Music Genre Classification Based on Signal Processing

© Stepan Evstifeev[1] © Ivan Shanin[2]
[1] Lomonosov Moscow State University,
Moscow, Russia
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences,
Moscow, Russia
minemile69@gmail.com v08shanin@gmail.com

Abstract. Music genre is a description, that allows to categorize music compositions into broader categories with similar characteristics. With the development of streaming platforms (iTunes music, SoundCloud, Spotify), the automatic classification of music is becoming increasingly important as a way to intelligently search in a large number of music files, and also as a support in building recommendation systems. In this paper, this approach is based on the extraction of information from a signal (timbre, rhythm, melody, pitch), as well as the construction of high-level features with subsequent classification by methods of machine learning, in particular, the gradient boosting trees and neural networks are considered. The GTZAN dataset is used to evaluate the performance of the algorithms with the best result of 78% precision. Comparison of algorithm with third-party systems is considered. To test algorithms on real data, a website has been developed that allows to automatically classify users' music.

Keywords: automatic music genre classification, information music retrieval, audio feature.

1 Introduction

The musical genre is a conventional category that determines to which type (compositional, stylistic, narratively) the musical composition refers. Listeners use genres to search for similar music, to organize music files into playlists. In the music industry, genres are used as a key way to determine the target market.

Currently, the number of music files in digital form, available on the Internet, is growing rapidly. There is an ability to download and save any song on the device, as well as the development of various services that provide the end user access to a large catalog of music on-demand by subscription (iTunes music, SoundCloud, Spotify). Typically, these services rely on manual classification, which is slow and time-consuming.

Moreover, features extracted by genre classification algorithms can be used in other problems of music information retrieval (MIR) based on its content: clustering tasks, segmentation, similarity analysis, recommendation systems and generation of similar music genre [1]. In addition, the automatic classification of genres of music is becoming increasingly important as a way of structuring and organizing a large volume of digital music, for example, in playlists or databases.

However, the task of unambiguous classification of the genre of music is complex for both human and computers. Often there is no generally accepted understanding of what characteristics a genre has, what genres should be used in genre's taxonomy, and how they relate to each other. An additional problem is that different people understand the genres differently, which leads to inconsistencies.

The division of music into genres is ambiguous and subjective task [11] because genres do not have a clear definition and over time, in different cultures, can be perceived differently (for example, differences in the perception of pop music in the 60's and 90's). There are a small number of genres that have a clear definition, and the available information about them is often ambiguous - some genres overlap considerably, and individual records can simultaneously belong to different, but similar genres. Between genres there are often complex relationships, some genres are broader, while others are narrower [15].

As has been shown, the genre is a subjective evaluation of music, but humans, usually accurately distinguish music genres based on a 250 millisecond - 3 second audio clip as investigated in [14, 19]. This suggests that human judge the genre using only musical characteristics, without using a theoretical, high-level understanding of music. Thus, to classify a genre, one can use the features associated with the characteristics of the musical composition: texture, sound instruments and rhythmic structure.

In this paper, an algorithm for automatic genre classification is proposed. A set of features is proposed based on the musical characteristics of the composition. Various machine learning models have been created, such as k-neighbors classifier, artificial neural networks, gradient boosting trees and gaussian mixture model, which were trained on vectors from these features. Those models were evaluated and compared on GTZAN dataset [6]. A website has been developed to classify music genres uploaded by users.

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

2 Related work

The basis of systems for automatic analysis of audio signals of any type is the extraction of the vector of features. A lot of works are devoted to extraction of features from signal. Consider the work on the extraction of low-level features - these are the features that were calculated on short signal intervals called “window”.

One of the fundamental works in this area is the work of Dannenberg et al. [3], based on the extraction of 13 low-level features, such as the tempo, loudness, height and duration of sound, followed by classification by the naive Bayesian and neural network methods. This approach correctly classified 98% of music among 4 genres, but it has greatly decreased on classification of 8 genres with a result of 77% accuracy. Although the result is not impressive, these ideas will be further improved in the future.

The development of ideas for the extraction of low-level characters from music is the classical article of G. Tzanetakis et al. [26]. They proposed three sets of features that represent the timbre, rhythm and pitch of the sound - Short-term Fourier Transform (STFT), Cepstral Mel-Frequency Coefficients (MFCC), and Wavelet Transform, respectively. They suggested using "texture" windows to generalize the timbral features by applying low-order statistics to larger windows. Such a generalization allows to reduce computational costs, but it is also closer to human perception [19]. These features were used in the Gaussian mixture model and k-Nearest Neighbor models with a precision result of 61% on modern music collections.

More recently, neural network approaches based on deep learning [4] have been increasingly used in audio informational retrieval areas. Convolutional and recurrent neural networks show very high results in audio processing, namely natural language processing [25], voice recognition [9], music generation and recognition of patterns in periodic data.

