Comparison and analysis of accuracy of traditional random forest machine learning model and XGBoost model on music emotion classification dataset

Jiulin, Song*
College of Engineering, Bucknell University, One Dent Drive, Lewisburg, PA 17837, USA
js131@bucknell.edu

ABSTRACT

The Turkish Music Emotion Dataset in the CL database is a dataset for music emotion classification, which contains many music samples and their corresponding emotion labels. This paper presents a comparative analysis of the performance of the traditional random forest machine learning model and the XGBoost model on the given dataset. The findings indicate that the traditional random forest machine learning model outperforms the XGBoost model in terms of accuracy, accuracy, and recall rate. The traditional random forest machine learning model achieves an accuracy of 80.8%, whereas the XGBoost model achieves an accuracy of 75%. The recall rate of the traditional random forest machine learning model is 80.8%, whereas the recall rate of the XGBoost model is 77.2%. The traditional random forest machine learning model achieved an F1 score of 80.5%, whereas the XGBoost model attained an F1 score of 75.3%. The comprehensive evaluation of the XGBoost model outperforms the traditional random forest machine learning model. Generally, the traditional random forest machine learning model demonstrates strong performance in the domain of music emotion classification. It exhibits favorable qualities such as robustness and interpretability, and possesses the ability to effectively handle noise and missing data. The XGBoost model, on the other hand, can be trained and predicted quickly, with high accuracy and generalization ability. Therefore, in practical application, it is necessary to select the appropriate model according to the specific situation, and optimize and adjust it to obtain the best classification effect.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning approaches.

KEYWORDS

Machine learning classification, XGBoost, Prediction accuracy

ACM Reference Format:
Jiulin, Song*. 2023. Comparison and analysis of accuracy of traditional random forest machine learning model and XGBoost model on music emotion classification dataset. In *2023 4th International Conference on Machine Learning and Computer Application (ICMLCA 2023), October 27–29, 2023, Hangzhou, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3650215.3650340

ICMLCA 2023, October 27–29, 2023, Hangzhou, China

ACM ISBN 979-8-4007-0944-9/23/10
https://doi.org/10.1145/3650215.3650340

1 INTRODUCTION

The Turkish Music Emotion Dataset in the UCL database is a dataset for sentiment analysis research that contains songs from Turkish traditional music and their corresponding emotional labels. The dataset comprises a comprehensive collection of 1,000 songs, each accompanied by an emotion label denoting various affective states such as happiness, sadness, anger, and relaxation. The dataset was constructed to represent a discrete model comprising four distinct classes, namely happy, sad, angry, and relaxed. In order to curate the dataset, a diverse range of Turkish music was chosen, encompassing both verbal and non-verbal compositions. The database contains a set of 100 musical works, with an equal distribution of samples across each class. The original dataset consisted of 400 samples, with each sample having a duration of 30 seconds.

The collection process of this dataset is completed by manual labeling, and professional musicians listen to and analyze each song, and then give the corresponding emotional label. Such labeling can ensure the accuracy and credibility of the data. In addition to the sentiment label, the dataset also provides some other information, such as the name of the song, the artist, the composer, and so on. This information is also very helpful for researchers to carry out more in-depth analysis and exploration. This dataset can be used for training and testing of sentiment analysis algorithms, as well as exploration of music emotion research. Through the study of this dataset, we can have a deeper understanding of the emotional characteristics of traditional Turkish music and explore the application of emotion analysis algorithms in the field of music.

