Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230273

A comparison of machine learning algorithms for music genre recognition

Yundi Xu

SDU-ANU Joint Science College, Shandong University, Weihai, Shandong, 264209, China

202000700021@mail.sdu.edu.cn

Abstract. Music information retrieval includes the classification of music genres as a key component. However, due to the complexity of music composition, existing artificial classification cannot achieve accurate results. In the music genre classification scenario, machine learning can effectively process large and complex data, bringing more accurate and personalized results for music recommendation. This article will start from three articles and use GTZAN as a data set to compare 3 traditional machine learning methods, including SVM, random forest and logistic regression, together with 2 deep learning models, involving convolutional and feedforward neural networks (CNN and FFNN) in music genres. classification accuracy. The results show that the classification accuracy will be affected by multiple factors such as the characteristics of the audio, whether data processing is performed and so on. In addition, the performance of deep learning has not been significantly better than traditional machine learning as expected. At the same time, the accuracy of CNN and FFNN is heavily dependent on whether the audio is processed into a spectrogram.

Keywords: machine learning, deep learning, music genre classification.

1. Introduction

Music information retrieval includes the classification of music genres as a key component [1]. The composition of music is very complex due to the diversity of accompanying instruments, the uniqueness of the singer's voice, and the differences in elemental harmonies. Earlier music classification needed to be performed manually, but due to the complexity of the features contained in audio classification, it was difficult to extract the appropriate audio features, making the automatic music classification results not particularly accurate [2,3]. Hence, the development of music audio information retrieval systems would benefit greatly from a technique that can automatically categorize musical genres.

Nowadays, machine learning-based music genre classification techniques have many commercial uses [4]. For example, music software can be designed to use the technology to create references for song classification so that users can easily hear their favorite music immediately. It can also identify groups of people who like to listen to specific songs and constantly recommend music of their favorite genres for such groups. In addition, it can help users filter out songs they do not like to avoid causing discomfort for them.

Machine learning is being widely known and applied for its advantages of speed, low cost and high accuracy [5]. In the music genre classification application scenario, machine learning can effectively

DOI: 10.54254/2755-2721/8/20230273

handle large and complex data to bring more accurate and personalized results for music recommendation. Deep learning, as a special machine learning technique, is used to quickly analyze various types of music information and achieve rich music recommendations by simulating the process of building automatic feature extraction, clustering labeling and model training in human brain neural networks.

Before 2002, the accuracy of people for music genre classification was about 70% [6]. Using the Gaussian mixture model (GMM) together with k-nearest neighbors (k-NN), Tzanetakis and P. Cook identified 10 musical genres in GTZAN in 2002 and achieved an accuracy of 61% [6]. With the refining and enhancement of machine learning technologies over the last decade, the accuracy of music classification has increased. Among them, Christine et al. used the CNN method in 2017 based on the GTZAN database to obtain the highest accuracy rate of 91% using the two late fusion systems FUSION1 and FUSION2 features, which once proved the high practicality of CNN in music genre classification [7]. In order to determine which classification method performs the best, using the same dataset, the author analyzes the classification accuracy of three standard machine learning algorithms as well as two deep learning classifiers.

2. Method

This section introduces the dataset used in the study, as well as the three classic machine learning classification techniques of support vector machine, logistic regression, and random forest, along with two deep learning techniques, including CNN and FFNN. A processing technique called Mel Spectrogram will also be described.

2.1. Dataset: GTZAN

The music genre classification results in this research were based on GTZAN dataset. As a publicly available dataset, it is extensively leveraged to validate the effectiveness of algorithms for music genre recognition (MGR). The dataset collects 10 music genres, including blues, disco, hip-hop, reggae, classical, metal, country, jazz, rock, and pop. Each has 100 audio files where each of them contains a song with 30 seconds. All tracks are saved as ".wav" format, with basic configurations demonstrated in Table 1

Table 1. Contents of GTZAN.

Item	Data
Genre	10
Length	30s
Sample rate	22,050 Hz
Audio files	16 bits
Song mode	Mono
Total	1000

2.2. Logistic regression

As a widely leveraged supervised algorithm, logistic regression is capable of calculating the probability in the dichotomous case. It first fits a decision boundary and then establishes a probability relationship between this border and the classification. While the calculation and derivation procedure are similar to those of regression, its primary use is to tackle binary classification issues (and also multiclassification problems).

2.3. Support vector machine (SVM)

SVM is broadly applied to address supervised classification issues. It aims at seeking the largest interval between features of different categories. It is rooted on a linear classifier, but optimized in different manner, via maximizing the interval is SVM's learning approach.

