Classification of Music Samples by Style

Ran Tian Mingze Liu Zhengqi Li

Abstract

In this paper, we investigate the problem of music genre classification. We applied five different machine learning algorithm to music genre dataset for classification purpose. Our results show that neural network and random forest algorithm achieve the best performance of classification on the music dataset. More importantly, we found there are four classes in the music dataset that are more discriminate than other music styles after we applied the algorithms on such classes.

Introduction

As increasing number of digital music arise in people's everyday life, an automatic classification method becomes necessary for distinguishing different music classes. When people open a music software or enter a music store, they can find the music are sorted in different classes. Comparing with sorting by time and by player, it's the most fundamental and most popular method to sort music. Classification by genre is one of most popular methods in the current website to help people to find out their favorite music. In the past, it requires professional musician to classify the class of music, which costs too much time and lack of objectivity; for example, different musicians may classify the same music into different classes. However, automatic musical style classification by machine can automate classification and provide an important information for a complete music information retrieval system in the internet.

A machine learning approach of music classification algorithm is one of the most popular and effective method to classify the audio signal. By analyzing the special music frame, for example, the features of audio signal within the music, the numeric dataset can be established. On this ground, many classification methods can be applied to determine the class of music. In addition, features extracted from the audio signal can also be used to describe other information of music. For instance, these features can be used to find similarity in the music database.

In this paper, we uses five different machine learning algorithm to address the music genre classification problem. Specifically, we firstly extract major features from each music sample in the dataset. Then we apply dimension reduction algorithm for feature selection. After extract principal components from the features of the music, we feed our training dataset into 5 classification algorithm: K-means, K-Nearest Neighbors, multi-class support vector machine, neural network and random forest. The performance of each algorithm is evaluated by the test validation error on test dataset.

Related Work

Before analyzing the classification algorithm, the first thing is to generate the dataset. A very commonly used speech recognition method, Mel-frequency cepstral coefficients[1][3], can be used on feature extraction. One of its disadvantage is only a short sample will be taken from each music, in this way many useful information may be lost if the sampling method is not appropriate. Another method to study audio is Spectrogram [2][4], which focuses on time-frequency of audio without losing any physical information.

Different classification methods have been tried in this field, including both supervised learning and unsupervised learning. Fisher Classier, linear classifier (LDC), Naive Bayes Classifier and K-Nearest Neighbors are most common supervised learning algorithm[6][7][8]. On the other hand, K-Means, Support Vector Machines and Expectation-Maximization (EM) algorithms have been adopted by machine researchers on the unsupervised clustering problems[9][10]. In addition, Some researchers also came up with several advanced methods such as Neural Network and Random Forest [1][2][5][11], which result in promising classification results.

Dimensionality reduction algorithm is mentioned by previous researches which is considered as a good way to achieve feature selection from the dataset which has larger number of dimension of features. In our project, we will try to reduce the dimensions of features space in the music dataset before the classification methodology been applied, if necessary.

Procedure and Measurements

1: Preprocessing Data

In this application, a method will be provided to measure the distance between incomparable music clips, in other words, it will transform the data (music) to feature vectors which can be handled in Matlab. Since the choice of feature is highly related to the classification algorithms later, it is critical to have a good metric to distinguish different classes. In this paper, we choose the Mel-frequency Cepstral Coefficients(MFCC) to extract the features of music.

Here, we use GTZAN Genre Collection dataset for our project. We preprocess 10 types of music in our sample - $Blues$, $Classical$, $Country$, $Disco$, $Hiphop$, $Jazz$, $Metal$, Pop, $Reggae$ and $Rock$, 100 for each type.

1.1: Mel-frequency Cepstral Coefficients(MFCC)

MFCC is a way to represent the music waveform in some coefficients in frequency domain. The general process is as following.

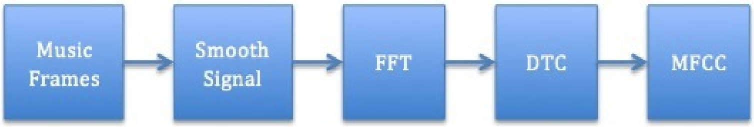


Figure 1: General Flow of MFCC

- Take about 20 seconds of frame and take a sample rate around 28KHz in the middle of every music.
- Use a smooth window to smooth the edges.
- Take the FFT to these samples.
- Map each of the power to mel frequencies and take the discrete cosine transform (DTC) of the list of mel log powers.
- Ignore the high frequency components and MFCCs are the amplitudes of the spectrum.

1.2: Principle Component Analysis(PCA)

If we used all the features that extracted from the music data set, the dimension of each feature vector of music sample is large, it will result in very slow speed performance and large number of computational burden. Thus, it's necessary to use method of dimensionality reduction to keep the only major dimension which have sufficient information.

