Classification of Music by Genre using Probabilistic Models and Deep Learning Models

Tshepo Nkambule and Ritesh Ajoodha

The University of the Witwatersrand, Johannesburg, South Africa,
1611821@students.wits.ac.za, ritesh.ajoodha@wits.ac.za,
WWW home page: https://www.wits.ac.za/

Abstract. The digital shift in the distribution of music has presented the need for effective automated classification of large volumes of music into various categories. In this paper automatic music genre classification is performed by first identifying and extracting representative aspects of a music piece. Subsequently music features are tested for their significance in the task of genre classification using mutual information gain in order to make the feature vector compact, comprehensive and efficient. After several off-the-shelve classifiers were used, Support Vector Machines with a radial basis function kernel turned out to be the best performing model achieving an accuracy of 80.80% in classifying music pieces into 1 of 10 genres in GTZAN.

Keywords: Music Features· Mutual Information Gain · Music Genre Classification· Support Vector Machines.

1 Introduction

Music is one of the most beautiful conceptions of the human mind. People associate emotional, cultural and spiritual meaning to music. While the distribution of music has shifted to digital platforms such as iTunes, Spotify, Google Play and other streaming platforms the criteria for classifying music have remained the same. Mood, style, and genre are popular criteria for music classification with genre being the most popular one. Musical genres are defined as categorical labels created by humans to distinguish music pieces [1]. Traditionally music genre annotation was performed by hand. Annotating by hand was quite a tedious and expensive task which resulted in inconsistencies. Inconsistencies amongst different annotators arise due to different cultural evolution of music genres, similarity amongst music genres, and the introduction of new music genres. Although music genres evolve the underlying music signals whether analog or digital of music pieces belonging to the same genre remain similar. This is because they are composed of similar instruments, have similar distributions in pitch, and similar patterns in rhythm [2]. The common characteristics have made it possible to perform automatic musical genre classification. Our aim is to perform automatic music genre classification by making use of only the underlying music signals.

2 Related Work

Music Genre Classification is the process of categorising music pieces using traditional and cultural aspects [3]. The presence of a reliable ground truth is essential in the implementation of accurate music genre classifiers. Ambiguity in ground truth results from: unclear definitions of music genres, inconsistencies in various music sources, similarity amongst music genres, introduction of new music genres and evolution of existing music genres [4]. As a consequence of the ambiguities that exist between genre definitions classification accuracy becomes inescapably bounded as many annotators may disagree on a particular genre classification of a piece of music [3].

To perform automatic music genre classification, musical aspects that have a significant contribution to perception of genre must be identified. The following aspects: harmony, melody, rhythm and sound (timbre, dynamics, and texture) are considered to have significant contributions to the notion of musical genre [3]. The musical aspects listed above are obtained from a music piece through feature extraction. Extracting features from a music piece involves identifying effective (not computationally taxing), comprehensive (represents music piece well) and compact (requires minimal storage) representation of the components of a music piece [2]. Classical music features are content-based features. Content-based features are divided into three categories as follows.

- **Timbre content-based features:** are used to distinguish sounds that have similar rhythmic and pitch content.
- **Rhythmic content-based features:** describe the movement of music signals over the time-domain. Examples include: beat and tempo.
- **Pitch content-based features:** pertain to harmony and melody of music. Their extraction is based on various pitch detection or extraction procedures.

A reliable dataset is essential for performing automatic music genre classification. The groundbreaking work of [1] resulted in the creation of the GTZAN dataset. The GTZAN dataset is one of the most popular dataset with a well established benchmark in automatic music genre classification. It consists of a 1 000 music pieces of 30 seconds duration each with 100 samples in each of 10 different music genres [5].

3 Music Features

In this section we present a number of features that could be used to perform automatic music genre classification. These features are considered to be effective, comprehensive and compact representations of components of a music piece. These representative features are categorized into four main groups

- **Magnitude-based features**
  these are mainly timbral features that represent music aspects such as loudness, compactness and pitch [3]. The timbre quality of a music piece allows

humans to group together different sounds originating from the same source such as two recordings made with the same instrument [8].