In the field of music genre classification, neural networks are used both for the extraction of features and for classification [20].

The article [21] shows that convolutional neural network (CNN) can learn spectral-timbre features, similar in efficiency with hand-crafted features. The input of the CNN is STFT of the audio signal, and its outputs are used for classification by classical machine learning methods, or by another neural network. The authors have trained a neural network, based on these features, with 500 neurons and 3 hidden layers with a result of 83% accuracy on the GTZAN dataset.

Another article [24] examines Long Short-Term Memory (LSTM) neural network architecture. To classify more than 6 genres, they used hierarchical divide-and-conquer strategy: they divided the genres into strong and mild, which were also divided into sub-genres. Each sub-genre was classified by a separate LSTM module, until it found the final genre. Their experiments showed that this architecture gives 50% accuracy.

The authors of the article [28] suggests using combination of max- and average-pooling to provide more statistical information to higher level neural networks and using shortcut connections to skip one or more layers. It was shown that these methods improve the accuracy of neural networks.

The work [22] presents the main types of low-level features: temporal, energy, spectral shape and perceptual features.

There is a number of high-level features that describe the entire song, which are presented in [16]: instrumentation, musical texture, rhythm, dynamics, melody, chords. It is shown that these characteristics correlate well with the task of classifying the genre.

3 Feature extraction

In this paper, we used the following features on each of the 20 milliseconds “analysis” windows:

- Zero crossing rate. The rate of sign-changes along a signal. It useful to detect the amount of noise in a signal.

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

  Where s is a signal of length T and $1_{R<0}$ is an indicator function.
- Spectral Centroids. The spectral centroid is defined as the “center of gravity” of the magnitude spectrum of the STFT.

$$C_t = \frac{\sum_{n=1}^{N}(M_t[n] * n)}{\sum_{1}^{N}(M_t[n])}$$

  Where $M_t[n]$ is the magnitude of the Fourier transform at the frame t and frequency bin n.
- Spectral Rolloff. The spectral rolloff is defined as the frequency R_t below which 85% of the magnitude distribution is concentrated. The rolloff is a measure of spectral shape.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n]$$

- Spectral Flux. The spectral flux is defined as the squared difference between the magnitudes of successive spectral distributions.

$$F_t = \sum_{n=1}^{N}(M_t[n] - M_{t-1}[n])^2$$

  The spectral flux is a measure of the local variation of the spectrum.
- Low Energy**.** The percentage of “analysis” windows that have energy less than the average energy of the “analysis” windows over the “texture” window.
- Mel-Frequency Cepstral Coefficients (MFCC) [13]. Cepstrum is the result of a discrete cosine transform from the logarithm of the amplitude spectrum of the signal. The mel-scale models the frequency sensitivity of the human hearing and Mel-Frequency Cepstral Coefficients are the values of the cepstrum, distributed on a mel-scale using multirate filterbanks.

  On each of the “analysis” windows, the mean,

variance, minimum, maximum, median and standard deviation of the corresponding multivariate values were calculated.

For the rhythmic features, a Discrete Wavelet Transformation (DWT) was used. It allows with small computational difficulties to find onset events for constructing beat histogram from the DWT coefficients [8].

We also extracted features from a convolutional neural network with 2 hidden layers as proposed in [21, 28].

4 Classification

4.1 Basic concepts and definitions

In general, the classification problem can be formulated in the following way: given the set of objects X and the set of answers $Y = \{1, \dots, M\}$ and there exists a target function $y^* = X \to Y$ whose values are known only on the finite subdomain of objects $\{x_1, \dots, x_l\} \subset X$. It is required to construct an algorithm $a: X \to Y$, that can classify an arbitrary object from X. In the problem of genre classification $M > 2$, which corresponds to a multiclass classification.

4.2 Machine learning algorithms

In this paper, we constructed and compared the classification algorithms that were most successfully applied in practice [6-8, 11]:

- k-neighbors algorithm (KNN). Memory-based classifier requires no model to be fit. Given a query point x_0, algorithm find the k training points $x_{(r)}$, r = 1, … , k closest in distance to x_0, then classify using majority vote among the k neighbors.
- Gaussian mixture model (GMM). For each class is assumed the existence of a probability density function, expressed as a mixture of a set of multidimensional normal (Gaussian) distributions. To evaluate the parameters of each component, an iterative algorithm expectation maximization (EM algorithm) is used;
- Artificial neural network (ANN) - classification algorithm, that consists of a large number of units called neurons. Neurons together receive and send information via weighted connections (synapses).
- Support vector machine (SVM) – machine learning method, that constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Kernel trick allows to construct non-linear separation.

