Music Genre Classification and Variance Comparison on Number of Genres

Miguel Francisco, miguelf@stanford.edu
Dong Myung Kim, dmk8265@stanford.edu

1 Abstract

In this project we apply machine learning techniques to build a music genre classifier. Focusing on the sonic properties of music, as opposed to metadata, we examine the classification accuracy of two types of features coefficients, Mel-Frequency Cepstral Coefficients and Chroma Features, using multi-class support vector machines. We hope to achieve significant improvements over random guessing and identify which coefficient yields higher accuracy. Additionally, we will classify songs into different numbers of genres, to analyze how the number of classes affects accuracy for each model.

2 Introduction

As music recommendation and auto-generated playlists become wildly popular, techniques to automatically classify the genre of any given song are anticipated to have important applications in such industries. Specifically, with the increasing volume of new music, it is much more efficient to rely on an accurate genre classification system rather than hiring musical experts or crowdsourcing to determine genre.

3 Data

The MARSYAS (Music Audio Retrieval and Synthesis for Audio Signals) open source software framework hosts the GTZAN Genre Collection [1], a collection of 30-second snippets of 1,000 songs in the .au file format covering ten different genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Although there are much larger music datasets such as the Million Song Dataset, the retrieval of actual audio is made much easier with the GTZAN dataset, as opposed to the Million Song Dataset, which is actually a collection of metadata. Furthermore, genre tagging in the Million Song Dataset is by artist, rather than by song, so it does not account for cases where the artists may span different musical styles.

3.1 MFCCs

There are many ways to numerically represent audio files. Mel-Frequency Cepstrum Coefficients (MFCCs) are one of the widely used feature extraction methods for machine learning in music-related fields (visualized below). These are based on cepstral representation (a nonlinear "spectrum of a

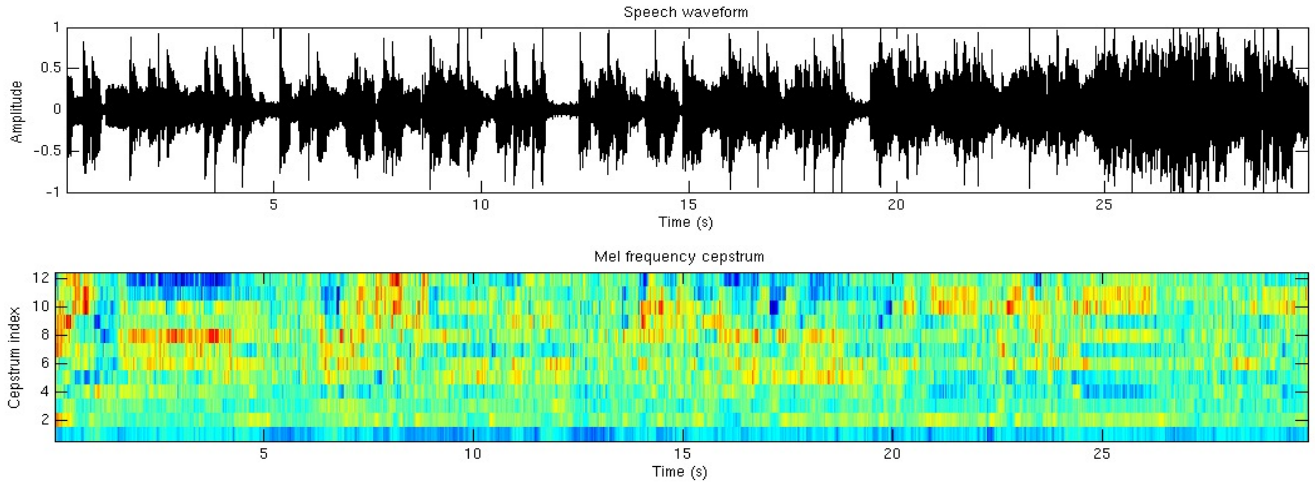


Figure 1: The sonic waveform and corresponding MFCC representation of "classical.00000.au" from the GTZAN dataset

