Genre classification of Music files

Puneet Girdhar
Department of Computer Science
The University of Texas at Dallas
Richardson, TX 75080
Email: pxg151330@utdallas.edu

Dhruvkumar Patel
Department of Computer Science
The University of Texas at Dallas
Richardson, TX 75080
Email: drp150030@utdallas.edu

***Abstract*—Music classification is an interesting problem with many applications. Music streaming services like Spotify, Apple Music and Google Play offer music catalog categorized into different genres and moods. It isn't clear however whether this process of music categorization is automated or manual. This feature is also not very commonly available for offline music files. The reason is that music genres are hard to describe systematically due to their subjective nature. In this project, we have applied audio signal processing and machine learning techniques to classify music files in different genres. We received comparable results to recent work in music genre classification task.**

Keywords- Music Genre Classification, MIR, Music Information Retrieval

I. Introduction

For our project, we have used GTZAN genre collection dataset, which is available as part of Marsyas audio processing framework. G. Tzanetakis and P. Cook curated this dataset in their work [1] on music genre classification. Even though they did not intend for this dataset to be standard dataset for work on music genre classification task [3], its easy availability on web has made it the primary dataset used by many [2] [3] [4]. This dataset consists of 1000 audio tracks, each of 30 seconds duration. These tracks are spanned over ten different genres namely: Classical, Blues, Country, Hip Hop, Disco, Jazz, Metal, Popular, Reggae, and Rock.

We investigate various machine learning algorithms on GTZAN dataset in this project, including K-nearest neighbours(KNN), K-means, multiclass SVM, and neural networks. For feature extraction task, we have relied on timbre-based features, namely Mel Frequency Cepstral Coefficients(MFCCs). This has been recommended by previous work on this problem [5]. There are many other audio features that can be used like power spectrum of the signal, FFT of audio signal, but it has been found that MFCCs represent given music file's timbre attributes, which is the most important aspect for genre classification problem. We have verified this by getting comparable accuracy values on given dataset as original work [1]. We have also included other features such as Zero Crossing Rate, Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Rolloff in different feature extractor plugins.

II. Feature Extraction

We investigate several machine learning algorithms on variety of features. We extracted different features from each audio file like Zero Crossing Rate, Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Rolloff, Chroma Vector etc. Then we save these features to a database, so we can use them in different combinations for different machine learning algorithms. These different combinations of features are 'Plugins'. There are 7 such plugins, namely AF,FEXTRACT,GMM,MFCC,MDELTA,RHYTHM,CHROMA. Analysis of each of these features and plugins is described below.

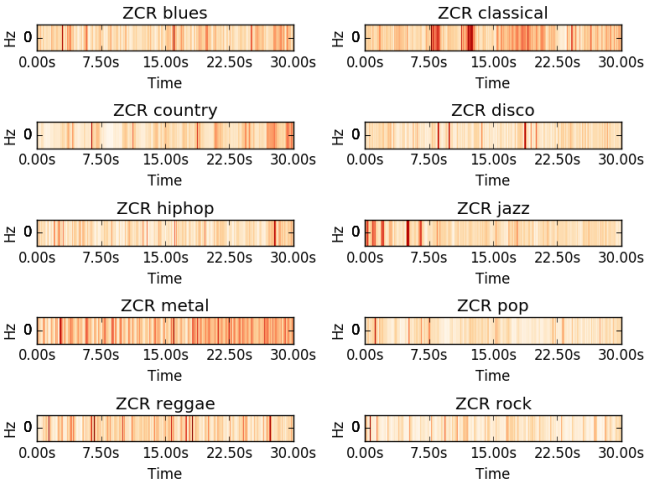
A. Zero Crossing Rate

Zero-crossings is the number of zero crossings of the signal in the time domain. It reflects the noisiness of the signal. Periodic sounds tend to have a small value of zero crossings while noisy sounds usually have a high value. The 'zero-crossing rate' is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. [**?**]

ZCR is formally defined as:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} f(s_t s_{t-1} < 0) \tag{1}$$

where f(t) is 1 if t is true.
Different genre represented as zero crossing rate vectors are represented in given figure.



