COMP90049 Introduction to Machine Learning
Project 2 Report

Anonymous

1 Introduction

In this day and age, music has become an integral part of everyone's life. There are now multiple sources on the internet where people can listen to music. This involves searching for them, downloading them or save them in a music playlist. However, searching for music always involves searching for music genres and the songs categorised by them. These genres are often labeled manually by humans with the help of meta tags such as ID3 tags. This task is often tedious (Zhen and Xu, 2010; Silla Jr. et al., 2008).

One solution to the manual approach is to utilise machine learning techniques to automatise the music genre classification process.

In this project, different machine learning classifiers are applied to different feature sets of a given training data set (Bertin-Mahieux et al., 2011; Er Schindler, 2012). The classifiers are then evaluated on a validation set from which the best classifier is applied to an unseen data set. The hypothesis of this project is that Lyrical(tags) features are the most predictive for predicting a music genre from a give data set.

2 Literature Review

Over the years multiple methods have been proposed by researchers. These methods incorporate different machine learning algorithms applied on different types of music data set containing different feature sets (audio, metadata, lyrics). Various results have been presented to support this field of study. In one study, Tzanetakis and Cook (2002) explored this field by applying classification on audio signals. Feature sets containing timbral texture, rhythmic content and pitch content were extracted from real world audio collections. After extraction, performance was measured by applying statistical pattern recognition classifiers on these sets of data. In another another study, Silla Jr. et al. (2008) further explored music signals with a newer approach applying pattern recognition ensemble classification over multiple features based on space and time decomposition schemes. Results show that ensemble approach always produced better quantitative results. In a parallel study, Kumar et al. (2018) tackled this problem by performing classification on lyrical texts to classify music genres. Word embedding techniques such as Word2Vec and TFIDF were used on this approach for processing lyrics before applying various machine learners algorithms. In another genre classification approach, classification was performed on available social tag data such as music-tag and artist-tag (Zhen and Xu, 2010).

3 Experiment

All the researches conducted previously by different researchers were on different feature sets. However, with a data set in our experiment containing a combination of three different types of feature for each song, it becomes necessary to evaluate the best set of features producing the best predictions.

3.1 Feature Selection & Extraction

3.1.1 Lyrical data

In the given data set, the tags contain a list of words for each song representing the lyrics. These tags are human annotated. To perform feature extraction, the Term Frequency Inverse Document Frequency (TFIDF) vectorisation was performed (Kumar et al., 2018). This strategy created a bag of words which were more specific to each song and helped to reduce the number of vectors for the tag features. Words that were of high frequency were filtered away and the resulted vector values were scaled between [0,1] using MinMaxScaler. To keep the array dimensions of the training and validation sets the same, the validation tag features were vectorised using the bag of words obtained from the training set.

