Automatic Music Genre Classification using Support Vector Machines

Aditya Mundle

1 Background

With the ever growing music database on the internet, it is extremely important to efficiently and accurately process all that music information. Music information retrieval (MIR) addresses this issue which includes music classification as one of its many areas. Classification addresses the problem of organizing, categorizing and describing music for retrieval. This is a major component of the online and electronic music databases like iTunes, Pandora, Last.fm etc.

Music genre can be used as a basis for music classification. Genre classification has traditionally uses human involvement which makes the process highly subjective, occasionally inaccurate and also difficult and time consuming. If this human involvement is removed (or limited), the automatic classification system will also allow the development of automatic analysis, segmentation, indexing and retrieval. Music from the same genre have similar characteristics like instruments used, rhythmic patterns, pitch distributions etc. [4]. These can be categorized into different features: Timbral features, originating from the speech recognition techniques, are computed for every short time frame of sound and use the short time fourier transform (STFT) [10]. The beat spectrum is used to represent the rhythmic feature [5]. The pitch content will have features related to the frequency information of the music signal.

The content-based features: timbral texture, rhythmic patterns and pitch distribution are used as input vector to a supervised classification system which then can be used to classify the music accordingly. K-Nearest Neighbors (KNN), Gaussian Mixture Models (GMM), Support Vector Machines(SVM), Tree-Based Vector Quantization, Least Squares Minimum Distance Classifier (LSMDC), Quadrature classifier have been used for the classification using the features listed above [12, 3, 9].

2 Aim

The aim of this project is to classify music into four categories of genres: classical, jazz, rock and electronic. Layering approach described in [13] will be used to discriminate the genres as shown in Figure 1. The proposed first layer of classification will be classical/jazz and rock/electronic using SVM1. The second layer containing the classical/jazz category will be classified further using SVM2 and the rock/electronic category will be classified into rock and electronic using SVM3 *(SVM1, SVM2, SVM3 refer to Support Vector Machines used to perform the classification task in each of the three cases.)*.

This layering approach will also be analyzed against the more conventional 2-way SVM classification approach used (i.e *Classical v/s non-classical etc.*) [2, 7]. Impact of the use of different audio codecs: WAV, MP3, AAC, OGG on performance will also be evaluated.

3 Methodology

This project treats the problem of music genre classification as a supervised machine learning problem. This requires that the training dataset be mapped into feature vectors. Using these feature vectors, SVM will be used to classify the test data into the four genres mentioned above.