- **Tempo-based features**
  these are features that explore and describe the rhythmic aspects of a music piece [3].
- **Pitch-based features**
  these are features that describe pitch the basic building block of key, melody, and harmony of a piece of music [8].
- **Chordal progression features**
  these explore chroma which can be a chordal distinguishing feature of music signals [3].

Before we explore the feature categories outlined above it is imperative that we introduce means in which the features in the different groups can be represented.

3.1 Feature Extraction and Representation

Most existing audio analysis systems involve two major stages: *feature extraction* and *decision, interpretation and classification*. Feature extraction serves the following 2 purposes

1. **Dimensionality reduction:** When processing an entire audio file, the raw audio data is too large to handle in a meaningful way. A feature set of is used to present this data with fewer values by discarding irrelevant information. Typically an instantaneous feature will produce a single feature value for each time frame in the audio signal or even a single value for the entire signal [8].
2. **More meaningful representation:** While all the information that can possibly be extracted is implicitly contained in the raw audio file, it is essential that we focus on representing music aspects in machine or human interpretable manner. It is not necessary for a feature to be meaningful in a perceptual way or musical way nor does it have to be interpretable by humans [8].

Each of the features that can be computed from a music piece usually result in a n dimensional vector as the feature value changes as the audio progresses in the time domain. The value of n is typically large depending on the length of the audio, this presents high dimensional feature vectors which are not optimal to have. Consider a feature F that takes on the values $(f_1, f_2, f_3, ..., f_n)$, the following statistics are chosen to represent the set of values for F

1. **Mean:** This is the average value of F. It is computed by

$$\mu_F = \frac{1}{n}\sum_{i=1}^{n} f_i \tag{1}$$

   The result is a value between the minimum and maximum value for F.

2. **Standard deviation:** This indicates how spread out the values of F are. It can be computed by

$$\sigma_F = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \mu_F)^2} \tag{2}$$

Strictly speaking this is the biased estimate of the standard deviation.

3.2 Magnitude-based Features

The magnitude spectrum, obtained by taking the absolute value of the fast Fourier transform of a music signal, contains a number of spectral features that can be used for automatic music genre classification [3]. In this section we present and briefly describe some of the spectral features embedded in the magnitude spectrum.

1. *Spectral Rolloff:* a frequency bin below which the accumulated magnitudes of the Short Time Fourier Transform (STFT) reach a certain percentage κ of the overall sum of magnitudes. Common values of κ being 85% or 95%. It measures the bandwidth of an audio signal [8].
2. *Spectral Flux:* the average difference between consecutive STFT frames. It measures the amount of change of the spectral shape [8].
3. *Spectral Centroid:* the frequency-weighted sum of the power spectrum normalized by its unweighted sum. It represents the center of gravity of spectral energy [8].
4. *Spectral Spread:* the standard deviation of the power spectrum around the spectral centroid. It describes the concentration of the power spectrum around the spectral centroid [8].
5. *Spectral Decrease:* an estimation of the steepness of the decrease of the spectral envelope over frequency. The result is a value in the range $[0, 1]$, with small values indicating the concentration of spectral energy at bin 0. This feature is not defined for silent audio blocks [8].
6. *Spectral Slope:* a measure of the slope of the spectral shape using a linear approximation of the magnitude spectrum [8].
7. *Mel Frequency Cepstral Coefficients (MFCCs):* a compact description of the spectral envelope of an audio signal. MFCCs are not defined for a silent audio signal [8].
8. *Spectral Flatness:* the ratio of geometric mean and arithmetic mean of the magnitude spectrum. It is a measure of the noisiness of a audio signal [8].

3.3 Tempo-based Features

Temporal aspects of audio signals such as the tempo and rhythm are important properties. A fundamental building block of tempo and rhythm is the onset. The onset marks the beginning of a musical sound event such as a tone or a stroke on a percussive instrument. The start time of an event is important in the human perception of music, as listeners seem to perceive musical events in terms of onset-to-onset intervals [8]. In this section we present a number of tempo and rhythm related features with brief descriptions.