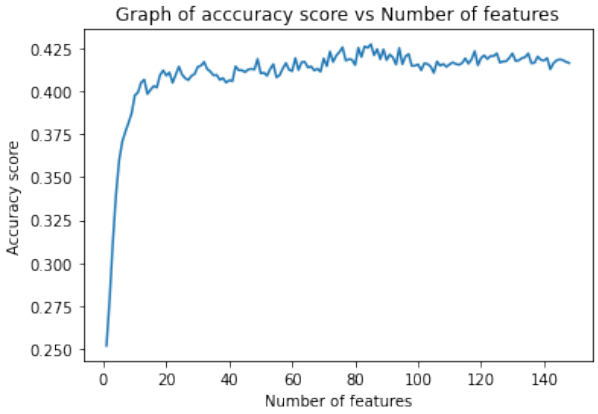

Figure 1: Graph of accuracy score vs number of features for audio feature set

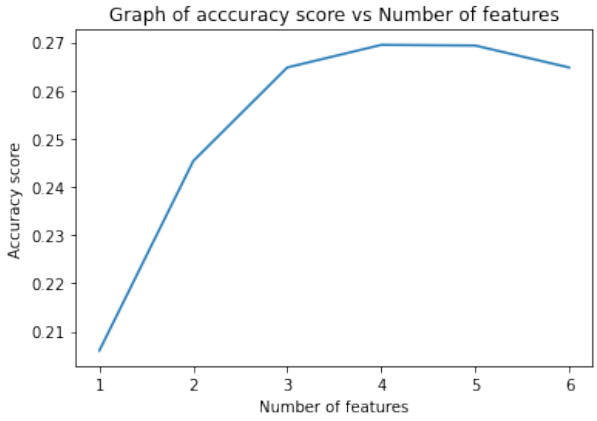

Figure 2: Graph of accuracy score vs number of features for metadata feature set

3.1.2 Audio data

To gather audio data, the 148 vectors of continuous, pre-computed audio features were processed differently. These features are pre-extracted in the data set and the values are not interpretable. Since these features are already in numerical form, the values were directly scaled between [0,1] using MinMaxScaler to remove negative values. After this, the best features were selected from 148 vectors using Recursive Feature Elimination with cross validation (RFECV). Default settings were kept for the selection process. From the graph the highest cross validation score value was produced by 85 audio feature vectors and thus were selected for the training of the classifiers (see Figure 1). Similar to the lyrical data set, the validation audio feature set was also transformed using the training feature set in order to keep similar array dimensions.

3.1.3 Metadata

The metadata features contain the title, loudness, tempo, key, mode, duration and time signature. Since the title is in textual form, it was first converted into a bag of words just like the lyrical feature set using TFIDF vectorisation and then scaled to between [0,1]. The other metadata features were also converted to scale [0,1] and 4 best features were selected through RFECV with default settings (see Figure 2). Finally after, processing of the data, all the metadata features were combined as a single metadata feature set. To keep the array dimensions of the training and validation title feature sets same, the validation title feature set was vectorised using the bag of words obtained from the training set. The remaining metadata features were also transformed by the dimensions of the training metadata features during the RFECV processing to keep the array dimensions similar (see Table 1).

Feature set	Number of Features
Audio	85
Lyrical(tags)	4827
Metadata	6353

Table 1: Number of features for each feature set

3.2 Learners Used

For the experiment, Decision Tree (DT) and Multilayer Perceptron (MLP) classifiers were selected for the learning process. Decision Tree was selected as it utilises tree traversal algorithm providing better performance. MLP was selected as it provides optimal results when no clear rules are available to solve a given problem (Ray, 2019). Moreover, Silla Jr. et al. (2008), and Tzanetakis and Cook (2002) have widely used MLP for music genre classification for audio signals. Zero-R classifier was also used as our baseline to compare how well our algorithms performed with the baseline values.

3.2.1 Hyper Parameter Tuning

Since the aim of the experiment was to find the most predictive set of features for Music genre classification, we fine tuned only few parameters for each learner (see Table 2). To obtain the best set of hyper parameters, we fine tuned the algorithms using GridSearchCV. GridSearchCV, tests all possible combinations of the hyper parameters selected for tuning. For Decision Tree, GridSearchCV was only tasked

to test combinations between 'gini' and 'entropy' criterion. Others were left as default values to allow the learner to scale till the max depth. For Multilayer perceptron, hyper parameter tuning was performed for 2 parameters, namely, the activation function and the solver. 'relu' and 'identity' options were removed from the parameter tuning for activation function as they utilised a linear model. The number of hidden layers were kept to 1 and the max iter was set to 2000 to allow the classifier to train for a longer period of time. Audio feature sets from the training features were used along with training labels for the tuning of the hyper parameters. Once tuned, the classifiers were re-trained with different feature sets before validating the learners with validation feature sets. The learners were not re-tuned for the remaining types of feature sets (lyrical and metadata) using GridSearchCV. Rather, they were directly used for training with the parameters obtained from the turning of audio feature sets.

Learner	Hyper parameters
Multilayer perceptron	activation='logistic' solver='adam' maxiter=2000
Decision Tree	criterion='entropy'

Table 2: Learner hyper parameters after tuning

4 Evaluation

Based on the results obtained (see Table 3), the Multilayer perceptron performed with a higher accuracy score across all the types of feature sets in our experiment. Moreover, it obtained a better accuracy score for the lyrical feature set than the audio feature set. On the other hand, it did not perform as well for the metadata feature set.
Decision Tree, performed slightly better for the audio feature set compared to the lyrical feature set. However, similar to Multilayer perceptron, it did not perform well for the metadata feature set.
Overall, both the learners performed better than the Zero-R baseline classifier.

