A comparative study on Music Genre Classification using Audio, Metadata, and Lyrical Features

Sharodh K R

1. Introduction

Music Genre Classification is an essential component of the Music Information Retrieval (MIR) system and has been drawing a substantial amount of attention with the advent of online music streaming services. The enormous amount of online music library has led to automatic music genre classification which enables users to stream stations of a particular genre. Conventional music genres are classified based on various parameters like region, instrument, style, and so on. The delineation of the genre is often vastly subjective and tends to overlap in certain instances. The key objective of this research paper is to analyse the performance of Audio, Metadata, and Lyrical features individually on different Machine Learning models to classify the genre of music.

2. Related work

Music genre prediction is a multi-class classification problem. In the past several attempts have been made to classify music based on content-based acoustic data which includes pitch, beat, and tonality, text-based features obtained from lyrics of the music, and also the metadata of the music which includes information like title, tempo, composer and so on. In 2011, Sander Dieleman, Philemon Brakel, and Benjamin Schrauwen (Sander Dieleman & Schrauwen, 2011) employed a convolutional neural network on audio features over musically significant timescales and observed an increase in accuracy for artist recognition, genre recognition and key detection. Similarly, 2003, Tao Li, Mitsunori Ogihara, and Qi Li (Tao Li & Qi, 2003) proposed a new feature extraction technique, DWCHs where global and local information of the music signals are captured using Daubechies wavelet coefficients. The resulting feature is compared using various machine learning techniques including Linear Discriminant Analysis and SVM. A significant increase in the accuracy has been observed on using DWCHs. Although features have been analysed individually using machine learning techniques, there are no conclusions arrived on the best feature to predict the genre of music.

3. Dataset

The dataset is a sample sourced from Million Song Dataset (MSD) (Bertin-Mahieux, et al., 2012) (Schindler & Rauber, 2014) which is collection of information on one million contemporary popular music tracks. The dataset comprises audio, lyrical and metadata features split across training, validation, and test instances. A sum of 7678 instances is present in training dataset which is used to train several machine learning models. The evaluation metrics are analysed using 450 instances present in the validation dataset. Finally, the best performing feature in conjunction with the best performing model is chosen to predict 428 instances of test dataset whose labels are masked. The performance of the model on the test dataset is ranked on Kaggle contest leaderbord[1].

The metadata features consist of, loudness, duration, key, mode, title, tempo, and time signature. Audio features comprise of 148 pre-calculated vectors that were pre-extracted from the 30- or 60-seconds snippets of every track, 'Mel Frequency Cepstral Coefficients' (MFCC), chroma, and capture timbre properties of audio. All the audio feature values are non-interpretable and continuous. The lyrical feature comprises of tags that occurred in the song lyrics. These tags are human-annotated and stemmed to reduce inflected word. Each song is classified with single genre label out of 8 possible genres which include 'Soul and Reggae', 'Punk', 'Pop', 'Dance and Electronica', 'Jazz and Blues', 'Classic Pop and Rock', 'Folk', and 'Metal'.

[1] https://www.kaggle.com/c/comp90049p2s2/

4. Proposed Methodology

After examining the dataset, the focus of the research paper is to build individual classifiers for Lyrical, Audio and Metadata features.

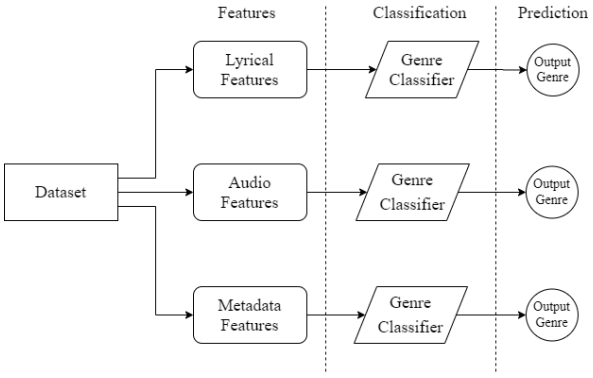

Figure 1- *Architecture diagram of the methodology followed*

Owing to the massive amount of tags present in lyrics, feature selection using information like PMI, Chi-Square, and MI, has been performed to select 300 best features out of all tags. Similarly, the title feature has been analysed and removed from metadata as it is less useful in predicting the genre of the music. Each

category of the feature was evaluated across different machine learning models to predict the genre of music.