Principal Component Analysis (PCA) is most common method for dimensionality reduction. The algorithm orthogonally projects the feature of the data onto a lower dimensional linear space such that the variance of the projected data is maximized. In the implementation, choose the major dimension which have the ratio of variance larger than 95% as the effective dimension which will be used for classification algorithm after applying PCA on the data. In our data, we choose dimension as 20.

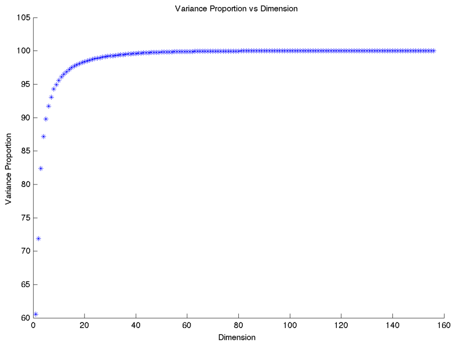


Figure 2: Variance Proportion vs Dimension

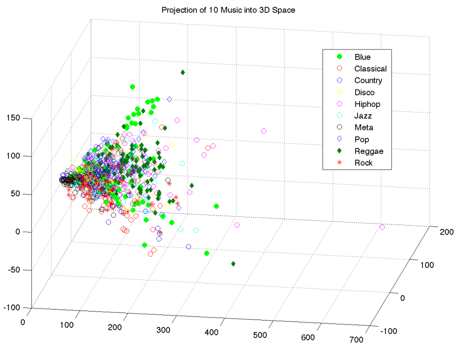


Figure 3: Projection of 10 Music Samples into 3D Space

2:Classification Stages

Here we propose 5 main algorithms for classifying music samples, including unsupervised learning - K-means, supervised learning - KNN, SVM, Random Forest and Neural Network. Firstly, we are trying to classify all 10 types with each algorithm. Then we particularly choose 4 types($Classical, Hiphop, Metal, Reggae$) to dwell on in case of failing classifying 10 types.

2.1:K-Nearest Neighborhood

K-Nearest Neighborhood is the easiest classification algorithm. The algorithm uses training data set to compute gaussian model for each sample, then for each new test data of song, computing the distance between new song and all other songs in the training set. Here, use two definition of distance and compare the performance based on different distance.

- Euclidean Distance
- City Block

In the implementation, choose k =10 $\sim$ 100, to see the classification error rate for each k value.

2.2:K Means Clustering

K-Means algorithm was used to cluster similar songs together, which is a unsupervised learning. Here, Euclidean distance has been set as the default distance. The experiments are performed with datasets with different number of labels. The comparison of the performances are provided in the conclusion part. Note that after we got the results from K-means, since its unsupervised learning, the clusters may not be the correct labels. In order to pair the clusters up with the right labels, conditional probability will be used. For each clusters, assign the label with the highest conditional probability.

2.3:Support Vector Machine

Support Vector Machine, known as the maximum margin classifier, is one of the most powerful tools for binary classification. For the music data set, because there are multiple music genres for classification, it is necessary to extend Support Vector Machine to multi-class classification.

In our samples,in order to classify 10 classes, multi-class SVM build 10 binary classifiers f1...f10, and each trained to classify one class from the rest.[13]Unfortunately, SVM fails to classify all 10 types of music, but for 4 types,it works pretty well.

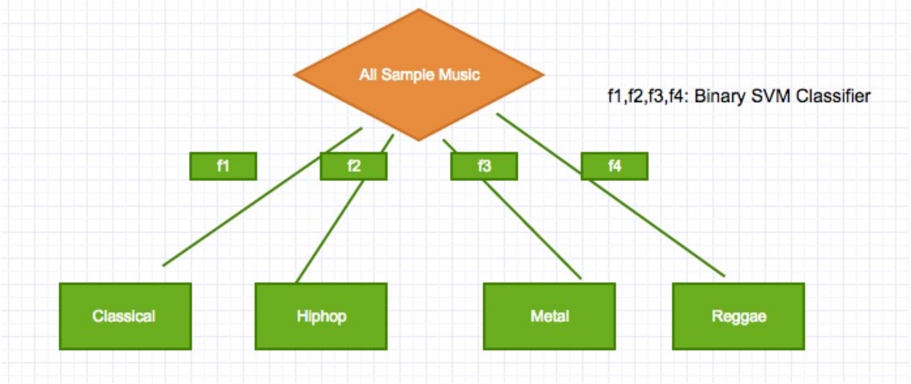


Figure 4: Multi-Class SVM With 4 Classification

2.4:Random Forest

Random forest, just as its name, is constructed by many independent classification and regression trees (CART). These classification and regression trees are picked from the training sample randomly using bootstrap sampling method and compose the new bootstrap samples. The number of bootstrap samples should be equal to the number of trees and the size of each bootstrap samples should be equal to the size of sample. Since the sampling method is with replacement, so some of training sample may appear more than once in the same classification and regression tree while some of them may not appear in any classification and regression tree. In this way, for each tree, the bootstrap samples are not the whole sample (even the size is the same), so the model over-fitting problem will be avoided.[11]

Now each bootstrap sample is corresponding to one classification and regression tree. Next is to train all the classification and regression trees. For each node, randomly sample the input variables and pick the best split from the selected sample, instead of choosing the optimal split from all the input variables.