The class distribution shows significant variations in the label predictions. For all types of feature sets, the Multilayer perceptron and the Zero-R predicted most songs as Classic Pop and Rock genre. Moreover, it predicted the least songs as Jazz and Blues genre for validation audio, lyrical and metadata feature sets.
The Decision Tree showed similar predictions except for the audio feature set where it predicted most songs as Folk genre rather than Classic Pop and Rock.

Learner	Accuracy score
Multilayer perceptron	0.455556
Decision Tree	0.304444
Zero-R	0.126667

Table 3: Accuracy scores obtained for audio feature set

Learner	Accuracy score
Multilayer perceptron	0.515556
Decision Tree	0.384444
Zero-R	0.131111

Table 4: Accuracy scores obtained for lyrical(tags) feature set

Learner	Accuracy score
Multilayer perceptron	0.311111
Decision Tree	0.233333
Zero-R	0.142222

Table 5: Accuracy scores obtained for metadata feature set

5 Contextualising Method behaviour

From the results above, it becomes evident that Multilayer perceptron is a better learner obtaining an accuracy score of 0.515556. This could be due to the utilisation of stochastic gradient descent as the solver for weight optimisation. This learner is fault tolerant and have the ability to handle noise(Ray, 2019). It tries to reduce error by changing its weighted score over time.
Decision Tree shows the next best performance when compared with all the learners used in the experiment. It had a faster computation during the tuning process using GridSearchCV. This could be due to the efficiency of the tree traversal algorithm (Ray, 2019).

6 Error Analysis

Overall, both the learners produced a locally optimal solution rather than a global one even though they performed better than the baseline. Upon closer inspection of the plots,the pattern becomes quite obvious. Both the learners show

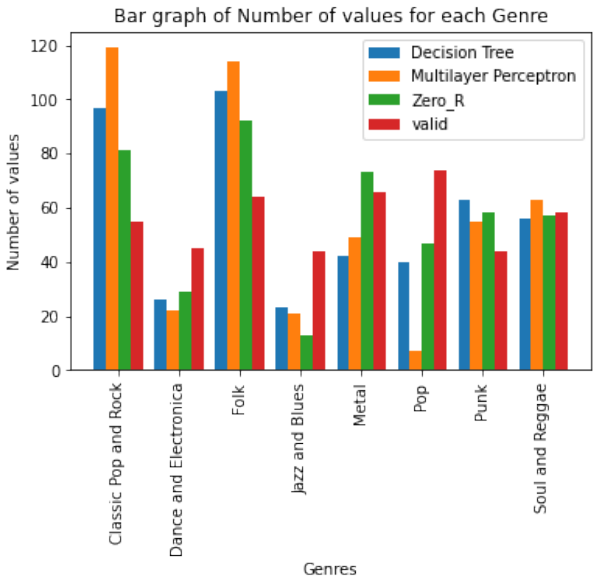

Figure 3: Bar graph showing number of genre for each class for audio feature set

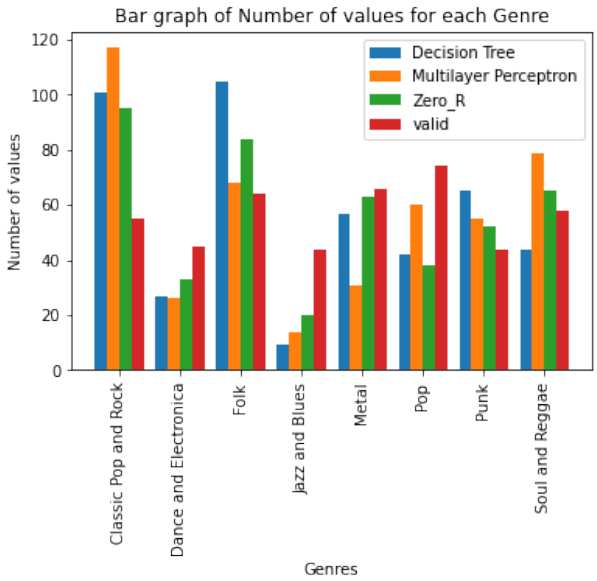

Figure 5: Bar graph showing number of genre for each class for metadata feature set