4.1 Data Pre-processing

Feature scaling is critical to eliminate bias generated due to range and standard deviation. Standardization is a technique used to scale the values with unit standard deviation. In gradient descent-based algorithm where the range is more sensitive to performance, standardization aids to scale the values of the dataset in the same range. Thus "StandardScaler" from Scikit Learn is chosen to scale for logistic regression and multi-layer perceptron as it is both based on gradient descent. (Pedregosa, et al., 2825–2830).

The lyrical features consist of a string which contains multiple commas separated pre-stemmed tags extracted from lyrics. These tags are transformed into numerical values using Scikit Learn "MultiLabelBinarizer" module. An N-dimensional vector of binary data is created for each tag present in input instance which denotes the presence of a tag in the instance. This results in the creation of a substantial amount of feature vectors which tends to be useless in predicting the genre of the music. The Scikit package "SelectKBest" is then employed to select the best set of 300 (arrived from multiple trails) features to predict the genre.

5. Implementation

The inherent implementation of Naïve-Bayes, Logistic Regression, and Multi-Layer Perceptron (MLP) from Scikit Learn library is utilized to implement the model. The python code attached to this research paper has the elaborate implementation of specifics. The upcoming section demonstrates the summary of the strategy performed to construct the respective machine learning models for music genre prediction.

5.1 Naïve Bayes

A generative probabilistic technique that estimates the best feasible probabilistic label for an instance as the one that increases the likelihood of seeing the label times the product of the likelihood of observing every individual feature value provided that label.

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Figure 2- *Naïve Bayes Formulation*

The Naïve Bayes classifier reduces the complexity of calculation by presuming the features of conditional independence. This technique functions well even in instances where the underlying assumption does not seem to hold well. Thus, the Naïve Bayes model is a respectable opening model for analysis (Silla, et al., 2008).

The metadata and audio features are continuous ranging between 0 and 1 after pre-processing. Thus, a Gaussian Naïve Bayes classification algorithm was suitable to create a model. On the other hand, binarized tags being discrete values was directly fed to Multinomial Naïve Bayes to predict class labels.

5.2 Logistic Regression

A discriminative model which computes the conditional distribution $P(y|x)$ eradicating the assumption of conditional independence. (Vilar, et al., 2004)

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Figure 3- *Logistic Regression Formulation*

A multinomial logistic regression model was performed for audio, tags and metadata features separately. As the features were normalised to lie between in the range of 0 and 1, the number of iterations required was substantially reduced from 1000 to 100 to achieve the same performance. After exhaustive search over different solver methods from Scikit Learn library, "lbfgs" is chosen as the best solver which maximises the performance.

5.3 Multi-Layer Perceptron

Multi-layer perceptron is a non-linear classification algorithm which can learn features on their own as transitional formulation. It has multiple layers of computational units which are interconnected with weights.

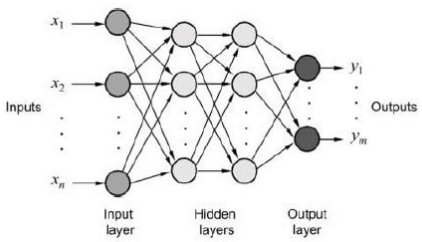

Figure 4- *Multi-layer Perceptron Schematic*

The neural networks use a backpropagation technique to compute the error from the output layer and transfer it back to the previous layers to implement the changes in the weights. The process is repeated until the satisfactory convergence of correct output predictions (Sander Dieleman & Schrauwen, 2011).

The MLP classifier has been implemented with an optimal number of hidden layers to balance both computation and performance. "SGD" solver is found to have the best performance with a learning rate of 0.01 running for a max iteration of 1000. The algorithm is used to predict the genre of the music for audio, lyrical and metadata features individually.

