International Journal of Research in Engineering, Science and Management
Volume 4, Issue 7, July 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Musical Sentiment Recognition from Row Audio using Various Machine Learning Methods

Siddhartha Chaki[1*], Shivkumar Goel[2]
[1]Student, Department of MCA, Vivekanand Education Society's Institute of Technology, Mumbai, India
[2]Professor, Department of MCA, Vivekanand Education Society's Institute of Technology, Mumbai, India

***Abstract*: The main objective of this paper is to present a Machine Learning model method to classify/tag any music based on its musical sentiment which it represents throughout each phase of time on it. In this paper, we try to map the emotions/moods with music using various machine learning methods. In preprocessing step its extracts various features like chroma frequencies, tonnetz, Mel-Frequency Cepstral Coefficients, root-mean-square energy, zero-crossing rate, spectral features collectively and their statistical calculations like mean, variance, min, max etc. At the final stage we apply various techniques to develop the best approach to solve this problem using extracted features.**

***Keywords*: Audio emotions/sentiment classification, Extreme Gradient Boosting, K-NN, Machine learning, Music information retrieval (MIS).**

1. Introduction

As the database of music increases on a day-to-day basis with each continent contributing with their style of the flow of music. With huge quantities, there are ways to classify them in new ways to make them manageable and more enjoyable to listeners with more and more custom and user-targeted selections. To accomplish this task, various music information retrieval (MIR) methods have developed over time, which focused on research and development of the system which targets new ways to the information retrieval from music, for better understanding of the music.

Classification of music based on emotions/sentiments is one of the most interesting research topics among music researchers, as it directly maps music to the human psychological state. Here we try to map the music to predefined emotional categories which include happy, sad, romantic, aggressive, and dramatic.

However, the classification of music based on sentiment is not an unsupervised task as it needs human intervention first to classify the music beforehand, which makes it more subjective towards understanding the music to the human first, and then to the mechanical systems.

One of the hardest parts in any music related problems which the outcome depends heavily on is the preprocessing step as one raw music data constrains about 44100 values per second which makes it a very large number of parameters to handle. Any 5-sec music will have 2,20,500 input dimensions which is too much to handle at a time. As the solutions many approaches have been developed which use statistical methods to reduce the input dimension while keeping music features intact. Those methods include feature extraction like tempo, timber, frequency-time features, harmonics, beats, spectrograms, and many more. After preprocessing steps the further steps are straightforward with many models like adaboost [6], convolutional neural network [2], recurrent neural network [7], etc. which have developed and provide promising results in many of the music-related problems.

2. Literature Survey

Classification of music based on emotions or moods is not a very new task, throughout the time many researches related to various music information retrieval approaches have been done to accomplish this task, this problem is a bit similar to music genre classification which uses the same set of features and methods.

Two approaches can be considered while solving the problem in the preprocessing step 1. Utilize various Music Information Retrieval (MIS) methods which focus more on extract statistics information based on timber, peach, frequencies, harmonies etc. [1], [5], [6].

Using the row music itself in which the row music data is in the form of row music wave values or spectrograms [2], [3].

In current solutions use both of the methods in different problems like music generation, genre classification, AI-music orchestrated music, etc. were in the case of music genre classification which closest/similar problem to emotions classification, are using the music information retrieval (MIS) features which is more usable as its faster using the lightweight machine learning models [5]. Although the deep learning models which use raw audio or spectrograms for classification also provide comparative results [2], [4].

Other than the processing part there are many models which are used for classification in the field of music; those include

*Corresponding author: siddharthachaki02@gmail.com minimum distance and K-nearest neighbor [9] AdaBoost [6], Support Vector Machine [8], and many more.

3. Proposed Models

The solution developed in three main general steps as preprocessing step or feature extraction, feature aggregation, and finally the classification.

As the primary step, the audio files are loaded into the system which in general has 5 seconds of tracks each. The tracks processed with 2048 sample rate at 22050 Hz data points for each track using librosa and the statistical features are extracted from it onwards. The features are listed below.

A. Chrome-based audio-features

Chrome features are related to twelve pitch classes, which divides the music into predefined categories based on their closest match with equal-tempered values on their own scales. Those features capture the melodies and harmonic characteristics of music with relation to changes in timber and instrumentations.

- *Chrome stft:* Represent chromagram from a short-time fourier transform.
- *Chrome cqt:* Represent chromagram from constant-Q transform.
- *Chrome cens:* Represent chroma variants “Chrome Energy Normalized” features.

Tonnetz: It represents the chrome features in a 6-D basis consisting of the perfect fifth, minor third, and major third each with two dimensions.

Mel-Frequency Cepstral Coefficient (MFCC): The Mel-frequency Cepstral (MFC) represents the short-term power spectrum of the sound and the Mel-Frequency Cepstral Coefficient is the coefficient that collectively represents the MFC. Generally, there are 10-20 MFCC’s which elaborate the entire shape of the spectral envelope.

Root-mean-square (RMS): this value represents RMS value each time frame, either from the audio samples y or a spectrogram S.)

Zero Crossing Rate: It represents the rate of significant changes along with the signal (the rate in which it changes from negative to positive or opposite).

Spectral Features (Centroid, Contrast, Bandwidth, Roll-Off): those features represent the statistical values of sound depend on the power of frequency where spectral centroid indicates the center, spectral contrast is the spectral peak and valley and their differences, roll-off is the frequencies which above the certain energy levels [11].

