Music Classification Report

Anonymous

1 Introduction

With the increasing of the amount of music, people find it is difficult to manage and categorize the music that they listen to. Therefore, finding a way to categorize the music is becoming much more important, which can help people to conveniently get access to their favorite songs. Our task aims to use machine to automatically categorize the music into different genres according to plenty of features by applying machine learning. The task is a multi-class classification task because there are 8 different genres in the music dataset. In addition, our music dataset includes 8556 different songs, and each song is categorized into one genre. Moreover, each song has three different types of features which are metadata features, audio features and text features. We hypothesized that the combination of audio and text features is more predictable than these two separate features for classifying the genre for music.

2 Related Literature

People have done many efforts in the music genre classification area and they also have achieved many successes in this area. Bahuleyan (2012) addressed this issue by adopting surprised machine learning classifier such as Convolutional Neural Network and Gradient Boosting classifiers. He made use of audio features which are sound clips of each songs. His study indicated the Mel Frequency Cepstral Coefficients (MFCC) did the most contributions to the performance of the classifiers. In addition, his study also pointed that it is significant to handle noisy data for improving the performance. Moreover, Clark, Park and Guerard (2012) have achieved 68 percent testing accuracy by using the combination of two Neural Network learning approaches.

3 Methodology

3.1 Classifiers

3.1.1 K-Nearest Neighbors

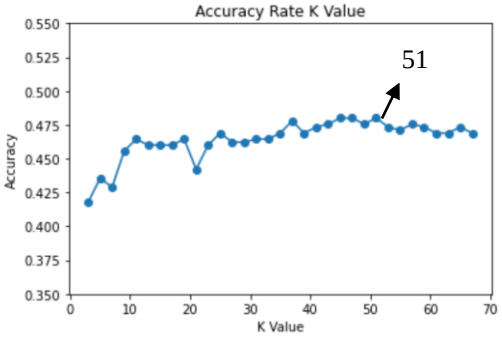

Figure 1- Accuracy results versus K value

K-NN classifier is a supervised machine learning approach that can be used for classification and regression. It stores the training dataset and measures the distance between testing and training instance. Moreover, it classifies the testing instance according to the K nearest instances. For improving the performance of the classifier, we set up the value 51 for K according to the accuracy trend shown in Figure 1 and apply weighted distance for classifying testing instance.

3.1.2 Gaussian Naïve Bayes

Gaussian Naïve Bayes is an extension of Naïve Bayes, which deals with continuous data. In addition, it does the prediction followed by the Gaussian normal distribution. However, the classifier requires independence assumption for features and instances.

3.1.3 Multi-layer Perceptron

Multi-layer perceptron is a special type of neural network, which are fully connected. Each layer contains many perceptron units. The main advantage of the classifier is that it can automatically do feature learning. For improving the performance of the Multi-layer Perceptron, we set the regularization parameter to 1 and the number of perceptron unit of hidden layer to 200 according to the result of GridSearchCV function which is to find the best parameters for the multi-layer perceptron classifier from the predefined

parameters list.

3.1.4 Decision Tree

Decision tree construct a tree representation by using the training data. Each node of the tree represents a feature. In addition, each leaf determines the classification. We implement the classifier by using information gain to do the attribute selection.

3.2 Baseline

We implement zero-R as the baseline, the accuracy of the baseline is 12.2% which performs poorly. The reason is that there are 8 different genres, therefore each genre takes up a small proportion. In addition, the distribution of each dataset is different.

3.3 Data Pre-processing

The section consists of the description of transforming the data into usable format. Each feature of the audio features takes a continuous value. However, the magnitude of the continuous value between different features is significantly different. Some of classifiers that are used in the study are sensitive to the magnitude of the values such as Multi-layer Perceptron classifier. Therefore, we re-scale the audio feature values so that each feature has a similar magnitude of values. The re-scaled values can improve the performance of the related classifiers. Moreover, for the text features, we treat each unique tag as an attribute. We assign 1 for the attribute if the text features of the song contain the related tag, otherwise assigning 0. Therefore, the text features are transformed into usable format which can be used in the classifier

4 Evaluation

4.1 Evaluation Strategy

The section consists of the description of the evaluation strategy and metrics that we have used. We implement the holdout strategy to evaluate the classifiers. Holdout strategy partitions data into two parts which are training part and validation part respectively therefore there is no overlap between datasets (Tan, Steinbach and Kumar, 2006). The study evaluates the performance of the 4 classifiers according to the accuracy metric. The accuracy metric refers to the percentage of the test instances that are correctly classified. The reason of choosing accuracy metric is that the aim of the task is to explore how accurate the songs can be categorized into correct genre.