The Turkish Music Emotion Dataset in the UCL database is a dataset for sentiment analysis research that contains songs from traditional Turkish music [1]. In the field of machine learning, researchers have developed some sentiment analysis algorithms based on machine learning through the analysis and research of this data set, and achieved some results [2, 3]. For example, Guo Feng used the support vector machine (SVM) algorithm to classify the emotion of the dataset [4], and achieved good classification effect. Xia Yu; Deep learning algorithms are used, such as convolutional neural networks (CNN) [5]. Das Sanchali used recurrent neural networks (RNN) [6] and also achieved good results. Some researchers have used traditional machine learning algorithms, such as support vector machines and naive Bayes, to carry out sentiment analysis [7, 8]. Other researchers use deep learning algorithms, such as convolutional neural networks, recurrent neural networks, etc. [9, 10], to conduct sentiment analysis. These algorithms have achieved many successful applications in the field of emotion analysis. The

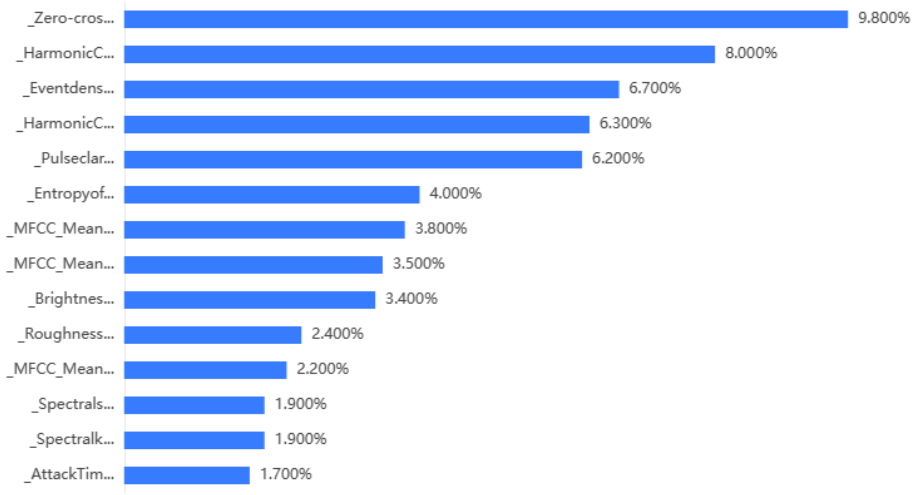

Figure 1: Feature dimension reduction. (Photo credit : Original)

application of these algorithms can help us more accurately analyze and understand the emotional characteristics of traditional Turkish music, and also provide some references and enlightenment for the development of the field of music emotion analysis [11]. In addition, these algorithms can also be applied to other fields of sentiment analysis, such as text sentiment analysis, image sentiment analysis, etc.

2 HANDLING OF DUPLICATE AND OUTLIER VALUES

In data preprocessing, the processing of duplicate values and outliers is a very important step. Duplicate values refer to the presence of multiple identical records in the data set, while outliers refer to some extreme value or unreasonable data in the data set that may adversely affect the accuracy of the model.

In terms of handling duplicated values, we can use the duplicated() function in the pandas library to determine if there is a duplicate value in the dataset and use the drop_duplicates() function to remove the duplicate value. For the handling of outliers, this paper adopts the 3σ principle. The 3σ principle means that the data in the data set is normalized according to the mean and standard deviation, and then the data above 3 standard deviations is treated as an outlier.

3 DATA STANDARDIZATION

Data standardization refers to scaling the data in a data set to a certain scale so that the data falls within a specific range. This paper uses Z-score standardization to standardize the data.

Z-score normalization is the process of subtracting the mean from each data point in the data set and dividing it by the standard deviation to convert the data into a standard normal distribution.

4 DATA STANDARDIZATION

4.1 Pearson correlation analysis

The Pearson correlation coefficient is a statistic used to measure the degree of linear correlation between two variables, and its mathematical concept is: Let X and Y be two random variables with covariance cov (X,Y) and variance Var(X) and Var(Y), respectively. The Pearson correlation coefficient ρ of X and Y is defined as:

$$rho = cov\,(X, Y)\,/\sqrt{(Var\,(X)} * \sqrt{(Var\,(Y)} \tag{1}$$

Where cov (X,Y) represents the covariance of X and Y, and Var(X) and Var(Y) represent the variance of X and Y, respectively.

