World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:7, 2021

Machine Learning for Music Aesthetic Annotation Using MIDI Format: A Harmony-Based Classification Approach

Lin Yang, Zhian Mi, Jiacheng Xiao, Rong Li

***Abstract*—**Swimming with the tide of deep learning, the field of music information retrieval (MIR) experiences parallel development and a sheer variety of feature-learning models has been applied to music classification and tagging tasks. Among those learning techniques, the deep convolutional neural networks (CNNs) have been widespreadly used with better performance than the traditional approach especially in music genre classification and prediction. However, regarding the music recommendation, there is a large semantic gap between the corresponding audio genres and the various aspects of a song that influence user preference. In our study, aiming to bridge the gap, we strive to construct an automatic music aesthetic annotation model with MIDI format for better comparison and measurement of the similarity between music pieces in the way of harmonic analysis. We use the matrix of qualification converted from MIDI files as input to train two different classifiers, support vector machine (SVM) and Decision Tree (DT). Experimental results in performance of a tag prediction task have shown that both learning algorithms are capable of extracting high-level properties in an end-to end manner from music information. The proposed model is helpful to learn the audience taste and then the resulting recommendations are likely to appeal to a niche consumer.

***Keywords*—**Harmonic analysis, machine learning, music classification and tagging, MIDI.

I. Introduction

IN the field of MIR over the last few years, the content-based approach [1] becomes more prevailing as it exploits the similarities of music content, serving as a supplement approach to those issues like popularity bias and the cold-start problem facing the conventional collaborative filtering method. The success of the content-based approach relies on finding ways of automatically classifying annotating music genre, mood or other descriptors. Basically, performing music classification and tagging tasks requires transforming diverse characteristics extracted from the raw music like these acoustic features and musical information into computer-interpretable numerical representations. But as discussed in [2], most of the effective music classification and recommendation systems are focusing on acoustic features extraction via digital signal processing with tailored feature-learning design, while musical information extraction still remains in the hand engineering stage. It becomes quite necessary to consider specific music elements when developing learning algorithms for MIR. So, in our study, it is aimed to make music aesthetic classification and tagging using the particular harmonic features obtained by MIDI format conversion to qualification matrix. Although the CNNs, the well-known two-dimensional (2-d) image classifier, have been successfully borrowed and applied to music genre classification and prediction with achievement of more accuracy than the traditional approach, such a deep learning technique is not able to outperform other machine learning methods significantly in numerous MIR tasks [3]. Some other learning techniques like SVM could also be trained as an alternative feature extraction and classifier, with comparable performance [4]. Here we would use the matrix of qualification converted from MIDI files reflecting the harmonic features as input to train two different classifiers, SVM and Cart-DT. Then the output shown by the Boolean tag values will determine music aesthetic classification and prediction.

Lin Yang, Zhian Mi, Jiacheng Xiao, and Rong Li are with the Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu Province, P. R. China, 215123 (phone: +86 0512-81884948; e-mail: Rong.Li@xjtlu.edu.cn).

II. Methods

In the MIDI file, the pitch, length, and strength of each note are clearly defined. MIDI format itself develops a standard symbolic representation for music as musical scores. Research on genre classification using MIDI format would date back to 20 years ago, see for example, [5], [6] and [2]. We firstly represent the music audio signal by music score in MIDI format as traditional analysis, then convert the MIDI file into matrix of qualification containing key information of the harmony logic, timbre and rhythm texture in the structure of a one - part musical form, as shown in Fig. 1.

Since any musical structure can be considered as the combination of several phrases in one-part form, this process enables various music pieces with different musical forms to be converted as the combination of different matrices, which would demonstrate the versatility of the proposed model. It should be noted that all the contents of this paper are based on tonal system.

A. Harmonic Analysis

The technique of harmony is the most crucial part in traditional music and how to quantify the harmony is a challenging problem. In music theory, the interval in a musical scale is defined with the term like ‘perfect’, ‘major’, ‘minor’, ‘augmented’, and ‘diminished’, which can be described mathematically by some positive integers, also known as value of distance (VOD). Clearly, the larger the VOD is, the further the two notes are apart [7]. The following are some examples of the most widely used natural modes, and each of them is labeled with largest VOD centered around tonic C:

Natural Major:
[C − B = 5]
Natural Minor:
[C − Ab= −4]
Mixolydian:
[C−Bb = −2; C − E=4]
Dorian:
[C − A = 3; C − Eb = −3]
Lydian:
[C − F# = 6]
Phrygian：
[C − Db = −5]
Gong:
[C − E =4]
Zhi:
[C − A = 3].

