Music Genre Classification

*

Nithyashree Senguttuvan
011808333, Machine Learning
School of Electrical Engineering and Computer Science
Voiland College of Engineering and Architecture
Washington State University
Pullman, United States of America
n.senguttuvan@wsu.edu

***Abstract*—Music is a go to remedy or tool which most of the people prefer to perform any of their desired activities. The general definition that could answer the question ”what?” is that music is the mix of vocal or instrumental sounds combined to produce a form or express emotion. In the modern era, people listen to all forms of music which nowadays have fusion of one or more forms of music. With various forms of being present already and evolving in time are hard to recognize when there is a geographical or modernised change in our course of living. We find people moving all around the world finding some indigenous music hard to grasp and categorize while getting to know the history and origin. It is hard to listen to something at a instant and categorize them, when we may not know all genre or may have never heard of them.This report is based on the classification system that has been developed from understanding the features of 10 genres in music, comparing with the accuracy of classification methods, classifying the GTZAN genre classification dataset and testing the same with a new input. From the results of accuracy, the classification is done using K nearest neighbours for the dataset.**

***Index Terms*—learning, model, genre, data**

I. Introduction

Machine Learning has grown and is continuing to grow to be solution for all kinds of need. Though the thought process of implementation of it in various questioning applications, it has shown to provide results and progress of improvements through models for its accuracy in prediction. The application here that is ”Music” , is a sense of art and in some defines culture in a way throughout the world. Its course of study has been an interesting and a thought provoking one on its journey because the means of communication at ease has been its beat of heart for centuries now. It is composed and performed for many purposes, ranging from aesthetic pleasure, religious or ceremonial purposes, or as an entertainment product for the marketplace.

Music has many different fundamentals or elements which depends on the definition of ”element” being used. They can include pitch, beat or pulse, tempo, rhythm, melody, harmony, texture, style, allocation of voices, timbre or color, dynamics, expression, articulation, form, and structure. Understanding these elements can aid in its application of cognitive science, as it has been proved that it can investigate human aptitude, skill, intelligence, creativity, and social behavior.

Music therapy is an interpersonal process in which a trained therapist uses music and all of its facets—physical, emotional, mental, social, aesthetic, and spiritual—to help clients to improve or maintain their health where the client’s needs are addressed directly through music. Conditions like psychiatric disorders, medical problems, physical disabilities, sensory impairments, developmental disabilities, substance abuse issues, communication disorders, interpersonal problems, and aging is being taken care of using music. It can improve learning, build self-esteem, reduce stress, support physical exercise, and facilitate a host of other health-related activities.

With the various modes to apply music and learn, there can be a difficulty in equipping with the right kind of music that is essential to perform these activities. With the growing knowledge and evolution of various forms of music adding to its density around the world, we might feel lost. To overcome this, we must derive an efficient mechanism to classify and make use of it application in the field of need. In the approach of paving the path to it, we find that the data to be collected, the songs or tunes, will be different format and range. We also find it hard to note the features that can efficiently differentiate and classify. Thus, we analyze the genre forms by visualization of graphs and methods to come to a meaningful feature to derive results.

The approach explained here is the data collection and exploration, Data preprocessing, knowing the features, performing PCA, comparison with different models, making the best model learn to classify the genre and testing the model with a new input. This method has been found to be best by the comparison on simpler models and is being used because of the computational feasibility for the application. The accuracy achieved has been satisfying and when tested on a new input for classification the model has accurately predicted the genre.

II. Problem Setup

A. Social wellness

Advancement of mobility makes people travel everywhere to learn, explore and experience various things of the world. Language and music are the two things you pass through inevitably during travel. Thus, many people maybe confused and cannot identify the genre or the type of music when they listen to something outright in that instance. The model explained can be made into an mobile application where one could know the song based on their genre.

B. Education

People in the field of musicology, Music theory, Zoomusicology or Ethnomusicology may find it difficult to classify new types of music which can be a fusion or a new genre of its type by its feature defining it. Thus, to clearly aid in the fields of academia to identify the features to describe, study and classify, the model may come in hand.

C. Psychological Studies

As discussed earlier, the medical field can open our parts of brain, to unknown discoveries through music. The model can help in choosing the right form of key to open the door to mysteries to unexplored parts of human psychology and origin to such sense and advancement.