We can clearly see that different genres have different zero crossing rate distributions.

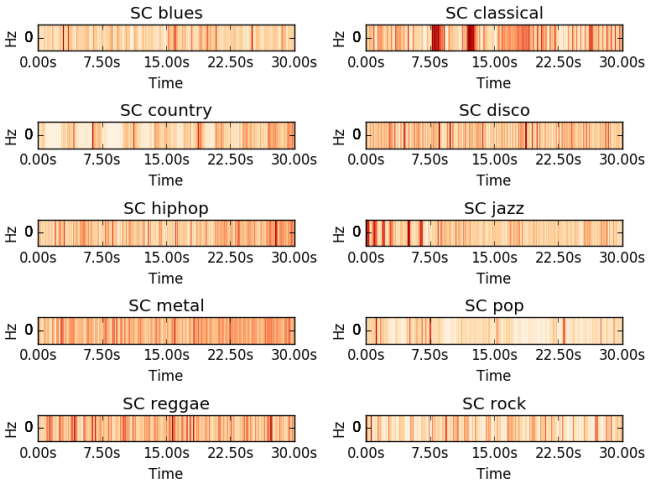
B. Spectral Centroid

The spectral centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the ”center of mass” of the spectrum is. Perceptually, it has a robust connection with the impression of ”brightness” of a sound. [**?**]

It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights:

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{2}$$

Spectral Centroid distribution for different genres is given in following figure.

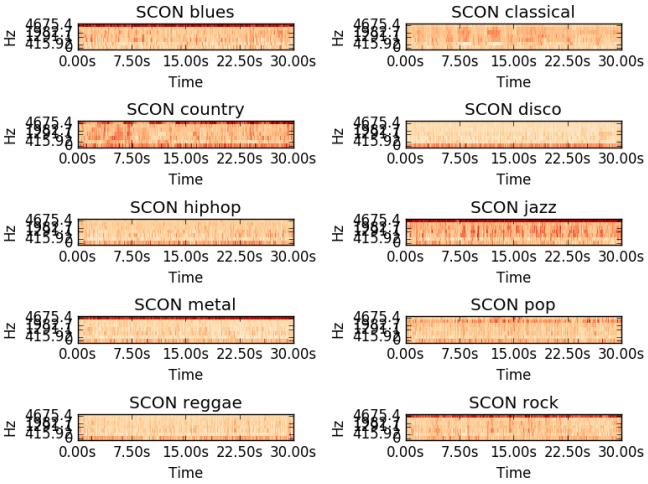


C. Spectral Bandwidth

Spectral Bandwidth is the Wavelength interval in which a radiated spectral quantity is not less than half its maximum value.

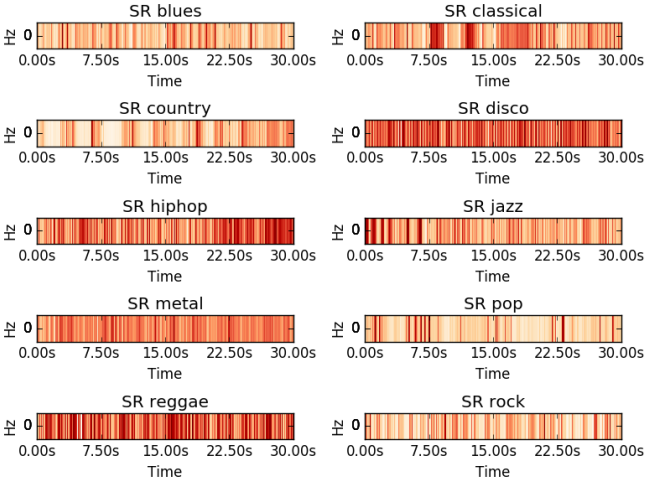
D. Spectral Contrast

Spectral contrast is defined as the decibel difference between peaks and valleys in the spectrum [**?**]. Spectral contrast distribution for different genres is given in following figure.