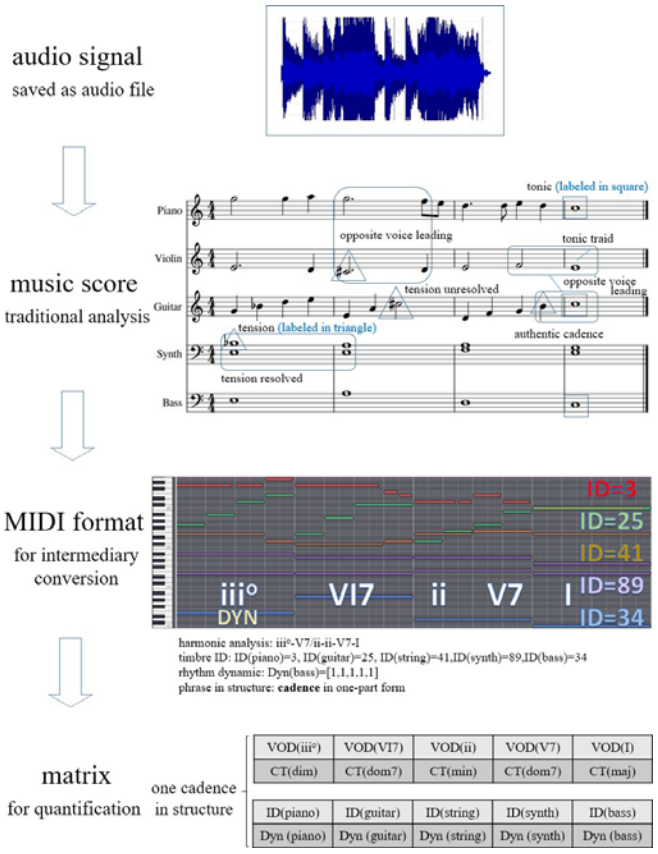


Fig. 1 The process flow chart

The concept of VOD can be further extended for the classification of harmony to cope with music pieces involving more complicated harmony structure by setting up a function called chord tensity (CT). Similar to the definition of the interval, the harmony structure, generally categorized as ‘major’, ‘minor’ or ‘diminished’, could be quantified in the way shown by Fig. 2. The first and second row of the matrix for quantification corresponds to the typical harmonic analysis, VOD and CT, respectively, of the musical phrase.

Structure	Maj	Min7	Min	Maj7	Dim	Majb7	Aug	MinM7	Dim7
Chord tension	1	1.5	2	2.5	3	3	4	4	5

Fig. 2 The CT for quantifying harmony structure

B. Form Structure

The music structure contains only one single section with obvious exposition, development and recapitulation with clear harmony progression and cadence is called the single one-part form, and this is the smallest for a complete musical form structure.

C. Classification by Using Machine Learning

Some classification algorithms which would be used are discussed in this section.

SVM is essentially used to geometrically classify data sets to two distinct categories by optimizing boundary region, namely, decision surface. It usually works with one hidden layer and is a typically effective shallow learning method for solving binary classification problem. When dealing with multi-class data sets, the techniques such as one-against-one and one-against-all are frequently used [8], [9].

DT is one of deep learning algorithms, which is much easier to understand and visualize. There are mainly three different kinds of ways in dealing with data in DT method: ID3, C4.5 and Cart. ID3 and C4.5 involve information entropy calculation while Cart relies on optimization for Gini index [10], [11].

III. Experimental Results

In this section, we use the sample data set of 30 MIDI music pieces to train two different classifiers, SVM and DT and report the prediction accuracy on music aesthetics of test data set of 10 MIDI music pieces. In Fig. 3, it is listed that a few examples of 30 inputs in the sample set with the associated qualification matrix of harmonic classification and Boolean tag values representing three music aesthetics terms: happy, mystical and intense.