III. Solution Approach

A. Understanding Data

We have known data to be in a structured table format , text or images and sometimes video. The mere idea of music as data was interesting and analyzing it for a model is in a way different from our usual method of study and knowledge. The format types of audio file are of three types - Uncompressed audio formats,Lossless compression and Lossy compression. There are like 35 to 40 file extensions available within these categories. Our data of study is in ".wav" format which is a standard audio file container format used mainly in Windows PCs. This extension is commonly used for storing uncompressed format, CD-quality sound files, which means that they can be large in size around 10 MB per minute. Wave files can also contain data encoded with a variety of lossy codecs to reduce the file size. For example the GSM or MP3 formats.

B. Pre-processing the Data

Not all data is clean and can be dumped for the model to train as it can vastly affect its credibility in learning the classification. There can be lot of aspects in the data making it untidy. In the case of music, it can be noise in between, or no audio in the data or the size of the data. Thus the data is preprocessed by getting rid of the things that can probably affect the learning of features to classify and making all of the data into a standard format that can be fed.

C. Feature Extraction

After making it a standard form of data, next is to find the points that could make the genre similar and different, such that the new file when uploaded can be classified accurately after the model learning it. The genres are visualized using LibROSA, which is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. when we talk about sound, we generally talk about a sequence of vibrations in varying pressure strengths, so to visualize sound kinda means to visualize airwaves. The basic preprocessing techniques are implied to see the small differences starting with Fourier Transform(FT).

$$Audio = amplitude(varying) + frequncies(varying) \quad (1)$$

in time domain. The plots visualized does not describe about the frequencies. Thus, to convert samples to amplitude vs frequency, we use FT. Discrete Fourier Transform(DFT) is performed as we know the values at different 't'. Short Term Fourier Transform (STFT) is done in place of it as is perform DFT in a faster way.

Using the methods we cannot see much as most sounds humans hear are concentrated in very small frequency and amplitude ranges, we develop spectrogram using Mel Spectrogram. The Mel Scale, is the result of some non-linear transformation of the frequency scale. This is constructed such that sounds of equal distance from each other on the Mel Scale, also "sound" to humans as they are equal in distance from one another.

D. Comparison

Using the spectral centroids, Spectral rolloff- for a frame, frequency below which a specified amount of spectral energy lies, Mel-frequency cepstral coefficients(MFCC) and Principal Component Analysis(PCA) the data is prepared for being the input with the features. The supervised learning models considered are Naive Bayes, Stochastic Gradient Descent, KNN, Decision trees, Random Forest, SVM, Logistic Regression, Neural Networks, Cross Gradient Booster, Cross Gradient Booster (Random Forest).

1) Naive Bayes: Naive Bayes is based on the Bayes Theorem. It is one of the simplest yet powerful algorithms in use and finds applications in many industries. To solve a classification problem and have created the features and generated the hypothesis, the best naive solution for this situation would be to use the Naive Bayes classifier, which is quite faster in comparison to other classification algorithms. It assumes that all predictors are independent and is suitable for solving multi-class prediction problems. If its assumption of the independence of features holds true, it can perform better than other models and requires much less training data. It is generally better suited for categorical input variables than numerical variables. The demerit is that all predictors are independent, which is not possible in real life. This limits the applicability of this algorithm in real-world use cases. Also, the algorithm faces the 'zero-frequency problem' where it assigns zero probability to a categorical variable whose category in the test data set was not available in the training dataset. It would be best if a smoothing technique is used to overcome this issue.

2) Stochastic Gradient Descent : Stochastic Gradient Descent (SGD) is a simple and very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features. SGD is merely an optimization technique and does not correspond to a specific family of machine learning models. It is only a way to train a model. The disadvantage is that SGD requires a number of hyperparameters such as the regularization parameter and the number of iterations and it is sensitive to feature scaling.

3) K-Nearest Neighbour: K-Nearest Neighbour(KNN) algorithm works by assuming that similar things exist close to each other. It utilises feature similarity between the new data points and the points in the training set to predict the values of the new data points. In essence, the K-NN algorithm assigns a value to the latest data point based on how closely it resembles the points in the training set. K-NN algorithm finds application in both classification and regression problems but is mainly used for classification problems. It is considered as a lazy learning algorithm because instead of learning from the training set immediately, the K-NN algorithm stores the dataset and trains from the dataset at the time of classification. The algorithm also a non-parametric, meaning it does not make any assumptions about the underlying data.