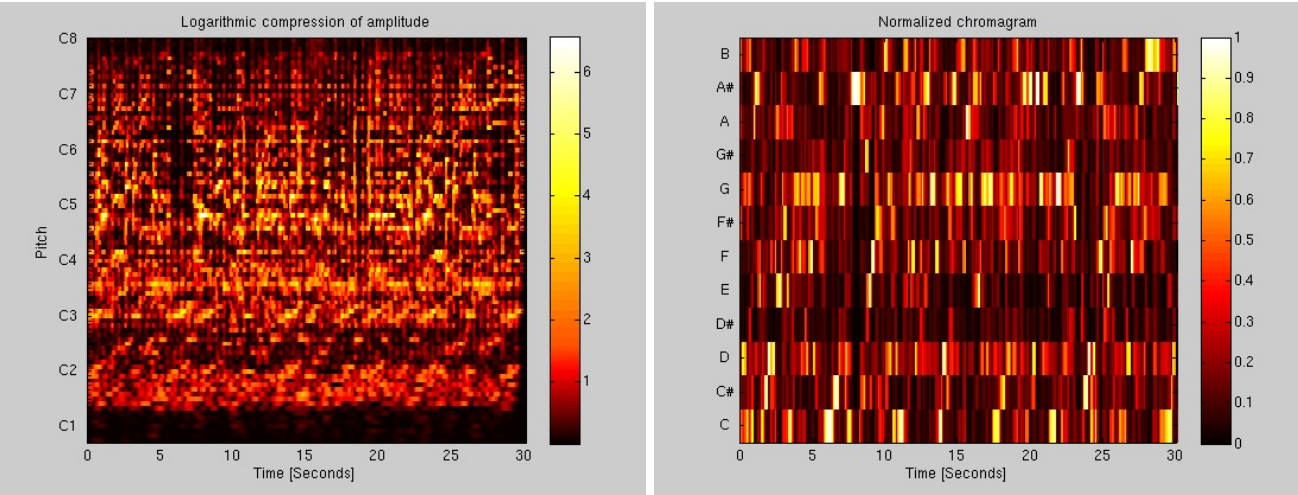


Figure 2: The pitch representation (left) and corresponding chroma (right) of "classical.00000.au" from the GTZAN dataset

spectrum"), but unlike the cepstrum, the Mel-Frequency Cepstrum uses frequency bands that are equally spaced over the mel-scale, and therefore are expected to produce a better representation of sound, closer to actual human perception.

Generating these coefficients requires a complex procedure including the Fourier transform, so we relied on open-source MFCC extraction software developed by Kamil Wojcicki[2]. As far as parameters are concerned, we settled on 12 cepstral coefficients and 18/15,000 (Hz) lower/upper frequency limits, a reasonable choice for investigating the audible sound spectrum.

MFCCs are based more on the timbre of the sound signals, and form an interesting contrast with chroma features (discussed more in depth later) which are based on pitch classes. We thought that contrasting the efficiency of two different coefficients in genre classification would be interesting, and also explored how increasing the number of features by combining both MFCCs and Chroma Features would impact the classification accuracy.

3.2 Chroma Features

which are generated using the short-time energy distributions over the chroma bands. The twelve chroma bands correspond to the twelve traditional pitch classes (C, C#, D, …) and as a result the chroma features have strong correlations with the harmonic progressions of the audio signals. We relied on the open-source Chroma Toolbox by Müller in order to generate chroma feature for our dataset.

3.3 MFCCs and Chroma Features Combined

Both MFCCs and Chroma Features generate a matrix for each 30 second snippet of audio. However the amount of features generated (number of entries in the generated feature matrix per snippet) is vastly different (~3000 for MFCCs and ~300 for Chroma Features). We have appended the Chroma Features to the MFCCs (after converting them to feature vectors) for simplicity, but the sheer volume of data that the MFCC holds naturally gives it more weight.