The aggregation part includes the features calculation and collection for all the segments in the whole audio in each seconds interval. Then the calculation of its mean, standard deviation, skewness, kurtosis, median, min, and max f which form the input vector for the model.

After extracting and processing different statistical values of the above features, we apply different machine learning techniques to check which works best for the solution. The algorithms which all are going to be used are Gaussian Naive Bayes (GaussianNB), regularized linear classifier with Stochastic Gradient Descent (SGD), K-nearest neighbor, Decision Tree, Random Forest with the number of child trees set to 1000 and max_depth set to 10, Support Vector Machine Classifier with RBF kernel with C set to 1.0 and gamma set to 1/(n_features* X.var()), Multi-layer Perceptron classifier which uses neural network hidden layers as (5000,5), alpha as 0.000001,lbfgs(Limited-memory Broyden–Fletcher–Goldfarb–Shanno) as weight optimizer, eXtreme Gradient Boosting with learning rate 0.05.

4. Experimental Results

The database used in the Music_Classification dataset from kaggle [12]. Where music is structured in mood-specific folders related to mood/emotions. After feature extraction, the resultant input vector size is 518, which corresponds to each feature. For the training, the dataset split into two parts with 70%(7090) for training and 30% (3039) for testing.

Table 1

Algorithm	**Accuracy**
Naive Bayes	0.37808
Stochastic Gradient Descent	0.59296
K-nearest neighbor	0.74729
Decision Tree	0.5742
Support Vector Machine	0.65548
Neural nets	0.54228
eXtreme Gradient Boosting	0.71866

Parameters of K-Nearest Neighbours Algorithm: leaf_size: 30, metric: Minkowski, n_neighbor: 19.

As of the algorithm comparisons the eXtreme Gradient Boosting and K-NN algorithm outperform others in the test dataset after getting trained. As for the K-NN parameter tuning options are a little limited so we try to optimize the XGBoosting algorithm to achieve decent accuracy. After a bit of trial on different parameter searches on hyper-parameters space using the GridSearch algorithm, XGBoost achieved approximately 75% accuracy which varies in some cases.

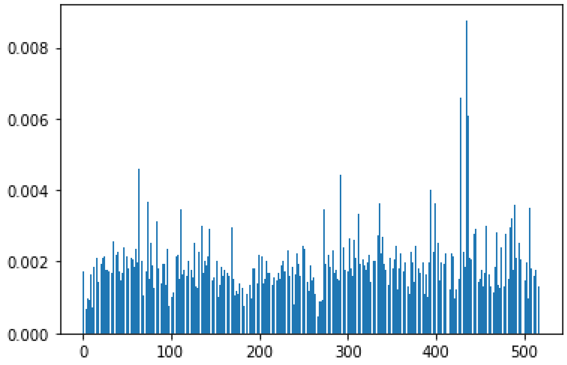

Fig. 1. Feature importance of all 518 parameters on the basis of XGBoost results

Current XGBoost Model Parameters:
colsample_bytree: 0.7456899658336619,
gamma: 3.1704577809079244,
max_depth: 5.0,

min_child_weight: 4.0,
reg_alpha': 4.0,
reg_lambda: 0.29676133204573224

5. Conclusion

Based on the above models the eXtreme Gradient boosting algorithm, which is robust, lightweight, and decent accuracy where it takes care of overfitting and less prone to noise which makes it best for the solution in the field of music with so many different types with every new day more and more music added to the list.

In this paper, one of the solution challenges is the data, as the data which is used in this approach is small in terms of variations which can be filled by manually selecting and increasing the dataset. Overall with the availability of the data and keeping the performance and other features of XGboost in mind this model works up to the mark although not greatly accurate. This algorithm with the availability of more data and power to search more parameters with grid search in more combinations values can provide more accurate results.

Day by day the Music classification algorithms are gaining a new level of maturity where they can utilize more Music information systems, newly developer features and new more powerful models with the capability of handling more data with more hand labeling of data and can verify large labeled collections automatically. However, when the performance and resources trade-offs come to be useful in a commercial setting, the requirement for algorithms to run quickly and be able to learn from datasets that are larger than the ones dealt with in these experiments.

References

[1] Eidenberger, Horst. “Fundamental Media Understanding”, 2011.
[2] Lee et al. “Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms”, 2017.
[3] Marolt et al., “Neural networks for note onset detection in piano music”, 2002, ICMC.
[4] Pons et al, “End-to-end learning for music audio tagging at scale”, 2018 ISMIR.
[5] C. N. Silla, A. L. Koerich, and C. A. A. Kaestner, “A machine learning approach to automatic music genre classification,” Journal of the Brazilian Computer Society, 2008
[6] J. Bergstra; N. Casagrande; D. Erhan; D. Eck; B. Kégl., “Aggregate features and ADABOOST for music classification. Machine Learning”,
[7] Choi, K., Fazekas, G., Sandler, M., Cho, “Convolutional recurrent neural networks for music classification”, 2017 ICASSP.
[8] Mandel, M. I., & Ellis, D. P, “Song-level features and support vector machines for music classification”, 2005.
[9] Beth Logan and Ariel Salomon “A Music Similarity Function Based On Signal Analysis”, 2001.
[10] Junqua, J., & Haton, J, “Robustness in Automatic Speech Recognition”. (1996)
[11] https://musicinformationretrieval.com
[12] https://www.kaggle.com/shanmukh05/music-classification