4) Decision Trees: Decision trees (DTs) are a way to vividly establish a chronological decision process. It contains decision nodes, each with branches for each of the alternative decisions. The random variables also appear in the tree, with the efficacy of each branch computed at the leaf of each branch. The expected efficacy of any decision can then be calculated based on the weighted addition of all branches from the decision to all leaves from that branch.

5) Random Forest: Random Forest is an ensemble method to discover the decision tree that best fits the training data by creating many decision trees and then determining the ”average” one. The ”random” part of the term refers to building each of the decision trees from a random selection of features; the ”forest” refers to the set of decision trees.The ”forest” built is an ensemble of decision trees, normally trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. It is a very accessible algorithm because the default hyperparameters it uses often yield a good prediction result. The demerit of random forest is that a great number of trees can make the algorithm too slow and unproductive for real-time predictions. In general, these algorithms are fast to train, but slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model. In most practical applications, the random forest algorithm is fast enough but there can certainly be situations where run-time performance is important and other approaches would be preferred

6) Support Vector Machine: Support Vector Machine(SVM) finds a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. The SVM kernel is a function that takes low dimensional input space and transforms it into higher-dimensional space, that is it converts non separable problem to separable problem. It is mostly useful in non-linear separation problems. In the kernel, it does some extremely complex data transformations then finds out the process to separate the data based on the labels or outputs defined. Its memory efficient as it uses a subset of training points in the decision function called support vectors and different kernel functions can be specified for the decision functions and its possible to specify custom kernels.

7) Logistic Regression: Logistic Regression(LR) goal is to estimate the posterior probability using continuous function and predict using discrete categorical value. This is an arithmetic technique like linear regression since it finds an equation that predicts the result of a twofold variable, Y, from one or more response variables,X. The response variables can be categorical or continuous, as the model does not necessarily require continuous data. To predict group association, the log odds ratio is used rather than probabilities and an iterative maximum likelihood technique is used rather than least squares to fit the model. Thus it can be more appropriate for non-normally distributed data or when the samples have unequal covariance matrices.

8) Neural Networks: Neural network can be understood as a network of hidden layers, an input layer and an output layer that tries to mimic the working of a human brain. The hidden layers can be visualized as an abstract representation of the input data itself. These layers help the neural network understand various features of the data with the help of its own internal logic. These neural networks are non-interpretable models. Non-interpretable models are those which cannot be interpreted or understood even if we observe the hidden layers. This is because the neural networks have an internal logic working on its own, that cannot be comprehended by us. The output of a neural network is a numerical vector, that bridges the gap between the actual data and the representation of the data by the network. An output layer can be understood as a translator that helps us to understand the logic of the network and convert the target values. A theorem named ‘Universal approximation theorem’ tells that a feed forward network that contains one hidden layer can be used to represent any function.

A neural network is a mathematical model that helps in processing information. The information is processed in the simplest form over basic elements known as ‘neurons’. Neurons are connected and help exchange signals/information between them with the help of connection links. This connection links between neurons could be strong or weak, and this strength of the connection links determines the method in which information is processed. Every neuron has an internal state which can be determined by the incoming connections from other neurons and has an activation function which is calculated on its state, and this helps determine its output signal. It computational graph of mathematical operations.

9) Cross Gradient Booster: Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. It is also known as gradient tree boosting, stochastic gradient boosting (an extension), and gradient boosting machines, or GBM for short. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "gradient boosting," as the loss gradient is minimized as the model is fit, much like a neural network.

E. Learning and Testing

Based on the comparison in the simplest methods, K-Nearest Neighbour is chosen based on their results and computational efficiency. The dataset is split into train and test data to get it trained and tested with a new audio file for its genre. As the audio signals are constantly changing, first we divide these signals into smaller frames. Each frame is around 20-40 ms long Then we try to identify different frequencies present in each frame Now, separate linguistic frequencies from the noise to discard the noise, it then takes discrete cosine transform (DCT) of these frequencies. Using DCT we keep only a specific sequence of frequencies that have a high probability of information.

F. Justification

The model that has been chosen is solely based on using the model assess function. Though XGBoost is found to have a greater accuracy, the choice was KNN due its generalised applicability for the given case and its results on it.

G. Drawbacks

The accuracy of the actual model has found to be lower than the one when used in comparison with the test set. This may differ if we had used XGBoost instead of KNN.

IV. Experiments and Results

A. Experimental Setup

1) Data: GTZAN genre classification dataset provide the data required to train the model with 10 classes - 'blues', 'classical', 'country', 'disco', 'hiphop', 'jazz', 'metal', 'pop', 'reggae', 'rock'. Each class has about 100 songs that can be used to train the model.