In this paper, we also evaluated the popular gradient boosting trees classification algorithm, which stably produces high results on data with a complex, nonlinear structure, and the results of this algorithm can be easily interpreted.

Boosting is an ensemble of algorithms that allows one of several weak models (usually a decision tree) to create one strong one. In other words, the goal of boosting is to consistently apply weak classification algorithms to the data. The predictions of each of the models are combined by a weighted majority to obtain the final prediction, $G_m(x)$, m = 1, 2, … , M are weak classifiers, α_m are the values of weights obtained by the boosting algorithm.

$$G(X) = sign\left(\sum_{m=1}^{M} \alpha_m * G_m(x)\right)$$

A popular implementation of the gradient boosting tree is xgboost [2].

5 Implementation and results

5.1 Implementation details

To implement the algorithms, the Python programming language was used with machine learning package scikit-learn [23]. The processing of audio and music, as well as the extraction of features was done using the librosa library [12].

To compare the performance of the algorithms, we used a classic GTZAN dataset [26] of music tracks, which contains 10 different genres of 100 tracks each: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock.

The dataset was stratified split into 3 parts: 55% for a training set for classifier training, 15% for a validation set for searching for hyperparameters and 30% for a test set for testing the quality of the algorithms.

The quality of the algorithms will be compared by 3 criteria:

- Precision. The fraction of relevant instances among the retrieved instances
- Recall. The fraction of relevant instances that have been retrieved over the total amount of relevant instances.
- F1-score. Harmonic average of the precision and recall.

It should be noted that all the presented algorithms have an accuracy of ~ 99%, if the dataset contained less than 4 genres, and hence the features presented in this work correlate with the genre characteristics of the musical composition.

The results of the algorithms on the test set are shown in Table 1.

The best model evaluated on validation set was soft voting model between SVM with radial basis function kernel, gradient boosting tree with 200 estimators and KNN with 15 neighbors. The model achieved F1-score on test set equal to 0.78.

Analysis of the model showed that the model is most often mistaken on genres such as rock and hip-hop with the corresponding precision of 0.62 and 0.68, perhaps it is due to fact that these genres have rather wide boundaries, as shown in the article [15].

Table 1 Evaluation of algorithms on test set

Classifier	Precision	Recall	F1-score
K-neighbors	0.65	0.64	0.64
GMM	0.68	0.68	0.68
ANN	0.67	0.66	0.66
SVM	0.76	0.75	0.75
GBT	0.74	0.74	0.74
Voting	**0.78**	**0.78**	**0.78**

5.2 Comparison with existing solutions

There are number of third-party systems for music genre classification.

GenreXpose [7] is open source implementation that uses mel-frequency cepstral coefficients [13] as features and logistic regression model as multiclass classification.

Another implementation [17] provides deep neural network approach. The network architecture is a convolutional neural network, that receive vector of mel-frequency beans and applies convolution and max-pooling operations. The network consists with 3 hidden layers and fully connected layer with softmax activation for multiclass genre classes prediction.

The comparison table of average precision and evaluation time, including preprocessing, feature extraction and prediction of this approaches on test set shown on Table 2. The algorithms were tested on Apple MacBook Air 2014 Core i5, 1.8 Hz, 8 GB RAM on virtual environment.

The Table 2 shows, that our approach is more accurate, than current open source implementations, but has solid runtime due the fact, that more complex features are being calculated (rhythm and energy features).

Table 2 Comparison of different approaches

Implementations	Precision	Runtime
GenreXpose [7]	0.71	**1m 10s**
Deep learning (CNN) [17]	0.46	1m 32s
Voting of SVM, KNN, GBT	**0.78**	1m 36s

6 Website

To test the algorithms on real data, a website [27] was created using the flask library [6], which allows automatically classify the genre of music that the user uploaded, based on already existing features and the gradient boosting tree algorithm. Moreover, in case of incorrect classification, it is possible to learn the algorithm on new examples. The home page is shown in Figure 1.

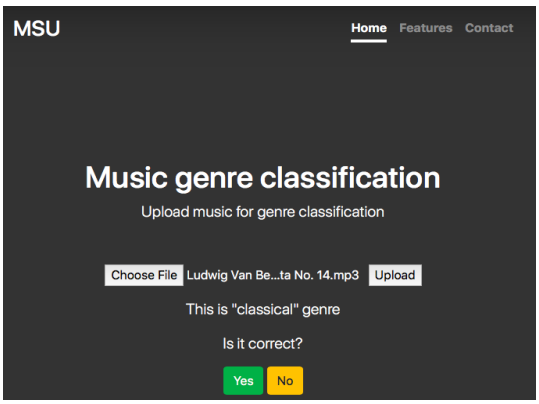

Figure 1 The home page of the website [27]

7 Conclusion

In this paper, one of the approaches to automatic music genre classification based on signal characteristics of music such as timbre, rhythm and pitch patters was studied, suggested and implemented. Modern methods of machine learning such as neural networks and gradient boosting tree were applied to these features and evaluated on the GTZAN open dataset.