1. *Tempo:* is the rate at which perceived pulses with equal duration units occur at a moderate and natural rate. A typical value for the natural rate is 100 *Beats per Minute (BPM)*[8].
2. *Energy:* a measure of the intensity of audio signals typically using the root mean square energy (RMS) of the signal [3,8].
3. *Beat histogram:* is a way to visualize rhythmic properties of a music signal. It is very similar to the magnitude spectrum, the frequency in the case of a beat histogram has the unit *BPM*. The computation of a beat histogram produces a very large design matrix, hence simple but quite meaningful features are required to represent the beat histogram [3]. A widely used set of features used to capture the beat histogram was introduced by [1] and consists of the following
   - the overall sum of the histogram.
   - the relative amplitude of the highest peak.
   - the relative amplitude of second highest peak.
   - the amplitude ratio of second highest to highest peak.
   - the BPM frequencies of the highest and second highest peak.

   Other representations of the beat histogram include statistical features such as the mean, standard deviation, kurtosis, skewness and entropy [8].

3.4 Pitch-based Features

A higher perception of pitch in nusic means that the underlying audio signal has a high frequency [8]. Humans find it hard to distinguish pitch, as even when a music signal is a combination of sinusoidal components with frequencies $f_0, 2f_0, 3f_0, ...$ the fundamental frequency f_0 will dominate pitch perception [8]. In this section we will explore a pitch related music feature and its brief description and applications.

1. *Zero Crossing Rate:* is the number of changes of sign in consecutive blocks of an audio sample. Since it is a thorough percussive descriptor, it has been used in speech recognition and in audio analysis [3].

3.5 Chordal progression Features

The intention of this section is to give an overview of the pitch chroma presentation of the tonal music property known as music key.

1. *Chroma:* is a histogram-like 12 dimensional vector in which each dimension represents one pitch class. It can be viewed as a a distribution of the pitch classes, in which the value of each dimension represents both the number of occurrences of that pitch and its energy. It is important to note that the pitch chroma is not a series of unrelated observations it is a distribution [8]. There is no standard set of features to be extracted from chroma, although the features used by [1] were simple and effective.

4 Feature Selection

Feature selection provides a way to discard irrelevant and redundant data, which can reduce computation time, improve learning accuracy and facilitate better comprehension of the leaning model and data. Information gain ranking eliminates features by computing the dependency of a target variable $\boldsymbol{Y}$ on a predictor variable $\boldsymbol{X}$, in which features that show high correlation with $\boldsymbol{Y}$ are kept [9]. The results of information gain ranking are shown in Table 1 below.

Table 1: Features maintained are shown on the upper region of the table while features discarded are shown in lower region.

Features maintained	**Rep**	**Dim** 54
Spectral Contrast	mean	7
Spectral Rolloff	mean + std	2
Spectral Flux	mean + std	2
Spectral Crest	mean + std	2
Spectral Flatness	mean + std	2
Spectral Decrease	mean + std	2
Spectral Flatness	mean + std	2
Spectral Kurtosis	mean + std	2
Spectral Slope	mean + std	2
Spectral Skewness	mean + std	2
Spectral Centroid	mean + std	2
Spectral Spread	mean + std	2
Spectral Entropy		1
Zero Crossing Rate	mean + std	2
Mel Frequency Cepstral Coefficients	mean	17
Root Mean Square Energy	mean + std	2
Beat Histogram	sum + mean + std	3
Auto correlation Coefficients	mean + std	2
Features eliminated	**Rep**	**Dim** 51
Spectral Crest Factor	mean + std	2
Spectral Tonal Power Ratio	mean + std	2
Mel Frequency Cepstral Coefficients	mean	35
Chroma	mean	12

5 Automatic Music Genre Classification

Figure 1 shows the results of taking the first m with $m \leq n$ highest ranked features to perform genre classification. We observe that there is a cut-off point such that a number of features can be discarded without having a significant impact on classification accuracy. The cut-off point indicated by a red line on Figure 1 is when $m = 54$, in which we keep the top 54 features and disregard 51 least significant features. Table 2 shows the performance and hyper=paremeters of different models.