Sample set for input

data of Matrix 1 (total of 30)

Input 1	I	Vi	ii	V	I	maj	min7	min	majb7	maj
Input 2	I	Ii	IV	viio	I	maj	min7	maj	dim	maj
Input 3	I	IV	ii	V	VI	maj	maj	min	maj	maj
Input 4	i	v	bVII	iv	i	min	min7	maj	min7	min7
Input 5	i	bIII	bVII	ii	i	min	maj	maj7	min	min
Input x	…	…	…	…	…	…	…	…	…	…

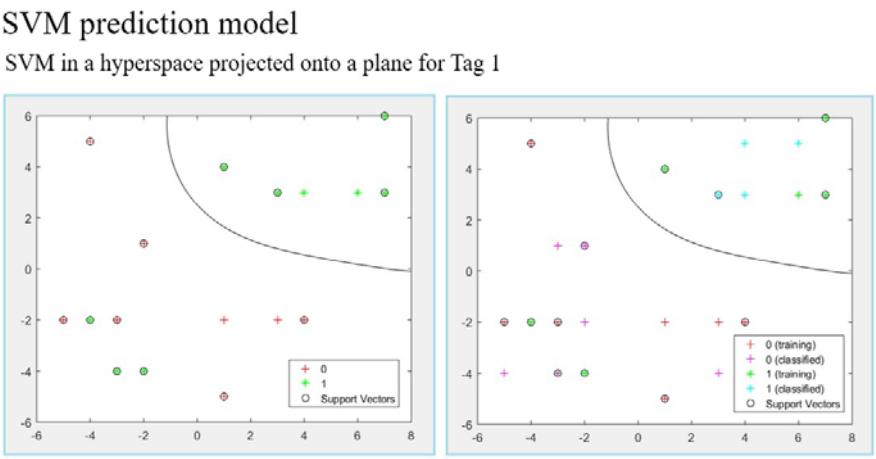
tags known

	Tag 1	Tag 2	Tag 3
Input 1	1	0	0
Input 2	0	0	1
Input 3	1	0	1
Input 4	0	1	1
Input 5	0	1	0
…	…	…	…

Fig. 3 The examples of sample data set

Fig. 4 visualizes the SVM classification and prediction model for tag 1 through projection the 10-dimension hyperspace onto a plane (models for tag 2 and 3 are similar). Linear kernel is used for SVM classifier. Fig. 4 (a) depicts the training model with 30 inputs from the sample set. And 10 more inputs from the test set are added into prediction model shown by Fig. 4 (b). The classification and prediction results are

differentiated with various colored signs.



(a) (b)

Fig. 4 The SVM classification and prediction for Tag 1

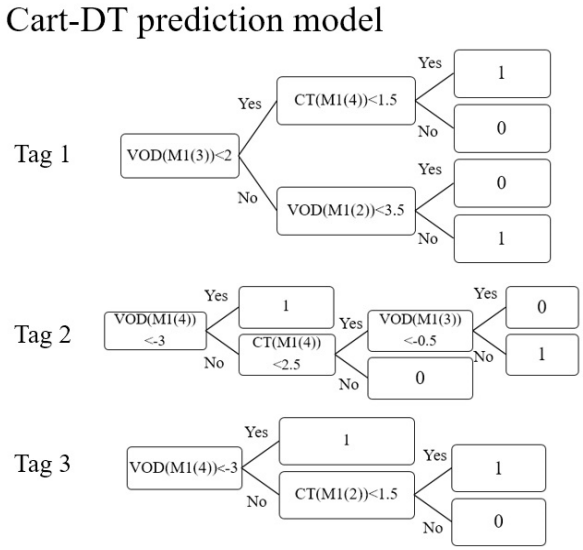


Fig. 5 The Cart-DT classification and prediction model

An alternative training model is generated by Cart-DT method, which is shown in Fig. 5. Compared to SVM approach, this model is much more interpretable while providing competitive results. By setting up a distinctive DT for each tag, one can clearly identify the most critical features that affect the decision-making on the classification of inputs. Then Fig. 6 summarizes the prediction results for the annotation task in the test set. The false predictions in tag values are highlighted in red. According to the table-n, the accuracy of SVM prediction results is 70%, 60% and 90% for each individual tag and that of Cart-DT are 80%, 70% and 70%, respectively. Both methods give the same overall prediction accuracy at rate of 73.33%. From users' experience, the prediction performance is decent in all the inputs but the input 3 and 5. Looking into the musical composition techniques of the two pieces, both involve complex harmony resolution such as deceptive cadence or altered cadence. This may increase the difficulty of prediction and eventually lead to false prediction.