In the future, it is planned to use current feature set and models on other open datasets, for example ISMIR2004 [16], which includes not only genres, but also sub-genres and MIREX [18] dataset with 22 thousand tracks. There are also ideas to improve performance of algorithms by extracting features from text and image of the music composition.

Acknowledgments. We thank Anton Bolychev, Moscow State University, for support in mathematical side of algorithms and implementation.

References

[1] Birmingham, W., Meek, C., O'Malley, K., Pardo, B., Shifrin, J.: Music information retrieval systems. Dr. Dobb's Journal, Sept. 2003.

[2] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp. 785-794, August, 2016.

[3] Dannenberg, R.B., Thom, B., Watson, D.: A machine learning approach to musical style recognition. In: Proceedings of the international computer music conference; 1997. p. 344–7.

[4] Deng, L., Yu D.: Deep learning: methods and applications. Foundations and Trends in Signal Processing, 7(3–4):197–387, 2014.

[5] Ezzaidi H, Rouat J. Automatic musical genre classification using divergence and average information measures. Research report of the world academy of science, engineering and technology; 2006.

[6] Flask. Microframework for python. https://www.palletsprojects.com/p/flask/

[7] GenreXpose. Quick music audio genre recognition. https://github.com/jazdev/genreXpose

[8] Germain, F.: The wavelet transform Applications in Music Information Retrieval. In: McGill University, December, 2009.

[9] Graves, A., Mohamed, A., Hinton, G.: Speech Recognition with Deep Recurrent Neural Networks, ICASSP 2013, Mar, 2013.

[10] Homburg, H., Mierswa, I., Moller, B., Morik, K., Wurst, M.: A Benchmark Dataset for Audio Classification and Clustering. In: ISMIR 2005, 6th International Conference on Music.

[11] Lee JH, Downie JS. Survey of music information needs, uses, and seeking behaviours: preliminary findings. In: Proceedings of the international conference on music, information retrieval; 2004.

[12] Librosa. Python package for music and audio analysis. https://librosa.github.io/librosa

[13] Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling, In: International Symposium on Music Information Retrieval

[14] Martin, K., D., Scheirer, E.D., Vercoe, B., L. Musical content analysis through models of audition. In Proceedings of the 1998 ACM Multimedia Workshop on Content-Based Processing of Music.

[15] McKay C, Fujinaga I. Musical genre classification: is it worth pursuing and how can it be improved? In: 7th Int conf on music, information retrieval (ISMIR-06); 2006.

[16] McKay, C., Fujinaga, I.: Automatic Genre Classification Using Large High-Level Musical Feature Sets. In: Conf. on Music Information Retrivial, ISMIR, 2004.

[17] Mlachmish. Music genre classification with CNN. https://github.com/mlachmish/MusicGenreClassification/

[18] Music Information Retrieval Evaluation eXchange (MIREX). http://www.music-ir.org/mirex/wiki/MIREX_HOME

[19] Perrot, D., and Gjerdigen, R.O. Scanning the dial: An exploration of factors in the identification of musical style. In Proceedings of the 1999 Society for Music Perception and Cognition pp.88

[20] Rajanna, A., Aryafar K., Shokoufandeh, A., Ptucha, R.: Deep Neural Networks: A Case Study for Music Genre Classification. IEEE 14th International Conference on Machine Learning and Applications, 2015.

[21] S. Sigtia, S. S. Dixon, S.: Improved music feature learning with deep neural networks". In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 6959–6963.

[22] Scaringella, N., Zoia, G., Mlynek, D.: Automatic genre classification of music content: a survey. In: Signal Processing Magazine, IEEE (Volume 23, Issue 2), March 2006.

[23] Scikit-learn. Machine learning in Python. http://scikit-learn.org/stable

[24] Tang, C., Chui, K., Yu, Y., Zeng, Z., Wong, K.: Music Genre classification using a hierarchical Long Short Term Memory (LSTM) model. In: International Workshop on Pattern Recognition IWPR, 2018.

[25] Tarwani M. K., Edem S.: Survey on Recurrent Neural Network in Natural Language Processing, International Journal of Engineering Trends and Technology (IJETT) – Volume 48 Number 6, June, 2017.

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

[27] Website for music genre classification. https://msumusic.herokuapp.com

[28] Zhang, W., Lei, Wenkang., Xu, X., Xing, X.: Improved Music Genre Classification with Convolutional Neural Networks. In: Interspeech, Sep, 2016.