E. Spectral Rolloff

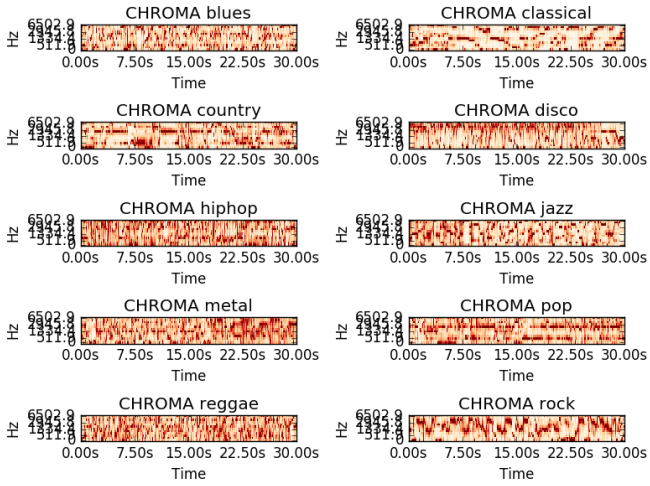
Spectral Rolloff is a measure of the amount of the right-skewedness of the power spectrum. The spectral rolloff point is the fraction of bins in the power spectrum at which 85% of the power is at lower frequencies [**?**]. Spectral Rolloff distribution for different genres is given in following figure.



F. Chroma Vector

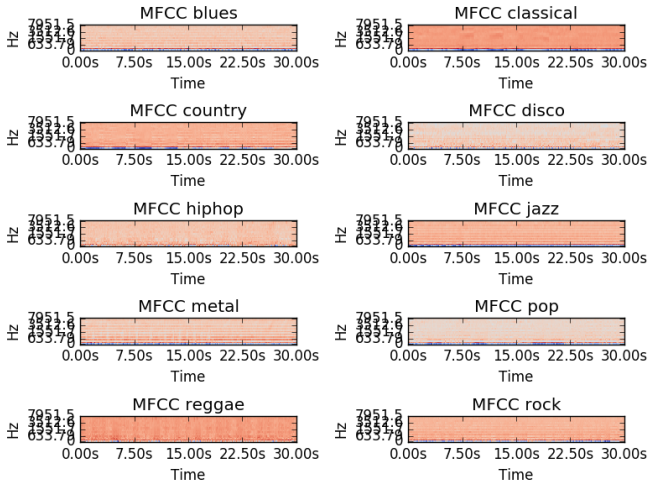
Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave. The term chroma closely relates to the twelve different pitch classes. Harmonic pitch class profiles (HPCP) is a group of features that a computer program extracts from an audio signal, based on a pitch class profilea descriptor proposed in the context of a chord recognition system. Often, the twelve pitch spelling attributes are also referred to as chroma and the HPCP features are closely related to what is called chroma features or chromagrams [**?**]. Chroma Vector distribution for different genres is given in following figure.

This feature looks more promising than others in images.



G. MFCC

MFCC, Mel-Frequency Cepstral Co-efficients, a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency [**?**]. MFCC vector distribution for different genres is given in following figure.



III. Feature Combinations- Plugins

IV. Data preprocessing pipleline

Librosa(python package to process music) is an open source software framework for python for music data analysis. It also provides basic building blocks necessary to create music retrieval systems. We downloaded GTZAN genre classification dataset popularly used for musical research purpose for our project. It contains 1000 audio tracks each 30 seconds long. There are 10 genres represented each containing 100 tracks. All tracks are 22050 Hz Mono 16-bit audio files in .au format.