6. Results & Evaluation

Using Audio, Metadata, and lyrical features, three machine learning models namely Naïve Bayes, Logistic Regression, and MLP were constructed and their performance measures have been tabulated below in the tables 2, 3, and 4, respectively. A 0-R classifier which predicts the most frequent class labels as the genre of the music is also used as a baseline metric to set the benchmark.

	Accuracy	Precision	Recall	F1
Baseline	0.12	0.01	0.12	0.02

***Table 1**- Evaluation metrics of 0-R classifier*

	Accuracy	Precision	Recall	F1
Audio	0.49	0.48	0.46	0.46
Lyrical	0.53	0.44	0.47	0.43
Metadata	0.26	0.20	0.26	0.20

***Table 2**- Evaluation metrics of Naïve Bayes Classifier*

	Accuracy	Precision	Recall	F1
Audio	0.49	0.57	0.49	0.49
Lyrical	0.49	0.43	0.45	0.43
Metadata	0.30	0.21	0.29	0.22

***Table 3**- Evaluation metrics of Logistic Regression*

	Accuracy	Precision	Recall	F1
Audio	0.41	0.63	0.48	0.48
Lyrical	0.28	0.24	0.28	0.22
Metadata	0.27	0.42	0.26	0.22

***Table 4**- Evaluation metrics of MLP*

After examining the training dataset, it is evident that there is an imbalance in the count (table 5) of class labels.

Genre	Frequency
Classic pop and rock	1629
Folk	1601
Metal	1143
Punk	937
Soul and reggae	930
Pop	657
Dance and electronica	478
Jazz and blues	303

***Table 5**- Label distribution of train data*

As the performance of the model is dependent on training dataset distribution of classes, the imbalance causes a lack of information about certain classes. It is exemplified in the confusion matrix of the Naïve Bayes classifier where most of the genres predicted are the most frequently class labels and least predicted genres are least frequently occurring class labels.

The evaluation metrics are analysed using accuracy, precision, recall and F-Score, calculated by predicting 450 validation instances using the model created with 7678 training instances. Precision signifies the accuracy of predicting the right genre. Whereas recall signifies the capability to correctly identify the right genre. F1 score is calculated using the harmonic mean of precision and recall.

From the experiments conducted, Naïve Bayes classifier has the best accuracy which is obtained using lyrical features. However, Precision and Recall are better for a logistic regression model using audio features. As a result of it, the Logistic Regression model using audio features produces the best F-score of 0.49 across three models. Since precision for audio and tag features are substantially more compared with metadata, they can be considered as best predictors.

Generally, F1 scores tend to be used as a key metric when handling imbalanced datasets (Quan, et al., 2015). Thus, the result of the experimentation is that audio features have the best chance of predicting the correct genre of music, irrespective of the machine learning model used.

8. Error Analysis

Error analysis has been performed by plotting accuracy over different training sizes. The gap between the training and cross-validation score can be used a metric to predict whether the model is overfitting or underfitting, Figure 5 shows the learning curve of Naïve Bayes classifier with negligible gap compared to figure 6 which represents the learning

curve of Logistic Regression model. On the other hand, MLP has a substantial gap which indicates overfitting of data. This reiterates that the logistic regression model has better performance with minimal overfitting.