4 Methodology

Genre classification is a supervised learning problem, hence we have limited our attention to the supervised learning algorithms such as

Another widely used feature representation of audio signals is the set of chroma features,

As mentioned above, both MFCCs and Chroma Features are in matrix form, but the classification method that we used (multi-class support vector machines) does not support multidimensional matrices (one feature matrix per training example). We instead turned each feature matrix into a feature vector by appending each row vector after another. To assess the accuracy of the classifications, we used 10-fold cross validation to estimate the generalization error.

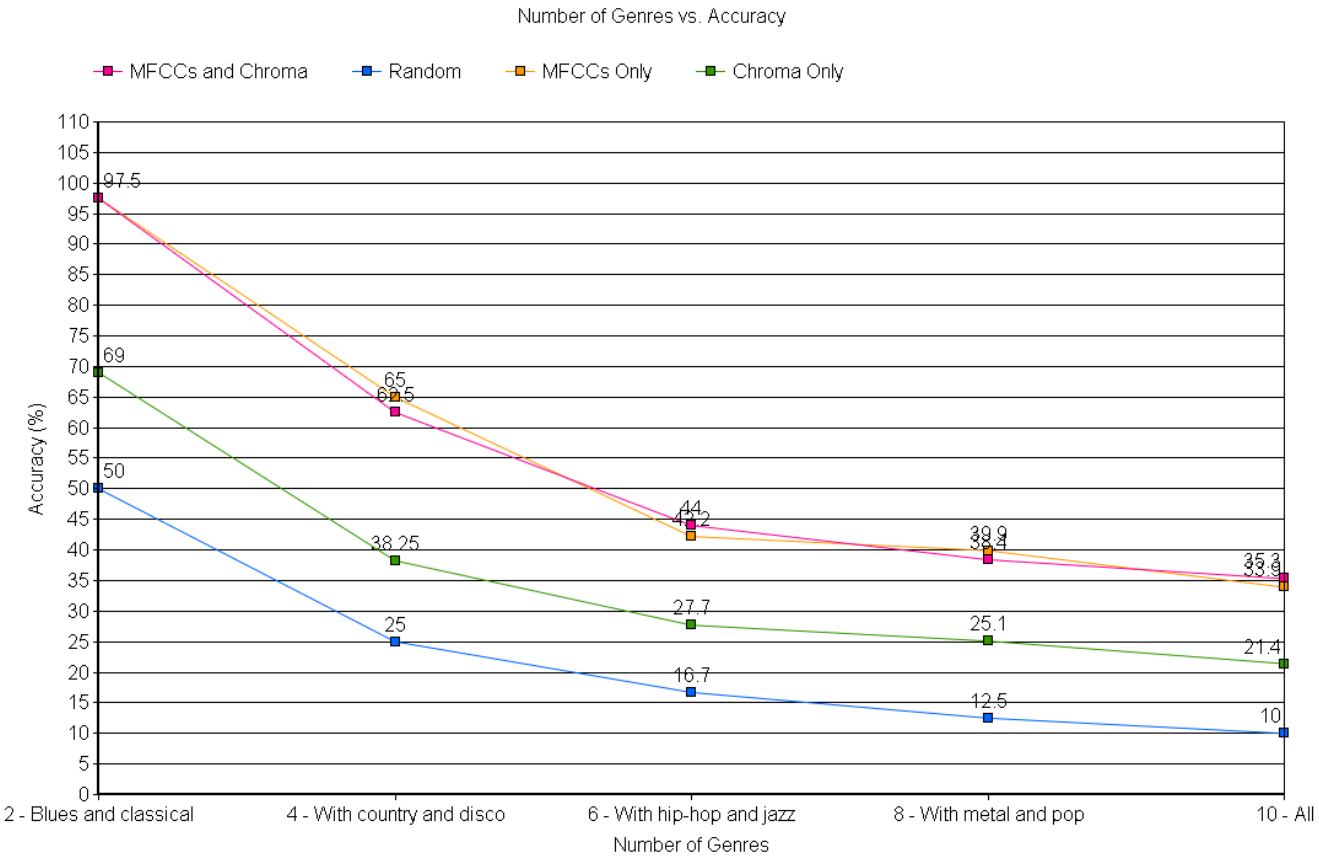
4.1 Multi-class SVMs

Since there are 10 total possible genres, we

support vector machines.
used multi-class support vector machines, a simple application of the traditional two-class support vector machines. A multi-class SVM uses a one vs. all approach; one support vector machine is trained for one class of training observations (the example is a member of this class, or is a member of any of the other classes). The testing example is tested against all classifiers, and we choose the class with the highest probability/confidence score.

We used multi-SVV code provided by Cody[3] which builds upon the svmtrain function provided by MATLAB.

Figure 3: Comparison of accuracy of the multi-class SVM on different genre subsets and feature representations



5 Evaluation

We ran three multi-class SVM tests: one using only chroma as features, one using only one MFCCs, and the third using both, as detailed in section 3.3. While we worked with a relatively small data set of only 100 snippets per genre where each snippet was the first 30 seconds of a song in the particular genre, the difference in accuracy shown above (Figure 3) seems to be statistically significant and therefore applicable to a larger dataset.

As the graph above shows, both MFCCs and Chroma Features showed statistically significant improvements over random guessing, which we used as the control to compare the classification accuracy, from binary classification to classifying snippets into all ten genres.

Chroma Features did not lead to a higher accuracy than that of MFCCs, and at times, even showed a lower classification accuracy. As discussed in section 3.3, this can possibly be a result of not equally weighting the two coefficients, and the lower classification accuracy than MFCCs only on four and eight genres is possibly due to having too many predictors and therefore overfitting.