We wrote a python script which reads audio files of 100 songs per genre , extract all features and saves in a sqlite database. Once feature vectors are stored, multiple models can be trained using combination of feature vectors. We further introduced custom matrix representation for each song which represents music feature mean vectors and covariances of those features as 1-D vector, effectively modelling features as Multivariate Guassian distribution. Lastly, we applied different supervised learning algorithms using the reduced mean vector and covariance matrix as the features for each song to train on.

** picture of song broken into feature vector

Given a piece of music, each section of song(20 msec) is transformed into a vector with N dimensions and hence complete song can be transfomed into a feature matrix of N rows and M columns where M = number of frames.

V. Classification

Once the feature vectors are obtained, we train different classifiers on diffrent set of feature vectors. We divided the dataset in 67% - 33% ratio for training and testing respectively. Feature selection is done manually by hit and trial method. Following are the different classifiers used:

1) Multi Label K Nearest neighbour
2) One-Vs-All approach with SVM
3) Neural Network
4) Non-Linear SVM
5) Logistic Regression
6) Ensembling

A. Multi Label K Nearest Neighbour

Intutively, Music belongs to one genre should share same audio features. If we consider audio featuers as N dimensional vector then music belongs to one genre should be close to each other. This is similar to the idea behind multi-Label KNN algorith in which it utilizes audio feature information of a music's k-nearest neighbour to infer its genres. We used above KL divergence measure (described below) for calculating similarity between two songs.

Consider $p(x)$ and $q(x)$ to be two song multivariate guassian distributions obtained from feature extraction. Then we have the following:

$$2KL(p||q) = \log\left(\frac{\sum_q}{\sum_p}\right) + Tr\left(\sum_q^{-1}\sum_p\right)$$

$$+(\mu_p - \mu_q)^T \sum_q^{-1}(\mu_p - \mu_q) - d$$

Since, KL divergence is not symmetric but the distance should be symmetric, we have:

$$D_{KL} = KL\left(p||q\right) + KL\left(q||p\right)$$

Here we experimented with multiple values of K and found out that K=33 yields the best output.

B. One vs All approach with SVM

Also known as OVR (One vs Rest) multi class strategy. For each classifier, the class is fitted against all other classes. In addition to its computationally efficency only $n_classes$ classifiers are needed, one advantage of this approach is interpretability.

We experimented with different cost penalties (1e-5 to 1e+5), loss functions (hinge and squared hinge) and penalty (L1 and L2) and observed that combination of L2 regularized hinge loss with C=100 yields the best result.

C. Neural Network

In order to train the neural network data set was first normalized. Normalization implies that all feature values from dataset should have zero mean and unit standard deviation.

$$X_n = \frac{X_n - \mu}{\sigma}$$

where μ is mean feature value and σ is standard deviation.

We trained neural network with 10% of training instances as validation set which is the used to tune model parameters. We built a model with two hidden layer (10, 5), input layer size (256) and output layer size(10). We chose activation function as sigmoid function which gives us smooth non-linear function and cost function with regularization.

We also attempted to utilize PCA to reduce the dimension of input data but didn't improve our accuracy much.

VI. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] G. Tzanetakis and P. Cook, *Musical genre classification of audio signals*. IEEE Transactions on Audio and Speech Processing, 2002.

[2] Michael Haggblade, Yang Hong and Kenny Kao, *Music Genre Classification*. http://cs229.stanford.edu/proj2011/HaggbladeHongKao-MusicGenreClassification.pdf.

[3] Bob L. Sturm, *An Analysis of the GTZAN Music Genre Dataset*. MIRUM 2012.

[4] Ioannis Panagakis, Emmanouil Benetos, and Constantine Kotropoulos, *Music genre classification: a multilinear approach*. ISMIR 2008.

[5] Fu, A., Lu, G., Ting, K.M., Zhang, D. *A Survey of Audio-Based Music Classification and Annotation*. IEEE transactions on multimedia 2010.