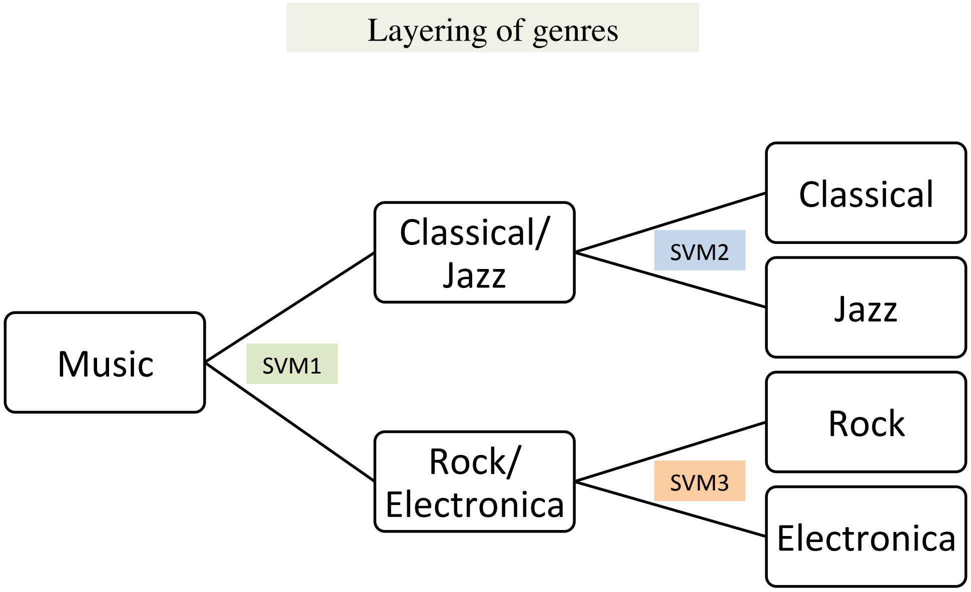

Figure 1: Layering Approach

3.1 Dataset

The dataset will consist of music from the four genres. The audio clip creation will be similar to the approach used in [7] - from each music track 30 seconds of sound will be extracted after initial 30 seconds. This procedure will be used for all the audio codecs: WAV, MP3, AAC, OGG will. The audio codecs give different feature vectors as MP3, AAC, OGG are lossy formats [1].

3.2 Features

3.2.1 Timbral Features

Timbral features have been used for music-speech discrimination and speech recognition. They are used to differentiate sounds with similar rhythmic and pitch content [6, 11]. To extract these, the signal is divided into frames that are statistically stationary using a Hanning window function. The spectral features used to represent timbral textures are as follows described below.

Spectrum Power
Using the Hanning window $h(n)$,

$$h(n) = \frac{\sqrt{8/3}}{2}\left[1 - \cos\left(2\pi\frac{n}{N}\right)\right] \tag{1}$$

the spectral power of the signal $x(n)$ is calculated as

$$S(k) = 10log_{10}\left[\frac{1}{N}\left|\sum_{n=0}^{N-1} x(n)h(n)\exp\left(-j2\pi\frac{nk}{N}\right)\right|^2\right] \tag{2}$$

Mel-Spectral Cepstral Coefficients (MFCCs)
MFCCs have been used to distinguish between speech and non-speech audio signals [8]. These are perceptually motivated and use short time fourier transform (STFT) and divide the signal according to the Mel-scale which is acoustically more relevant. The resulting feature vectors are decorrelated using Discrete Cosine Transform (DCT).[12] notes that 5 coefficients are sufficient for genre classification whereas [3] uses 12 MFCCs. The performance of the SVM with these 2 numbers will be evaluated.

Zero Crossing Rate
Zero crossing measures the noisiness of the signal. A zero crossing occurs when the successive samples change signs (time domain).

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \tag{3}$$

where the *sign* function is 1 for positive arguments and 0 for negative arguments and $x[n]$ is the time domain signal for frame t.

3.2.2 Rhythmic Content Features

Rhythmic content features characterize the movement of signals over time. They contain information about the regularirty of the rhythm, the beat, the tempo and the time signature [12]. The beat spectrum measures the rhythm and tempo of the music [13]. It can distinguish between the different kinds of rhythms at the same tempo. The beat spectrum is computed as follows [13]:

1. Music is parameterized using a spectrum which gives a sequence of feature vectors.

2. A similarity matrix is formed using a distance measure to calculate the similarity between all pairwise combinations of feature vectors.

3. Using diagonal sums or auto-correlation, the beat spectrum is obtained from the periodicities in the similarity matrix.

3.2.3 LPC Derived Spectrum

Linear predictor is important due to its accuracy with vocal signals in music [13]. A music sample can be approximated as a linear combination of past music samples. By minimizing the sum of the square of the actual music and linear predictor samples, a set of predictor coefficients can be determined.

3.3 Classification

Support Vector Machines (SVM) are used in binary classification tasks. Suppose given a set of training data $(x_1, x_2, x_3, ...x_n)$ and the class labels $(y_1, y_2, y_3, ...y_n)$ where $x_i \in R^n$ and $y_i \in \{-1, -1\}$. The input vectors are transformed into a high dimensional feature space using non-linear transformation ϕ and then a linear separation is performed in feature space. If a non-linear SVM is to be used, a kernel function $K(x, y)$ will be required. I this case, a Radial Basis Function (RBF) is used with Gaussian kernel of width $c > 0$.