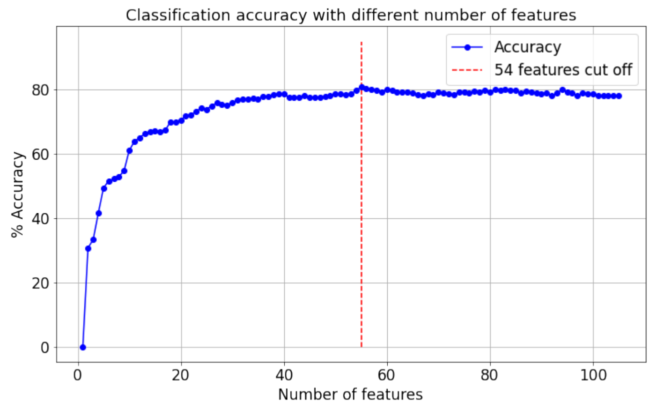

Fig. 1: Classification accuracy with different number of significant features.

Table 2: Model perfomance and hyper-parameters.

Classifier	Accuracy	Time to build (s)	Parameters
Naive Bayes (**NB**)	54.50%	0.0019	Gaussian naive bayes with smoothing
k Nearest Neighbour (k**NN**)	69.70%	0.011	k = 7, with manhattan distance metric, weighting = distance
Random Forests (**RF**)	72.40%	61.08	split function=gini , number of tress = 100 max depth = 100
Logistic Regression (**LR**)	75.80%	0.08	solver = newton-cg , max iterations = 500,
Support Vector Machines (**SVM**)	80.80%	0.3	radial basis function kernel, tolerance= 0.001, regularization = 0.17, tolerance = 0.0001
Multilayer Perceptron (**MLP**)	77.30%	0.23	hidden layers = 2, learning rate = 0.02, activation = relu, max iterations = 200 solver = adam, tolerance = 0.0001

6 Conclusion and Recommendations

In this paper automatic music genre classification was performed by first identifying, extracting and selecting representative aspects of music piece. After several off-the-shelve classifiers were used, **SVM** turned out to be the best performing model achieving an accuracy of 80.80%. The experimental results in this work are on par with benchmarks that have been set by several authors including [2,3]. Limitations that may be capping model performance are the absence of data and a reliable ground truth. Since music genre annotation is typically done by hand, different annotators may not agree on which genre a particular music piece belongs to due to their different cultural backgrounds, this creates inconsistencies in ground truth and affects model evaluation. While the GTZAN dataset is a popular dataset, it is only confined to 10 genres which is significantly small compared to the number of existing genres, this means that it is difficult to develop models that will generalise to cover a wide range of genres. In future a hierarchical genre structure can be adopted towards the construction of a large corpus dataset similar to GTZAN that will span a significant number of existing genres. This new dataset could improve classification accuracy as data sensitive models will perform better due to presence of more data and models will generalize and cater for a wider range of genres. Researchers should also consider augmenting the psychological and social aspects of music such as emotion and danceability as features to be considered in automatic music genre classification. As these psychological and social aspects may be indicators of the association of a music piece with a particular genre.

Acknowledgments. This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835).

References

1. George Tzanetakis and Perry Cook. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5):293–302, 2002.
2. Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 282–289, 2003.
3. Ritesh Ajoodha, Richard Klein, and Benjamin Rosman. Single- labelled music genre classification using content-based features. The Pattern Recognition Association of South Africa (PRASA), pages 66–71, 11 2015.
4. Cory McKay and Ichiro Fujinaga. Musical genre classifica- tion: Is it worth pursuing and how can it be improved? In ISMIR, pages 101–106, 2006.
5. Bob L Sturm. An analysis of the gtzan music genre dataset. In Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies, pages 7–12, 2012.
6. S. Sigtia and S. Dixon. Improved music feature learning with deep neural networks. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6959–6963, 2014.
7. Mingwen Dong. Convolutional neural network achieves human-level ac- curacy in music genre classification. ArXiv, abs/1802.09697, 2018.
8. Lerch(auth.) 2012] Alexander Lerch(auth.). An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. Wiley-IEEE Press, 2012.
9. Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, RobertTrevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective.ACM Computing Surveys, 50, 01 2016.