4.2 Feature dimension reduction

Feature dimensionality reduction refers to reducing the feature dimension of the data set by some method in order to reduce the complexity of the model and improve the accuracy of the model. In this paper, the correlation coefficient of each feature is calculated using correlation analysis, and then the redundant features with high linear correlation are screened out. Select the first ten features with greater weight, and the results are as shown in the Figure 1

5 COMPARISON BETWEEN TRADITIONAL RANDOM FOREST ALGORITHM AND XGBOOST ALGORITHM

The data set was loaded and divided according to the ratio of 7:3. The Random forest classification model and XGBoost classification model were used to train the data set, and the accuracy, accuracy, calculation recall rate and F1 score of the models were compared to compare the classification effects of the two models.

A Random Forest is an ensemble learning algorithm composed of multiple decision trees. In the random forest, the training data of each decision tree is randomly extracted from the original data set, and the features of each node are also randomly selected from all the features for calculation, thus reducing the overfitting degree of the model to the training data and improving the generalization

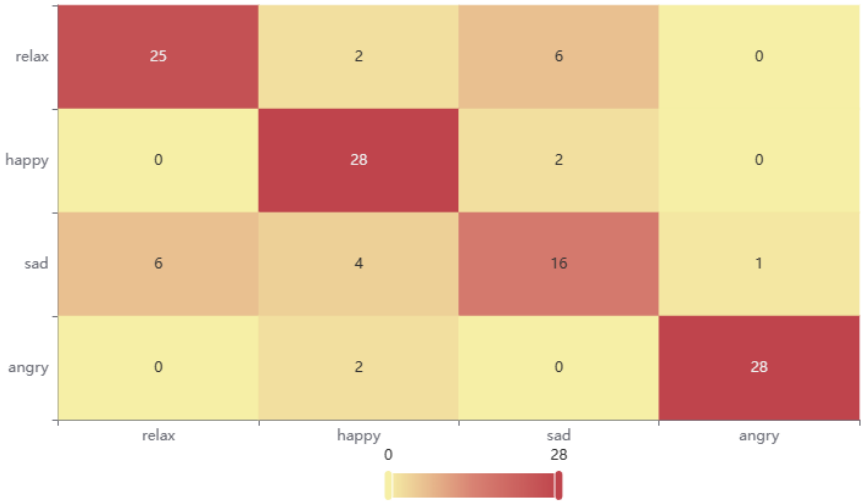

Figure 2: Random forest confusion matrix. (Photo credit: Original)

Table 1: Model evaluation

	Accuracy	Recall	precision	F1
Training set	1	1	1	1
Test set	0.808	0.808	0.808	0.805

Table 2: Model evaluation

	Accuracy	Recall	precision	F1
Training set	1	1	1	1
Test set	0.75	0.75	0.772	0.753

ability of the model. Random forest model has good performance in processing high-dimensional data and nonlinear feature fitting.

XGBoost (eXtreme Gradient Boosting) is an optimization implementation of Gradient Boosting Decision Tree (GBDT). Compared with the traditional GBDT, XGBoost introduces regularization term and second derivative information, optimizes the training process of the model, and improves the accuracy and generalization ability of the model. XGBoost also has a good ability to handle high-dimensional data and nonlinear feature fitting. At the same time, XGBoost also supports parallel computation and missing value processing, which makes the training process of the model more efficient and stable. The results are shown in Figure 2, Table 1, Figure 3 and Table 2 below:

5.1 Pearson correlation analysis

5.2 XGBoost predicts results

Music is an important part of human culture, and emotions are at the heart of human emotional experience. Therefore, it is of great significance to study the relationship between music and emotion. The Turkish Music Emotion Dataset is a dataset for music emotion classification, which contains many music samples and their corresponding emotion labels. The analysis of this dataset can help us better understand the relationship between music and emotion, and provide data support for the construction of music emotion classification model.

In constructing the classification model of this dataset, we used the traditional random forest machine learning model and the XGBoost model, and compared their accuracy, recall rate and other indicators. The results show that the accuracy, accuracy and recall rate of the traditional random forest machine learning model are better than the XGBoost model, 80.8%, 80.8% and 80.5%, respectively. The comprehensive evaluation shows that the traditional random forest machine learning model has better comprehensive classification effect.