6 Conclusion and Future Work

Our results suggest that a multi-class SVM is a very effective way to distinguish between a smaller amount of genres, but it loses accuracy when more genres are added, specifically when the added genres have similar musical qualities. Although chroma features have significant roles in other situations such as music transcription or song identification, the higher classification

The classification accuracy declined exponentially as the number of genres increased, but the classification accuracy for MFCCs and MFCCs with Chroma remained ~2.5 times higher than that of the control regardless of the number of genres. Chroma Features showed higher classification accuracy than the control, but significantly lower accuracy than that of MFCCs. This could be due to the fact that the MFCCs have almost ten times the amount of predictors compared to Chroma Features, but large number of predictors compared to observations might also lead to overfitting, which tends to produce worse prediction results for a test set. The large difference in accuracy is probably due to the musical property that these coefficients are based on; MFCCs are based on timbre, and Chroma Features are based on pitch class. The combination of MFCCs and accuracy of MFCCs, even sometimes when combined with chroma features, signifies that the timbre of the sound, as opposed to the note values, holds greater weights in differentiating between genres. Furthermore, this proves that even if similar melodies and chord structures are shared between songs, the combination of instrumentation and vocal tone determine genre.

In further work, we hope to use principal components analysis to determine the 300 most representative components of each song's MFCC in order to better test the effectiveness of chroma features when combined with MFCCs. Additionally, we would like to explore additional algorithms that might possibly have higher classification accuracy even on a larger amount of genres. Although the multi-class SVM is significantly better than random, it is not entirely effective.

Works Cited

[1] G. Tzanetakis and P. Cook (2002). *GTZAN Genre Colection*. Retrieved from http://marsyas.info/download/data_sets/ , Dec. 2013
[2] Wojcicki, Kamil. "HTK MFCC MATLAB." *MATLAB Central- File Exchange*. N.p., 11 Sept. 2011. Web. 23 Nov. 2013
[3] Cody. "Multi Class SVM." *MATLAB Central- File Exchange*. N.p., 7 Dec. 2012. Web. 27 Nov. 2013.