4.2 Results and Analysis

The section consists of the results of different classifiers and the analysis of the performance of the classifiers on different datasets.

4.2.1 K-Nearest Neighbors

Feature for training	Accuracy
Audio Features	48%
Text Features	26.7%
Audio and Text Features combined	54.4%

Table 1- Performance comparison of K-NN classifier

K-NN classifier which is trained by combination of audio and text features generates the best performance in term of accuracy based on the values in Table 1. The performance of the K-NN classifier that is trained by text features individually is poor, which points to that the distance between instances that is measured by text features individually is inaccurate. However, the combination of audio feature and text feature significantly improve the performance. This illustrates that the audio feature plays an important role in measuring accurately the distance between instances. The reason is that we re-scale the audio values so that the magnitude of the values will not influence the performance of the classifier

However, the overall performance of the K-NN classifier does not reach an acceptable accuracy. One reason is that there are 8 different genres in dataset, therefore the probability of including unrelated classes is high which can lower classifier performance. The other reason is that the thousands of attributes contain a great number of noisy values, which influences the distance function to generate accurate results

4.2.2 Gaussian Naïve Bayes

Feature for training	Accuracy
Audio Features	37.8%
Text Features	41.6%
Audio and Text Features combined	41.8%

Table 2- Performance comparison of Gaussian Naïve Bayes classifier

The accuracy of Gaussian Naïve Bayes classifier which is trained by the combined features is 4 percentage higher than the audio feature classifier and 0.2 percentage higher than text feature classifier as shown in Table 2. However, the performances of these three classifiers are poor. This implies that the audio features and text features are not suitable for the Gaussian Naïve Bayes classifier. One reason of it is that the Naïve Bayes classifier assumes that each feature is strongly independent with other features, therefore it cannot include the interaction of features into classification (Rish, 2001). However, the audio features are extracted from the snippets of each track, therefore, there should be a relationship between each audio feature even they are not interpretable. In addition, the text features are the lyrics of the songs which points to each text feature is related to others. The other reason is that the Gaussian Naïve Bayes assumes that the values of each feature are normally distributed (Langley and John, 1995). However, there are two values for tag attributes which are 1 and 0. In addition, the ratio of these two values are significantly unbalanced. Therefore, these two reasons have led to poor performance of the Gaussian Naïve Bayes classifier

4.2.3 Multi-layer Perceptron

Feature for training	Accuracy
Audio Features	50%
Text Features	56.9%
Audio and Text Features combined	66%

Table 3- Performance comparison of multi-layer perceptron classifier

Multi-layer Perceptron classifier: The accuracy of Multi-layer Perceptron classifier which is trained by combination of audio and text features is 16 percentage and 9.1 percentage higher than the audio and text features classifiers respectively. This illustrates that the combination of audio and text features are more predictable.

Moreover, the overall performance of the multi-layer perceptron classifiers reaches an acceptable accuracy, especially the combination features one. The capability of feature learning is the key factor in doing great performance of multi-layer perceptron classifier in this task. The reason is that there are thousands of different attributes in the audio and text features and each attributes are related to others, the multi-layer perceptron has the ability to capture the iterative features and generate more representative features for the task by feature learning. In addition, as shown in Table 3, the accuracy results point that the multi-layer perceptron classifier can generate more useful features by using combination of audio and text features compared to the separate features.