Traditional random forest machine learning models are ensemble learning algorithms that combine multiple decision trees for classification or regression. The advantages of random forest model are that it can process high-dimensional data, has strong nonlinear feature fitting ability, and has good robustness and interpretability. The robustness of the random forest model comes from the construction process of its decision tree, which is generated by

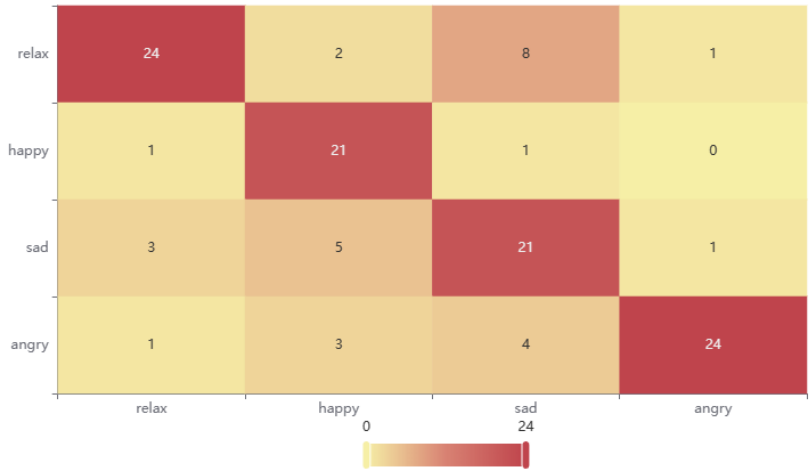

Figure 3: XGBoost confusion matrix. (Photo credit: Original)

random sampling and therefore has a certain tolerance for noisy data and missing values. In addition, random forest models can also provide an assessment of the importance of features, which helps to understand the decision-making process of the model.

XGBoost (eXtreme Gradient Boosting) is a gradient enhanced decision tree algorithm that improves the accuracy of the model by iteratively training multiple decision trees. XGBoost model has the advantages of fast training and prediction, high precision, and strong generalization ability. The fast training and prediction speed of the XGBoost model is due to its parallel computation and cache optimization mechanism, while it can also handle sparse data and missing values. In addition, the XGBoost model can provide an assessment of the importance of features, which helps to understand the decision-making process of the model.

In the field of music emotion classification, the traditional random forest model and XGBoost model have been widely used. Traditional random forest models can classify music by extracting various features, such as rhythm, tone, timbre, etc., while XGBoost models can classify music by learning emotional features, such as happy, sad, angry, etc. To sum up, based on this application scenario and data characteristics, the random forest machine learning model is better.

The traditional random forest machine learning model and XGBoost model have their own advantages and disadvantages. In practical application, it is necessary to choose the appropriate model according to the specific situation. For the Turkish music emotion dataset, the traditional random forest machine learning model performed better, but for other datasets, perhaps the XGBoost model would be more appropriate. Therefore, in practical application, we need to choose the right model according to the specific situation, and optimize and adjust it to obtain the best classification effect.

In practical application, we can choose a suitable model according to the needs and data characteristics of music emotion classification. If you need to deal with high-dimensional data and nonlinear features, you can choose the traditional random forest machine learning model; If you need to train and predict speed quickly and deal with sparse data and missing values, you can choose the XGBoost model. At the same time, in the process of model training, we also need to screen and optimize the features to improve the classification effect and generalization ability of the model.

To sum up, music emotion classification is a challenging and significant problem, and both traditional random forest machine learning model and XGBoost model are effective solutions. In practical application, we need to choose the right model according to the specific situation, and optimize and adjust it to get the best classification effect.

6 CONCLUSION

The Turkish Music Emotion Dataset is a dataset for music emotion classification, which contains many music samples and their corresponding emotion labels. The analysis of this dataset can help us better understand the relationship between music and emotion, and provide data support for the construction of music emotion classification model.