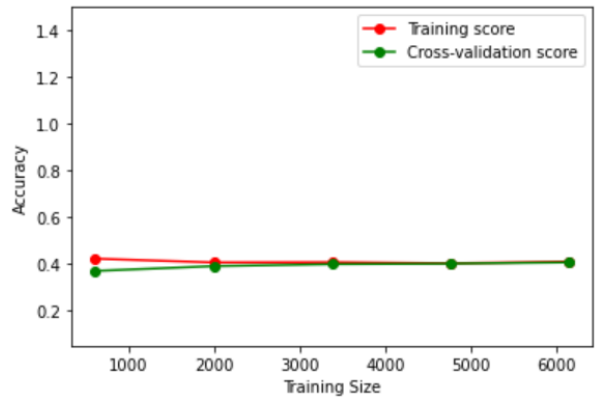

***Figure 5**- Naive Bayes - Learning Curve*

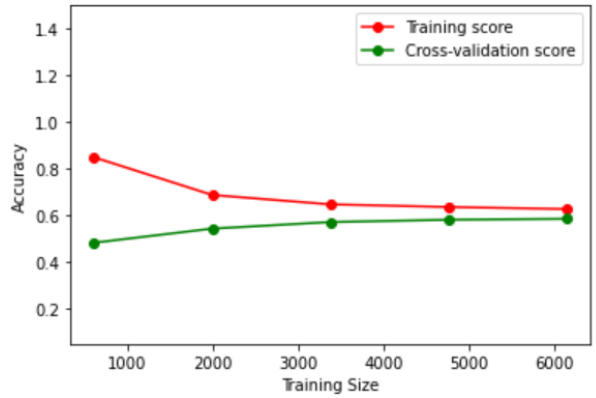

***Figure 6**- Logistic Regression - Learning Curve*

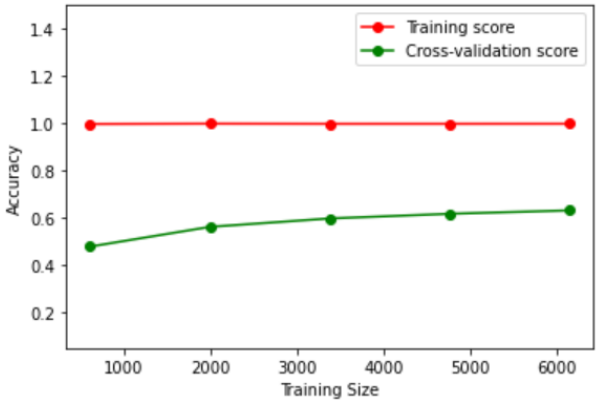

***Figure 7**- Multi-Layer Perceptron Learning Curve*

8. Conclusions

The evaluation results confirm that audio (Oramas, et al., 2017)features were able to best predict the genre of the music. Future work on this research could be using done using the combination of features with models which are more immune to the imbalance in the dataset. Another extension to this research is to remove the imbalance in the data using one of the many popular techniques like oversampling, under-sampling and so on to improve the accuracy of music genre prediction.

References

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B. & Lamere, P., 2012. *The Million Song Dataset.* Miami, Florida, University of Miami.

Oramas, S., Nieto, O., Barbieri, F. & Serra, X., 2017. *Multi-label Music Genre Classification from Audio, Text, and Images Using Deep Features.* s.l., Proceedings of the 18th International Society of Music Information Retrieval Conference (ISMIR 2017).

Pedregosa, Fabian & Varoquaux, 2825–2830. *Scikit-learn: Machine learning in Python.* s.l., Journal of machine learning research.

Quan, Z. et al., 2015. *Finding the Best Classification Threshold in Imbalanced Classification.* s.l., Elsevier, pp. 2-8.

Sander Dieleman, P. B. & Schrauwen, B., 2011. *Audio-based music classification with a pretrained convolutional network.* Miami, FL, USA, University of Miami, pp. 669 - 674.

Schindler, A., Lidy, T. & Rauber, A., 2016. *Comparing Shallow versus Deep Neural Network Architectures for Automatic Music Genre Classification.* s.l., FMT, pp. 17-21.

Schindler, A. & Rauber, A., 2014. *Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness.* s.l., Springer, Cham.

Silla, C. N., Koerich, A. L. & Kaestner, C. A. A., 2008. *A Machine Learning Approach to Automatic Music Genre Classification.* s.l., s.n., p. 7–18.

Tao Li, M. O. & Qi, L., 2003. *A comparative study on content-based music genre classification.* Toronto, Canada, Association for Computing Machinery, pp. 282-289.

Vilar, D., José, M. & Sanchis, C., 2004. *Multi-label Text Classification Using Multinomial Models.* s.l., Springer, Berlin, Heidelberg, pp. 220-230.