Guess The Genre

Atalay Gürel [1]
Muhammed Fidan [1]

Abstract

We could not even imagine a life without music. Throughout history wherever there were people, there was music too. Empires fell and rose, revolutions were made and humankind has come to these days.The music was with them for all these moments. Humankind took shape with music but also the music took shape with humankind. Music has evolved for generations and generations and found its form in many ways. Today we use the term "genre" for different categories of music. The goal of this paper is to give machines a chance to predict music genres given input features from music tracks. To do that, we apply various techniques based on machine learning on the dataset called FMA which consists of 161 subgenres among 106,574 tracks.

1. Introduction

As a part of being human, we all have different tastes for any kind of hobby. There is even a popular expression for this: "Tastes and colors are indisputable". This also includes our music tastes; the difference may come from the person's upbringing, reliefs as well as the environment and era in which the person grew up.

What we aim in this project is to successfully make a Music Genre Recognition when we have input as an audio file by utilizing Machine Learning algorithms. Popular music streaming services and platforms are using these algorithms to understand the preferences of their audience and provide them better recommendations so that the audience will enjoy their music more and spend time on these mediums more.

In this project we tried to implement and discuss baseline classification models to solve the problem of music genre classification. These models are:

- K-Nearest Neighbor Classifier
- Logistic Regression through the one-vs-rest
- Support Vector Machines

To represent the audio tracks we planned to use Mel Frequency Cepstral Coefficients(MFCC) and Spectral Contrast features to begin with, which have been shown to be effective in the task of predicting genres.

2. Related Work

For the music genre classification one of the most common datasets are FMA: A Dataset For Music Analysis (Defferrard & Bresson, 2017), and GTZAN (Tzanetakis & Cook, July 2002.). However FMA is the one more up to date and fresh dataset.

Kim et al (Jaehun Kim & Liem, 2018) used the same FMA Medium dataset but took a different approach to the genre recognition problem. This work claimed that genre labels are noisy, subjective, and not clearly separable from one another. Genre labels may be inter-correlated, and a generally accepted taxonomy of musical genres does not exist. Given this claim, the authors asserted that artist labelings do not suffer from the same faults that genre labelings do. Furthermore, they contended that since artists tend to stay within one or several related genres, musical characteristics of the artists may be indicative of certain musical genres

Automatic classification of audio has also a long history originating from speech recognition. Mel-frequency cepstral coefficients (Davis & Mermelstein, Aug. 1980), are a set of perceptually motivated features that have been widely used in speech recognition. They provide a compact representation of the spectral envelope, such that most of the signal energy is concentrated in the first coefficients.

In the academic study titled "A comparative study on content-based music genre classification", DWCH(Daubechies Wavelet Coefficients Histogram) (T. Li & Li, 2003) which is a new method of extracting attributes on music genre classification, is mentioned. This method is based on feature extraction by calculating the Daubechies Wavelength Coefficients of the signals. Using the Support Vector Machine, K-NN, Gaussian Mixing

[1]Department of Computer Engineering, University of Hacettepe, Ankara, Turkey. Correspondence to: Muhammed Fidan <muhammed.fidan77@gmail.com>, Atalay Gürel <atagrl98@gmail.com>.

BBM406 Fundamentals of Machine Learning, Spring 2021.

Model and Linear Discriminant Analysis algorithms, the success percentage of this new method compared to the previous methods was compared.

3. Dataset Analyze

The FMA dataset, a dump of the Free Music Archieve, includes 106,574 tracks with 161 sub-genres. In this task, we use 38,990 of the tracks with 15 top-genres sampled considering their metadata and popularity for computational efficiency. Our data includes clips of 30s and an unbalanced distribution among genres that differ from 24 to 14,182 clips per top-genre (figure 1).

In FMA dataset, the features are generated using librosa, and stored as statics, including kurtosis, max, min, mean, median, std and skew, for each feature. For pitch feature, chroma representations are a preferred way to encode harmony, and suppressing perturbations in octave height, loudness, or timbre. Chroma features could be extracted in different ways, e.g. by convolving a fixed size window Short Time Fourier Transform (chroma stft), or a variable sized window, constant-Q transform (chroma cqt), over the audio signal to extract the time-varying frequency spectrum

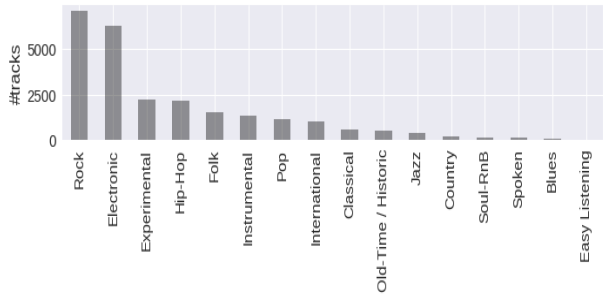

Figure 1. Top Genres

We split our data preserving the percentage of tracks per genre as a reflection of population (stratified sampling) into training, validation and test by 80/10/10%. Thus, our training data turned into a matrix of 6400 rows and 140 columns.

4. The Approach

4.1. NEAREST NEIGHBORS CLASSIFIER

KNN is a non parametric algorithm (meaning, it does not make any underlying assumptions about the distribution of data)belonging to supervised learning community. KNN algorithm can also be used for regression problems.The only difference will be using averages of nearest neighbors rather than voting from nearest neighbors. Although kNN is an easy implemented algorithm and performs well in a large number of classification problems, it suffers from the curse of dimensionality. Our model has a dimension space of 140 features which makes kNN a bit vulnerable.