Test set for prediction

data of Matrix 1 (total of 10)

Input 1	I	vi	IV	V	I	maj	min	maj	majb7	maj
Input 2	I	vi	iii	V	I	maj7	min	min	maj	maj
Input 3	I	ii	IV	V	VI	maj	min	maj7	maj	maj
Input 4	I	bVII	V	iv	I	maj	maj	maj	min	maj
Input 5	I	vii	iii	II	I	maj	min7	min7	maj	maj
Input x	…	…	…	…	…	…	…	…	…	…

Model prediction result

user's potential experience SVM prediction Cart-DT prediction

	Tag 1	**Tag 2**	**Tag 3**	**Tag 1**	**Tag 2**	**Tag 3**	**Tag 1**	**Tag 2**	**Tag 3**
Input 1	1	0	0	1	0	0	1	0	0
Input 2	1	0	0	1	1	0	1	1	0
Input 3	1	0	1	0	1	0	0	1	0
Input 4	0	1	1	0	1	1	1	1	1
Input 5	1	0	1	1	1	0	0	1	1
Input 6	0	1	1	0	1	1	0	1	1
Input 7	0	1	1	0	1	1	0	1	1
Input 8	0	1	1	0	1	1	0	1	1
Input 9	0	0	1	1	0	0	0	1	1
Input 10	0	1	1	0	1	1	0	1	1

Fig. 6 The prediction results for both SVM and DT methods

IV. Further Discussion

The aesthetics on music may be subject to culture background and personal taste, therefore it is quite necessary to make further improvement on the predictive result of the learning model. One way to solve this problem is to adjust the weight of different parameter values as small as possible until the result as desired. Fig. 7 shows an example of VOD calibration in linearity, and it can be further modified and customized for better accuracy. For instance, we can define a new nonlinear function instead of original linear VOD function in order to approximate the varied music tastes of different users. That is, by setting up more nonlinear functions and adjusting the parameters in the functions, the model can learn

the tastes of various users for improving the prediction results of tonal music.

V. Conclusion

In this paper, we proposed a music aesthetic tagging model to learn audience's taste and the model design is much intuitive. The two learning methods, SVM and DT have been explored in the proposed model exploits. Both methods provide comparable results. In future works, we will continue to study on extracting high-level features like timbre, texture together with harmony analysis via a couple of deep learning architectures design. For example, in terms of typical timbres, we could apply the Fast Fourier Transform (FFT) to generate intuitive graphs which show the basic quantification of timbres (see Fig. 8, an example of violin sample of timbre). Running FFT on hundreds of typical timbres, learning models can be developed to find out the patterns and regularities. In this way, it is expected to produce better results than using the conventional category IDs of timbres alone. Moreover, lyrics are believed extremely important especially for commercial music because they make it easier for the audience to understand what the authors want to express. From the aspect of learning and prediction, lyrics also matters; while lyric is not essential comparing with other features of music mentioned in the literature review [12]. In addition, the scale of dataset used in research is always a matter of concern [13]. The industrial-scale datasets are usually expensive to acquire and restricted to publish because of licensing issues. In our future work, we are now actively applying for government-supported grants and looking for appropriative cooperation with industry participants.