Feature for training	Accuracy
Re-scaled Audio Feature	50%
Original Audio Feature	27.8%

Table 4- Performance comparison of multi-layer perceptron classifier between re-scaled and original dataset

Regularization Perceptron	Accuracy
1e-07	41.3%
1e-06	46.2%
1e-05	44.2%
0.0001	41.1%
0.001	42.2%
0.01	40.9%
0.1	45.1%
1	50%
10	44.9%

Table 5- Performance comparison of multi-layer perceptron classifier among different parameters

Dataset for predicting	Accuracy
Validation Dataset	66%
Training Dataset	99.6%

Table 6- Outcome of multi-layer perceptron classifier's training and validation accuracy

However, there are many reasons that can decline the accuracy of the multi-layer perceptron classifier. The first reason is the magnitude of the values. As shown in Table 4,

the accuracy results point that there is a significant difference between re-scaled audio features and original audio features, which points out the multi-layer perceptron is sensitive to the magnitude of feature values. The second reason is the regularization parameter which is used to control the capability of the classifier (Collobert and Bengio, 2004). As shown in Table 5, the accuracy results point that there is a significant gap between the best result and the worst result. Therefore, it is important to tune the regularization parameter for improve the performance. However, the most importance reason is the overfitting of dataset because the multi-layer perceptron is a powerful tool which can capture noisy of data to generate completed classifier (Eisenstein, 2019). As shown in Table 6, there is a significant gap between the outcome of training and test instances, which illustrates that the classifier tunes itself to be the characteristics of the training dataset instead of learning the general pattern of it.

4.2.4 Decision Tree

Feature for training	Accuracy
Audio Features	31.1%
Text Features	36.9%
Audio and Text Features combined	43.3%

Table 7- Performance comparison of decision tree classifier

Dataset for predicting	Accuracy
Validation Dataset	43.3%
Training Dataset	100%

Table 8- Outcome of decision tree classifier's training and validation accuracy

Decision Tree classifier: The classifier which is trained by combination of audio and text features generates the best performance based on the results in Table 7. However, the overall performance of the decision tree classifiers is poor. One reason is that the values of audio features are continuous number which cause the classifier loss information. In addition, a great number of values in each audio feature generate many branches which leads to the bias of the feature. The other reason is the thousands of features make the classifier be complex which results in the overfitting of training dataset. As shown in Table 8, there is a significant gap between the outcome of training and test instances.

Given the above, the multi-layer perceptron classifier does the best performance compared to other classifiers. The reason is that there are a significant number of features which must contain noisy data and irrelevant data, the multi-layer perceptron has the feature engineering ability which can generate useful features based on the original features for predicting.

5 Conclusions

This study categories the genre of music based on audio features, text features and combination of audio and text features. We use 4 different models which are K-Nearest Neighbors, Gaussian Naïve Bayes, Multi-layer Perceptron and Decision Tree models to classify the music. The multi-layer perceptron model was confirmed to be the best model. However, the overfitting issue of multi-layer model is serious in this study, which declines the performance. Therefore, we can identify methods to remedy the overfitting issue of the classifiers in the further study which may improve the performance. Moreover, the evaluation results support our hypothesis that the combination of audio and text features is more predictable than two separate features for classifying the genre for music. In the further studies, we can identify ways of finding the important features and ignoring the noisy features before training the classifiers.

References

Bahuleyan, H. (2018), Music Genre Classification using Machine Learning Techniques, University of Waterloo, Canada.

Bertin-Mahieux, T., Ellis, D. P.W., Whitman, B., Lamere, P. (2011). The million song dataset. In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)

Clark, S., Park, D., Guerard, A. (2012). Music Genre Classification using Machine Learning Techniques

Collobert, R and Bengio, S. (2004). Links between perceptrons, MLPs and SVMs. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 23

Eisenstein, J. (2019). Natural Language Processing. MIT Press. Chapters 3 (intro), 3.1, 3.2

John, G. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (UAI'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 338–345

Rish, I. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3

Schindler, A. and Rauber, A. (2012). Capturing the temporal domain in Echonest Features for improved classification effectiveness. In Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR)

Tan, P., Steinbach, M., Kumar, V. (2006). Introduction to Data Mining