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b) \tag{4}$$

$$K(x,y) = exp(-\left|x-y\right|^2/c) \tag{5}$$

The process is shown in the form of a flowchart in Figure 2.

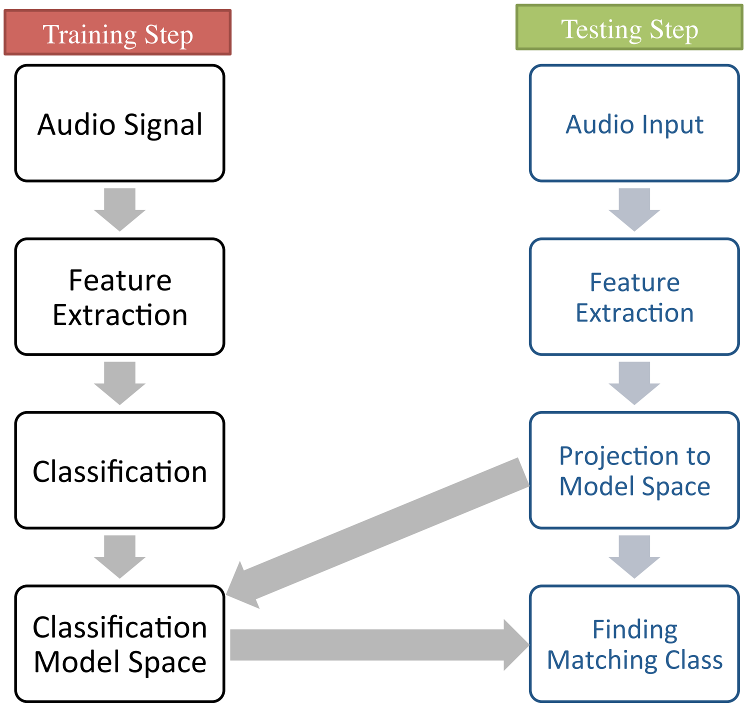

Figure 2: The process consists of 2 parts: Training step and Testing step. In the training step, features described before are extracted from the audio signal dataset. This is used train the SVM model. In the testing step, the features are extracted from the data and projected into a higher dimension. Using the model space from the training step, the data is labeled.

It would be good to include a discussion on performance assessment.

References

[1] Zehra Cataltepe, Yusuf Yaslan, and Abdullah Sonmez. Music Genre Classification Using MIDI and Audio Features. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9, 2007.

[2] H. Deshpande, R. Singh, and U. Nam. Classification of music signals in the visual domain. In *Proceedings of the COST-G6 Conference on Digital Audio Effects*, pages 6–9. Citeseer, 2001.

[3] Hrishikesh Deshpande, Unjung Nam, and Rohit Singh. MUGEC : Automatic Music Genre Classification. pages 1–13, 2001.

[4] W. Jay Dowling, Dane Harwood, and Mark C. Gridley. Music cognition by w. jay dowling and dane harwood. *The Journal of the Acoustical Society of America*, 83(2):840–840, 1988.

[5] J. Foote and S. Uchihashi. The beat spectrum: a new approach to rhythm analysis. *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 881–884, 2001.

[6] N J Hunt, N Lennig, P Mermeletein, and Bell Northern Reeeerch. Stllablebased recognition. *English*, (3):880–883, 1980.

[7] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, page 282, 2003.

[8] B. Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, volume 28. Citeseer, 2000.

[9] D. Pye. Content-based methods for the management of digital music. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, pages 2437–2440, 2000.

[10] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[11] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1331–1334, 1997.

[12] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[13] Changsheng Xu, Namunu C Maddage, Xi Shao, Fang Cao, and Qi Tian. Classification using support. *Power*, pages 429–432, 2003.