4.2. LOGISTIC REGRESSION

Logistic regression is a statistical machine learning algorithm that classifies the data by considering outcome variables on extreme ends and tries makes a logarithmic line that distinguishes between them. It is a go-to method for binary classification problems.

We used multinomial logistic regression. It uses for three or more categories without ordering. Linear regression uses mean squared error as its cost function. But in multiclass classification problems, it would be more logical to use the maximum likelihood cost function. Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

4.3. SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

We tried to preprocess the dataset before applying Linear SVC using standard scaler to represent standard normally distributed data. Beside from Logistic Regression, SVM uses one-against-one approach for multi-class classification. This method is consistent, which is not true for one-vs-rest classification.

5. Experiment and Conclusion

In our models, we tried to approach our results with three classification method. These are; Nearest Neighbors(kNN), Logistic Regression and Support Vector Machine (SVM) with linear kernel. We split our dataset into train and test subsets. In our train set there are 6400 clips meanwhile test set included 800 clips.

We used these subsets for all our models, all the methods we used, have different kind difficulties while capturing the non-linearities of the data, for this reason we achieve less accuracy than we expected.

The following tables show the accuracy performance obtained with all the features and Mel Frequency Cepstral Coefficients (MFCC) alone

Table 1. Classification methods for MFCC Only.

MODELS	TRAIN ACC.(%)	TEST ACC(%)
KNN	57.8	36.1
LOGISTIC REG.	56.0	42.1
SVM LINEAR	75.8	46.3

Our kNN and Logistic Regression models are outperformed by SVM regardless of the chosen features. The reason behind of the issue is that Nearest Neighbors algorithm treats vectors as an input which makes this method work uneffi-ciently in complex dimensionals while also logistic regression does not try to find the best margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data. Instead it can have different decision boundaries with different weights that are near the optimal point.

Support vector machine with linear kernel performing a classification with a linear decision boundary. So when facing an input set of high dimensions, we can reach to a conclusion that we can improve our decision boundary by using a radial basis kernel in order to perform a well non-linear decision boundary.

As an addition to the MFCC feature set of 140 dimensions used in Table 1, Despite the increase in dimension, all of our methods except for kNN, we achieved better results in both train and test accuracy. Therefore, with better feature and model selection, improved accuracy results can be achieved.

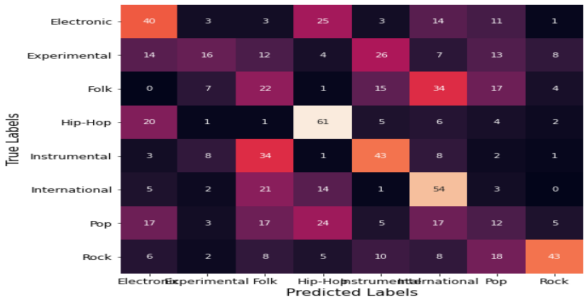

Figure 2. kNN Confusion Matrix

We run the model with different k values to find the best k value of KNN. After trials between 1-25, we observed that the most optimal k value is 13. We have seen that the most wrongly predicted music genres in KNN are experimental and pop on the confusion matrix. Meanwhile International and Hip hop music genres were predicted more accurately than other music genres.

Our SVM model found clips in hiphop, rock, and electronic genres highly accurate. However, as in KNN and logistic regression, SVM also made unsuccessful predictions in pop

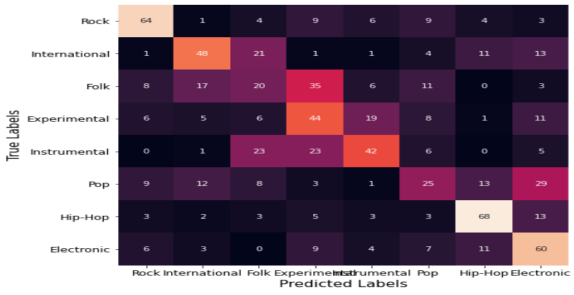

Figure 3. SVM Confusion Matrix

and folk genres. Music such as hip hop, rock tend to have very different sounds such as auto-tuning and therefore the model is easier to distinguish. The same style of instruments is commonly used in pop and experimental music, mostly guitar. This causes the model to mix them easily, resulting in worse performance with both types.

Due to unbalanced data we have, we were expecting our confusion matrixes will not be healthy. Because we figured out that we were trying to preserve the percentage of the population while doing that we also misspredicted most of the minority genres. As a future work we might consider creating genre-specific models for minority genres.

References

Davis, S. and Mermelstein, P. Experiments insyllable-based recognition of continuous speech. 28:357–366, Aug. 1980.

Defferrard, M., B. K. V. P. and Bresson. Fma: A dataset for music analysis. *18th International Society for Music Information Retrieval Conference*, 2017. URL https://arxiv.org/abs/1612.01840.

Jaehun Kim, Minz Won, X. S. and Liem, C. Transfer learning of artist group factors to musical genre classification. *Companion Proceedings of the The Web Conference 2018*, pp. 1929–1934, 2018.

T. Li, M. O. and Li, Q. A comparative study on content-based music genre classification. *SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 282–289, 2003.

Tzanetakis, G. and Cook. P. musical genre classification of audio signals. volume 10:293–302, July 2002. doi: 10.1109/TSA.2002.800560.