DOI: 10.54254/2755-2721/8/20230273

Given a p-dimensional vector as the input sample of the SVM, a p-1 dimensional hyperplane would be used to divide these points. Yet, a sizable number of hyperplanes could be able to classify data. The optimal hyperplane is likely to be the one that separates the two classes by the largest margin. SVM decides on the hyperplane and then optimizes the distance to the hyperplane for the data points that are closest to it.

2.4. Nonlinear support vector machine

By utilizing kernel techniques and soft margin maximization, a nonlinear SVM may be taught when the training data is linearly inseparable. The kernel approach is separated into two steps: To learn the model in the new space using the training data, first conduct a transformation to transfer the data from the original space to the new one. The nonlinear SVM mentioned in this article is processed by kernel techniques. The kernel functions used in this article are polynomial kernel and RBF kernel.

2.5. Random forest

A random forest method uses the concept of ensemble learning to integrate many trees. Its fundamental component is a decision tree. Random selection of features and samples is referred to as random: Each tree is generated from the complete training sample set using a predetermined number of samples before a fixed number of features are chosen to form each decision tree in the random forest. A forest is a model made up of multiple decision trees. In its design, each tree algorithm will deliver a prediction result, i.e., n predicted probabilities will be generated by n trees. Afterwards, a voting will be conducted, where the most predicted outcome will be regarded as the final prediction when the random forest incorporates all the classification voting results.

2.6. Convolutional neural network (CNN)

CNN is a widely used artificial intelligence model, which leverages the neural network as the basic constructive unit, and its three layers—convolutional, pooling, and fully connected—can be separated into several categories. The convolution layer plays the main role to extract features. Features are downsampled by the pooling layer without affecting the recognition outcomes. The fully connected layer assigns a classification to the extracted features.

Convolution, which is the superposition of two functions applied to an image, can be thought of as applying a filter to the image in order to identify specific features, as one needs to locate numerous features in order to recognize a specific object. The pooling layer reduces the data size and has no impact on the result of recognition, i.e., it downsamples the output of the convolutional layer. The fully-connected layer's primary function is categorization. These summary characteristics are classified in the fully connected layer using the features obtained from the preceding convolution and pooling layers.

2.7. Feed-forward neural network (FFNN)

The connections between the nodes in a feed-forward neural network do not create a loop. Each neuron in a feed-forward network is only connected to the neurons in the layer preceding it. There is no feedback between the layers; instead, information is received from the previous layer and output to the following layer in a single direction

2.8. Mel spectrogram

Due to the structure of human hearing, people pay special attention to specific frequencies and only let certain frequencies of sound pass through. In order to process sounds realistically, people use logarithmic scales by Mel scales and decibel scales when working with the frequency and amplitude of the data. The Mel spectrogram can largely preserve the information needed for the human ear to understand the original speech, which makes it particularly important for deep learning methods. Converting raw audio into spectrograms requires the use of short-time Fourier transform methods.

DOI: 10.54254/2755-2721/8/20230273

3. Result

The algorithms studied in this paper are mainly from three articles, which are written by Yukta Padgaonkar et al., Derek A. Huang et al., and Dhevan S. Lau and Ritesh Ajoodha [7-9]. Among them, the results of SVM, shown in Table 2, were obtained from Yukta Padgaonkar et al [8]. Dhevan S. Lau and Ritesh Ajoodha provided the random forest and linear regression findings shown in Table 3 [7]. The results of CNN and FFNN, displayed in Table 4, were obtained from Derek et al [9].

Table 2. Comparison of SVM with different kernels.

Algorithm	Accuracy with all features	Algorithm with top 20 features
SVM with polynomial kernel	69.0%	66.0%
SVM with RBF kernel	74.0%	65.5%

Table 3. Comparison of conventional machine learning algorithms.

Algorithm	Accuracy with 30-second input features	Algorithm with 3-second input features
Logistic Regression	66.5%	67.5%
Random Forests	74.5%	80.3%

Table 4. Comparison of neural network-based algorithms.

Algorithm	Accuracy with data processing	Algorithm without data processing
Convolution Neural Network	82.0%	25.0%
Feed-forward Neural Network	54.0%	33.0%

The tabular results show that when leveraging polynomial and RBF kernel in SVM, the performance based on all features in the dataset exceeds the performance based on the first 20 features, and the accuracies using all features are 69% and 74%, respectively. When studying logistic regression and random forest methods, the accuracy using 3-second input feature set was higher than that using 30-second input feature set in both cases, with 67.5% and 80.3% accuracy using 3-second input feature set, respectively. The results of studying CNN and FFNN methods show that the results with data processing are significantly better than those without data processing, and the accuracies with data processing are 82% and 54%.

4. Discussion

The accuracy calculated by different authors using the same machine learning model can be influenced by many factors.