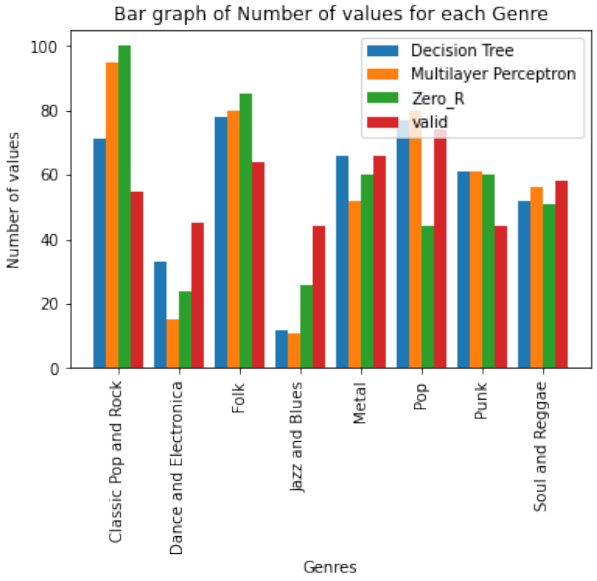

Figure 4: Bar graph showing number of genre for each class for lyrical (tags) feature set

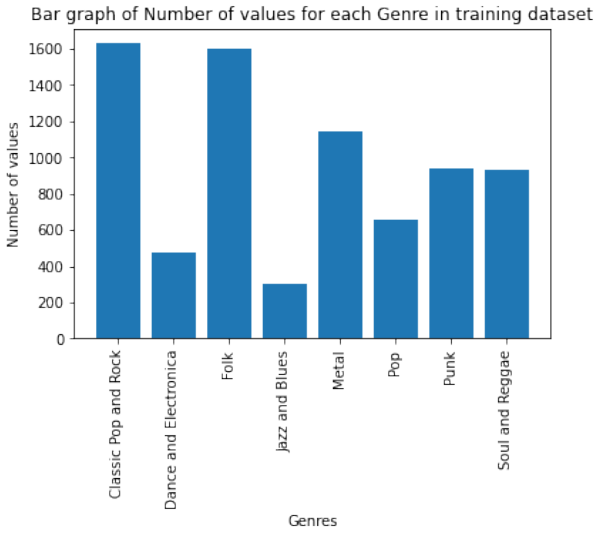

Figure 6: Bar graph showing number of genre for each class in the training data set

similar performance by predicting similar genres for each type of feature set. This shows signs of bias existing in the learners as they were only tuned using the audio features of the training set. Moreover, certain steps in the experimental process could also have been the factors introducing bias in the results. All the textual data from the validation set had been transformed into a frequency vector based on the frequency of words from the training set. Since the training frequency scores were used by TFIDF, this introduced scaling that was designed for the training set. As a result, words that may have little importance in the training set would have obtained a better score leading to misleading results. Also, feature selected from the RFECV method were also based on the training data set values. This could also have selected features that showed stronger correlation in the training set but not in the validation set.

For both the training and the validation set, there exists an uneven distribution of songs for each class label(see Figure 3 - 5). The uneven

distribution could also affect results as it does not provide enough instances for each class labels that could have allowed the learners to generalise. Values in the data sets could also be a reason for a biased performance. For instance, in the data, there are many songs with audio vectors equal to 0. Moreover, some of the values in a feature appear anomalous when compared to the entire feature column. These values could be human errors when taking readings for song samples.

7 Conclusion and future work

In this report, different feature sets from a given data set were extracted and classified with 3 different machine learning classifiers. Accuracy score was used to observe the performance of the learners on each type of feature set and the genre predictions were compared for analyse. Multilayer perceptron showed better performance using the lyrical feature set. Therefore, our hypothesis is proven that lyrical data set serves as a better predictor for classifying music genres. More work such as enhanced feature selection and hyper parameter tuning of the algorithms would be required for further analysis in the future.

References

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Andreas Rauber Er Schindler. 2012. Capturing the temporal domain in echonest features for improved classification effectiveness. In *In Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR 2012)*.

A. Kumar, A. Rajpal, and D. Rathore. 2018. Genre classification using word embeddings and deep learning. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2142–2146.

S. Ray. 2019. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39.

Carlos N. Silla Jr., Alessandro L. Koerich, and Celso A. A. Kaestner. 2008. A machine learning approach to automatic music genre classification. *Journal of the Brazilian Computer Society*, 14:7 – 18, 09.

G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

C. Zhen and J. Xu. 2010. Solely tag-based music genre classification. In *2010 International Conference on Web Information Systems and Mining*, volume 1, pages 20–24.