2) Features: The features that has been obtained based on the visualization are taken to classify they are STFT values, RMS, Spectral Centroid, Spectral Bandwidth, Spectral rolloff, Zero crossing rate, Harmony, Tempo and MFCC.

3) Base Learners: Identifying the neighbours in the actual model act as the base learners to implement the classification. The non-parametric supervised learning method whose output is a class membership for classification where an object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

B. Evaluation Methodology

1) Performance metrics: The functions of Scikit learn which is accuracy score() , which returns the accuracy classification score is used for the comparison part. It calculates the classification accuracy. And as with the KNN it is mathematically calculated with the test set.

2) Experiments: The experiment is carried out in the general way, where the dataset is divided into train and test and the evaluation based on the test data results with the actual value.

3) Hyper-parameters: The hyper parameters in our approach has been set manually, to give a general view of the training process as the different models has already been compared beforehand based on the mentioned features.

C. Baseline Approach

The approach initially was aimed to carry out was implement a model onto the data and find its accuracy. But after learning the features and its process, the exploration was to compare different models on the data and develop a complete application based model that can be incorporated into anything to train and test.

Musical sounds can comprise a wide range of sound components with different acoustic qualities. In particular, we consider two broad categories of sounds: harmonic and percussive sounds. A harmonic sound is what we perceive as pitched sound, what makes us hear melodies and chords which is the acoustic realization of a sinusoid, which corresponds to a horizontal line in a spectrogram representation. Example: The sound of a violin. Most of the observed structures in the spectrogram are of horizontal nature, even though they are intermingled with noise-like components.

Thus, for a percussive sound is what we perceive as a clash, a knock, a clap, or a click. Example: The sound of a drum stroke or a transient that occurs in the attack phase of a musical tone. It is the acoustic realization of an impulse, which corresponds to a vertical line in a spectrogram representation.

D. Results

With the features , the following comparison results have been obtained that has led to choose the KNN classifier for the discussed dataset.

The model of KNN that has been trained and tested has a accuracy of 70.6% and the testing on a new input has been verified to be accurate. Thought the actual model differs in accuracy, the method of how it is implemented differs that when taken into account makes up the decrease.

The takeaway is that the aspects of how a data in spite of its various formats or forms can be inspected for its features

TABLE I
ACCURACY FOR DIFFERENT MODELS

Model	***Accuracy***
Naive Bayes	0.52386
Stochastic Gradient Descent	0.64298
KNN	0.8038
Decission trees	0.65065
Random Forest	0.79479
Support Vector Machine	0.74274
Logistic Regression	0.67434
Neural Nets	0.69703
XGboost	0.89656
XGboost Random Forest	0.74708
XGboost1	0.8979

and implemented to be a active model that can be used in various applications is far more worth in the process of skills and time required to develop this. The learning put together is that machine learning though is considered for every real world application, it is what we make it with the use of data and how we define it to be. In the above, if XGBoost has been implemented to train and test, the whole process may have enhanced in accuracy. Thus, we make the learning what we require it to be.

V. Conclusions and Future Work

Supervised learning based approach has been taken for the dataset and the model for its high accuracy. The model can be changed and in order to improved, semi-supervised or unsupervised learning approach can be taken to verify its credibility of accuracy and for future unexplored applications.

Acknowledgment

The inspiration was the different kinds of music which was not known to many of the students due to the transition for studies to different countries to be culturally acknowledged. Also, from the Data Science class that had helped in making the possibility of such music genre classification visible.

References

[1] G. Tzanetakis and P. Cook, ”Musical genre classification of audio signals,” in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002.

[2] Panagakis, Yannis & Benetos, Emmanouil & Kotropoulos, C. Music Genre Classification: A Multilinear Approach.. 583-588, 2008.

[3] A. Pooransingh and D. Dhoray, ”Similarity Analysis of Modern Genre Music Based on Billboard Hits,” in IEEE Access, vol. 9, pp. 144916-144926, 2021.

[4] K. Markov and T. Matsui, ”Music Genre and Emotion Recognition Using Gaussian Processes,” in IEEE Access, vol. 2, pp. 688-697, 2014.

[5] R. Yang, L. Feng, H. Wang, J. Yao and S. Luo, ”Parallel Recurrent Convolutional Neural Networks-Based Music Genre Classification Method for Mobile Devices,” in IEEE Access, vol. 8, pp. 19629-19637, 2020.