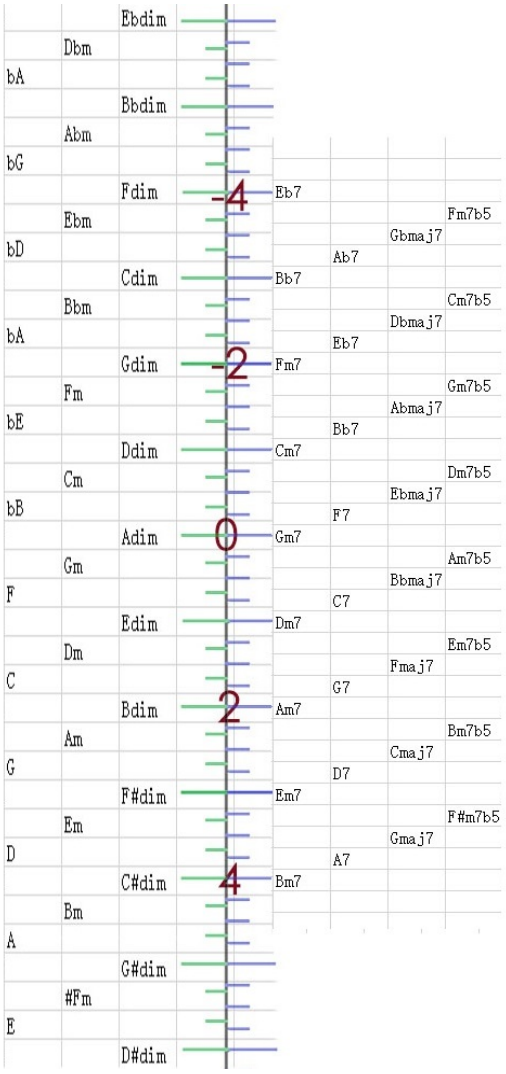


Fig. 7 An example of VOD calibration

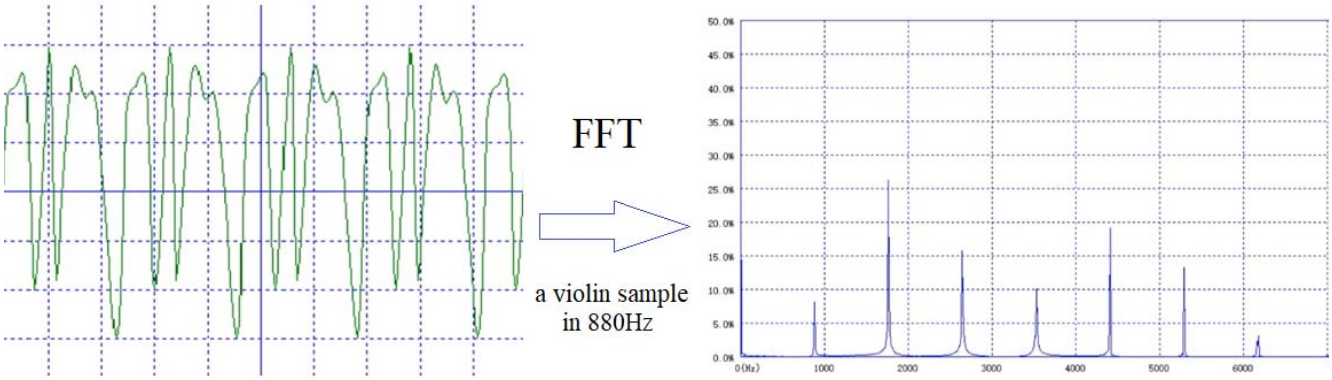


Fig. 8 An example of violin sample of timbre

Appendix

Software Used:
Audacity
Matlab
MISO
Musescore
MusicScope
Sonic Visualiser
Studio One
Tracktion Waveform

ACKNOWLEDGMENT

We would like to express our gratitude to Yushuo Liu for generously sharing ideas to improve the quality of the paper. We also would like to express special thanks to all reviewers for their time and efforts to improve the quality of this paper.

REFERENCES

[1] Rodrigo Capobianco Guido Antonio Jose Homsi Goulart and Carlos Dias Maciel. 2012. Exploring different approaches for music genre classification. The Egyptian Informatics Journal 13, 2 (2012). https://doi.org/10.1016/j.eij.2012.03.001

[2] Yaslan Y. Cataltepe, Z. and A. Sonmez. 2007. Music Genre Classification Using MIDI and Audio Features. (2007).

[3] N. Cristianini and J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

[4] Cambridge Press, New York, NY. https://doi.org/10.1017/CBO9780511801389

[5] Edward Dunne and Mark Mcconnell. 1999. Pianos and Continued Fractions. Mathematics Magazine 72, 2 (1999), 104–115. https://doi.org/10.1080/0025570X. 1999.11996712

[6] A. Elbir and N. Aydin. 2020. Music genre classification and music recommendation by using deep learning. Electronics Letters 56, 12 (2020), 627–629. https://doi.org/ 10.1049/el.2019.4202

[7] Cory Mckay and Ichiro Fujinaga. 2004. Automatic genre classification using large high-level musical feature sets. In In Int. Conf. on Music Information Retrieval, ISMIR 2004. 525–530.

[8] J. Nam, K. Choi, J. Lee, S. Chou, and Y. Yang. 2019. Deep Learning for Audio- Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach. (2019).

[9] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep Content-Based Music Recommendation. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (Lake Tahoe, Nevada) (NIPS'13). Curran Associates Inc., Red Hook, NY, USA, 2643–2651.

[10] Flores M.J. Ramírez, J. 2019. Machine learning for music genre: multifaceted review and experimentation with audioset. (2019).

[11] André C.P.L.F de Carvalho Rodrigo C. Barros and Alex A. Freitas. 2015. Automatic Design of Decision-Tree Induction Algorithms. Springer. https://doi.org/10.1007/ 978-3-319-14231-9

[12] George Tzanetakis, Andrey Ermolinskyi, and Perry Cook. 2003. Pitch Histograms in Audio and Symbolic Music Information Retrieval. (2003).

[13] lan H. Witten, Eibe Frank, and Mark A. Hall. 2011. Chapter 4 - Algorithms: The Basic Methods. In Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)