To begin with, when building a machine learning model, the selection of features is important for effective prediction of results. The methods leveraged for the selection of featuers are: filter, wrapper, embedded and hybrid methods [10]. Yukta et al. adopt the Random Forest Importance approach. It is a kind of embedded models [8]. By this method, the authors extracted the top 20 features that have the most influence on the output labels. At the same time the authors trained another set of classifiers with all the features. In the process of comparing the results, it was found that the accuracy of including all features was always higher than the accuracy of the top 20 features, regardless of whether it was LR or SVM. This indicates that the selection of features affects the final result. Moreover, Derek et al. limited the window to 2 seconds in order to extract features and found that 44100 features allowed the length of the audio samples to be balanced with the dimensionality of the data [9]. In the article by Dhevan and Ritesh, to enhance the quantity of training data, the authors split the dataset into two types: 1000 30-second audios and 10,000 3-second audios, from which 57 features were extracted for the study [7].

DOI: 10.54254/2755-2721/8/20230273

Besides, the validity of the results is significantly influenced by the pre-processing of the data. By converting the raw audio into a graphical form, Derek et al. showed that all models performed significantly better after the data transformation. Mel spectrogram and comparing it with the accuracy of the models without data pre-processing [9]. In addition, it is possible that the training set to test set ratio has an influence on the findings. Dhevan S. Lau and Ritesh Ajoodha and Yukta et al. chose a training set to test set ratio of 80:20 [7].

Surprisingly, both the conventional machine learning approach and the deep learning method have benefits and limitations of their own, and neither one has a clear edge over the other. The conclusions with excellent performance are contradictory. CNN is a method specially used for image feature recognition, but in this study, it performs similarly to random forest under 3-second input feature set. Therefore, further research on the optimization of CNN and FFNN is necessary in the future. In addition, it is hoped that more machine learning methods can be compared in the future. For example, KNN, as a relatively old machine learning algorithm, has relatively good classification results in previous literature. For example, some RNN methods (GRU, LSTM, etc.), the advantage of RNN is that it can record past data and use past information to predict the current state. Ultimately, it is hoped that a mix of classical machine learning and deep learning may be investigated in order to deal with different music genres. To increase accuracy and efficiency, neural network features are employed to extract data set characteristics, and typical machine learning methods are used for training and testing.

5. Conclusion

Using the GTZAN data set, this article compares the accuracy of two deep learning models CNN and FFNN, with three standard conventional classification algorithms: support vector machine, logistic regression, and random forest. It is concluded that the accuracy of SVM when all features are selected is higher than the accuracy of selecting the first 20 features. Moreover, when testing with logistic regression and random forest, the performance is higher than that of using the 30-second input feature set. The accuracy validated on the second feature set, accuracy results of CNN and FFNN after data processing is significantly higher than that without data processing. It is shown that the accuracy will be affected by many factors such as feature selection, whether data processing is performed or not. At the same time, in this paper, the deep learning method did not perform significantly better than the traditional machine learning method as expected, and the performance of the two deep learning approaches is heavily influenced by whether or not the audio is converted into a spectrogram. In the future, it is necessary to improve CNN Further research on the optimization of deep learning methods. Meanwhile, it is important to research deeper machine learning techniques that combine classical machine learning with music genre categorization.

References

[1] Neumayer, R., & Rauber, A. (2007). Integration of text and audio features for genre classification in music information retrieval. In Advances in Information Retrieval: 29th European Conference on IR Research, 724-727.

[2] Zhang, J. (2021). Music feature extraction and classification algorithm based on deep learning. Scientific Programming, 2021, 1-9.

[3] Doraisamy, S., Golzari, S., Mohd, N., Sulaiman, M. N., & Udzir, N. I. (2008). A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, 331-336.

[4] McKay, C., & Fujinaga, I. (2006). Musical genre classification: Is it worth pursuing and how can it be improved? 101-106.

[5] Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., & Serra, X. (2015). Acousticbrainz: a community platform for gathering music information obtained from audio. 16th International Society for Music Information Retrieval Conference, 786-792.

[6] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), 293-302.

DOI: 10.54254/2755-2721/8/20230273

[7] Lau, D. S., & Ajoodha, R. (2022). Music genre classification: A comparative study between deep learning and traditional machine learning approaches. In Proceedings of Sixth International Congress on Information and Communication Technology, 4, 239-247.
[8] Padgaonkar, Y., Gole, J., & Tekwani, B. (2022). Music Genre Classification using Machine Learning, 139-143.
[9] Huang, D. A., Serafini, A. A., & Pugh, E. J. (2018). Music Genre Classification. CS229 Stanford, 1-6.
[10] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.