When constructing the classification model for this dataset, we use the traditional random forest machine learning model and XGBoost model, and compare their accuracy, recall rate and other indicators. The results show that the accuracy, accuracy and recall rate of traditional random forest machine learning model are better than XGBoost model, 80.8%, 80.8% and 80.5% respectively. According to the comprehensive evaluation, the comprehensive classification results of the traditional random forest machine learning model are better.

The traditional random forest machine learning model is an ensemble learning algorithm that combines multiple decision trees for classification or regression. The advantages of the random forest model are that it can process high-dimensional data, has a strong ability to fit nonlinear features, and has good robustness and interpretability. The robustness of the random forest model comes from the construction process of its decision trees, which are generated by random sampling, so they have a certain tolerance for noisy data and missing values. In addition, the random forest model

can also provide an assessment of the importance of features, which helps to understand the decision-making process of the model.

XGBoost (eXtreme Gradient Boosting) is a gradient boosting decision tree algorithm that improves the accuracy of the model by iteratively training multiple decision trees. The advantage of XGBoost model lies in its fast training and prediction speed, high accuracy and generalization ability. The fast training and prediction speed of the XGBoost model is due to its parallel computation and cache optimization mechanism, while it can also handle sparse data and missing values. In addition, the XGBoost model can provide an assessment of the importance of features, which helps to understand the decision-making process of the model.

In the field of music emotion classification, both traditional random forest model and XGBoost model have been widely used. The traditional random forest model can classify music by extracting various features, such as rhythm, tone, timbre, etc., while the XGBoost model can classify music by learning emotional features, such as happiness, sadness, anger, etc. In summary, based on this application scenario and data characteristics, the random forest machine learning model is better.

In summary, both the traditional random forest machine learning model and the XGBoost model have their own advantages and disadvantages, and it is necessary to choose the appropriate model according to the specific situation in practical application. For the Turkish music emotion dataset, the traditional random forest machine learning model performed better, but for other datasets, perhaps the XGBoost model would be more suitable. Therefore, in practical applications, we need to choose the right model according to the specific situation, and optimize and adjust it to obtain the best classification effect.

REFERENCES

[1] Feng G, Zhiyuan L, Xiaoliang L, *et al.* Does air pollution influence music sentiment? Measuring music sentiment by machine learning [J]. Journal of Asian Economics, 2023, 87.
[2] Sanchali D, K. B B, Swapan D. Retraction Note to: Building a computational model for mood classification of music by integrating an asymptotic approach with the machine learning techniques [J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 14 (Suppl 1).
[3] Yu X, Fumei X. Study on Music Emotion Recognition Based on the Machine Learning Model Clustering Algorithm [J]. Mathematical Problems in Engineering, 2022, 2022.
[4] Zijin W. Research on Automatic Classification Method of Ethnic Music Emotion Based on Machine Learning [J]. Journal of Mathematics, 2022.
[5] Quasim T M, Alkhammash H E, Khan A M, *et al.* Emotion-based music recommendation and classification using machine learning with IoT Framework [J]. Soft Computing, 2021, 25(18).
[6] S S U, P K. A review on machine learning based counterfeit integrated circuit detection [J]. Engineering Research Express, 2023, 5(4).
[7] Selini N H, Guido C, Constantinos A. Machine learning for activity pattern detection [J]. Journal of Intelligent Transportation Systems, 2023, 27(6).
[8] Chen W, Huiying Y, Huafeng L, *et al.* Anthropogenic contributions to antibiotic resistance gene pollution in household drinking water revealed by machine-learning-based source-tracking [J]. Water Research, 2023, 246.
[9] Li, Y., Nie, X., and Huang, R .2018. Expert Systems with Applications, 96, 261-270.
[10] Goh, K. L., and Singh, A. K. 2015. Procedia Computer Science, 70, 434-441.
[11] Bai J, Luo K, Peng J, *et al.* Music Emotions Recognition by Machine Learning With Cognitive Classification Methodologies [J]. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 2017, 11(4).