic Regression	"tag" (TF-IDF), "audio" and "visual" vector	0.413
Perceptron	"tag" (TF-IDF), "audio" and "visual" vector	0.361
Multilayer Perceptron	"tag" (TF-IDF), "audio" and "visual" vector	0.435

Table 3- Performance of applying different learner.

5 Contextualises the behaviour of the method(s)

From the above table, TF-IDF representation of "tag" feature beats the count vector representation when combining with "audio" and "visual" pre-processed features and tested through logistic regression. This can be due to that TF-IDF balances the term frequency with its inverse total appearing frequency in all words appeared in the "tag" feature values. The most appearing "tag" terms are not meaningful. TF-IDF assign these tag terms with a very small weight so that these tag terms would not be treated as the important terms when learning.

From the above measures, input dataset ("tag" TF-IDF, "audio" and "visual" vector form) is the best combination of features for movie genres classification.

Zero-R has been used to obtain the baseline which was 0.171 with every instance predicted to "Romance" genres.

From various types of Naive Bayes model, Multinomial Naive Bayes algorithm was implemented, since that each $p(fi|c)$ is treated as a multinomial distribution and it works well with cleaned text data as words can easily be turned into frequency. It achieved an accuracy of 0.388 which is reasonable. To further improve the performance, Complement Naive Bayes algorithm was applied and reached an accuracy of 0.395. This is because Complement Naive Bayes learns parameters using data from all classes but not the true labelled class. Rennie, Shih, Teevan and Karge (2003) have found that CNB works better for multi-label classification problems as it reduces the bias of the weight estimates.

Logistic regression achieved performance 0.413, which is slightly higher than Complement Naive Bayes' performance. The logistic regression works well particularly with frequency-based features and large dataset. Unlike Naive Bayes models, it does not consider the feature independence assumption, therefore feature relations can be captured. For example, for a horror movie, "audio" and "visual" data can be related to its "tag" value. The other reason is that the generative model (Naive Bayes) tends to have a higher asymptotic error when the dataset is large (Ng & Jordan, 2001). In other words, Naive Bayes reaches a greater absolute error when working given dataset.

Before moving to the neural network, a simple Perceptron method was built as the baseline for Multilayer Perceptron and achieved an accuracy of 0.361. Multilayer Perceptron enables hidden layers which helps to improve the baseline accuracy. However, there are many

hyperparameters and can have a strong impact of the learner's performance. In this project, only the following ones were concentrated and turned through GridSearchCV (Table 4).

Hyperparameter	Optimized value
Hidden layer size	(200, 100, 50,), three hidden layers, with 200,100 and 50 hidden units each layer.
Solver	lbfgs

Table 4- Optimized Hyperparameter of Multilayer Perceptron.

With optimized hyperparameters, the Multilayer Perceptron outperforms all other models with an accuracy score of 0.435. Backpropagation is responsible for its learning and weights updating helps the network with automatic feature learning. The result implies that Multilayer Perceptron has a better ability to learn data with diverse features. However, the result cannot be explained in an easily comprehended form due to its multiplicity from heuristics, compared to classical statistical machine learning algorithms.

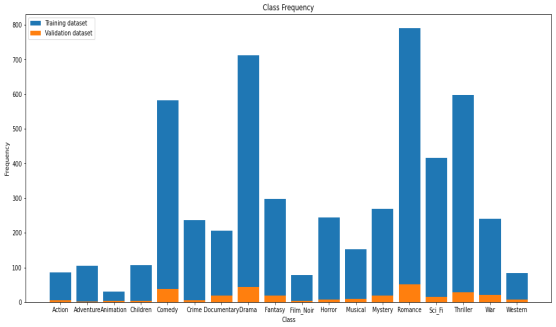
6 Error analysis of the method(s)

Although all attempts have a drastic improvement on the baseline, they are still far from being feasible in real-world application with the current performance. This section will mainly focus on the error analysis of Multilayer Perceptron's performance as it ranks the best among systems.

The backpropagation algorithm used in the Multilayer Perceptron can be problematic as it may lead to non-optimal results if it ends up finding the local minimum. In the experiment, the learning rate was set to be 0.01 and no further adjustment was made to test whether the learning rate was appropriate.

From previous discussion, feature selection was only performed on the combination of "audio" and "visual" data. The selected 37 columns have little weights when contacted with the large, vectorized "tag" feature columns. Moreover, the selected feature columns cannot guarantee to be the best fit when merging with the "tag" vector as the feature distribution is biased. More experiments should have been done to select the most appropriate feature columns.

The dataset has an imbalanced class distribution was discovered (Plot 4), applying instance selection methods can have a strong impact on the accuracy score of Multilayer Perceptron (Zhao, Xu, Kang, Kabir, & Liu, 2014). Nonetheless, the same training instances have been used for training throughout the analysis which could constraint further performance improvement.



Plot 4- Class distribution of Training and validation datasets.

7 Summarises the principal conclusions and future work

Multiple learners have been tested with the same combination of data features. From the result, Multilayer Perceptron outruns both of Naive Bayes and Logistic Regression. It is an understandable result as Naive Bayes suffers from the fast speed of reaching its asymptotic in large dataset. Although Logistic Regression reached can be considered as the most suitable model for this classification problem, considering its acceptable performance and decent computational time. On the other hand, Multilayer perceptron its remarkable ability to learn features and is proved to be more favoured when working under diverse feature datasets. For future work, it can be interesting to build a hybrid algorithm and compare its performance with both Naive Bayes and logistic regression. It is also essential to spend time on turning the hyperparameters of Multilayer Perceptron and record its improvement.

8 A bibliography

Deldjoo, Y., Constantin, M. G., Ionescu, B., Schedl, M., & Cremonesi, P. (2018, June).

MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference* (pp. 450-455).

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, *5*(4), 1-19.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and Naive bayes. In *Advances in neural information processing systems* (pp. 841-848).

Pieters, M., & Wiering, M. (2017, November). Comparison of machine learning techniques for multi-label genre classification. In *Benelux Conference on Artificial Intelligence* (pp. 131-144). Springer, Cham.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of Naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 616-623).

Soria, D., Garibaldi, J. M., Biganzoli, E., & Ellis, I. O. (2008, December). A comparison of three different methods for classification of breast cancer data. In *2008 Seventh International Conference on Machine Learning and Applications* (pp. 619-624). IEEE.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., & Liu, Y. (2014). Investigation of multilayer perceptron and class imbalance problems for credit rating. *International Journal of Computer and Information Technology*, *3*(4), 805-812.