After finishing the training process, random forest now can predict the outcome from the new data. The prediction is taking the average of the prediction from each classification and regression tree.

The number of tree is set to 500, which is enough for this dataset because the outcome converges. Set the argument doBest=TRUE, it optimize the sample size which minimize the "out-of-bag" error rate. Set the argument importance=TRUE, it do the variable selection automatically.[14]

2.5:Neural Network

Neural Network is a machine learning model that mimics the biological neural networks of human brains. The neural network model consists of multiple layers, which can be described as a series of functional transformation. The two layer neural network is shown in figure 5. Suppose we have input features of dimension D, M hidden units and K output units. We can combine all the stages into overall network function:

$$y_k(\mathbf{x}, \mathbf{w}, \mathbf{v}) = h_2(\Sigma_{j=1}^{M} v_{kj} h_1(\Sigma_{i=1}^{D} w_{ji} x_i + w_{j0}) + v_{k0}) \tag{1}$$

where v_{kj} and w_{ji} are weights for input units and hidden units. h_1 and h_2 are nonlinear transformation. Our music style classification problem is a multilcass classification problem. We use logistic sigmoid function as activation function of hidden units (equation 2) and softmax function as activation function of output units (equation 3).

$$h_1(x) = \frac{1}{1 + exp(x)} \tag{2}$$

$$h_{2j}(z) = \frac{exp(z_j)}{\Sigma_{m=1}^{M} exp(z_m)} \tag{3}$$

We use gradient descent algorithm for optimization in order to get optimal weight parameters (equation 4). Furthermore, we use back-propagation algorithm to obtain the error-function derivatives.

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \beta\nabla E(\mathbf{w}^{\tau})) \tag{4}$$

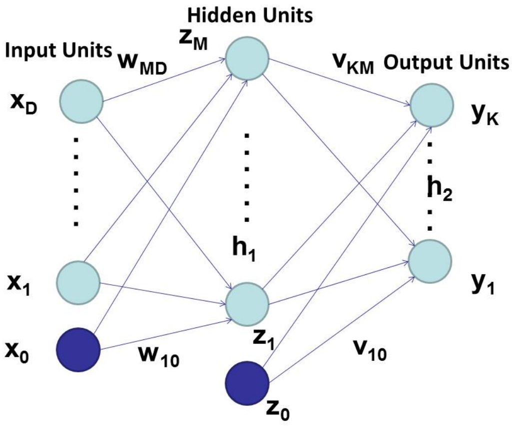


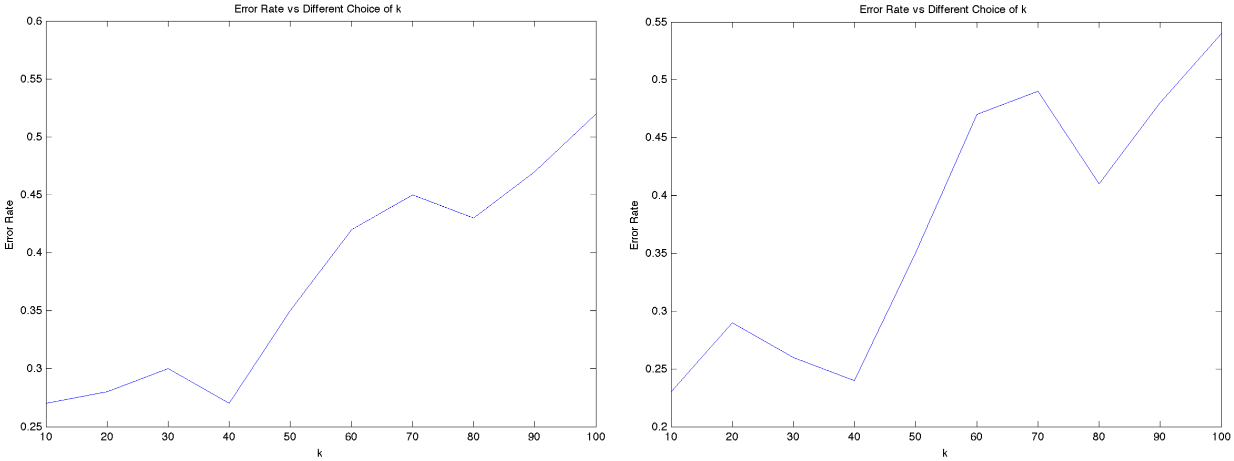
Figure 5: Two layer neural network structure. The input has D dimension. There are M hidden units and K output units. h_1 and h_2 is nonlinear transformation function. w and v are weights for linear combinations of inputs units and hidden units respectively

Results

Each type of music contains 100 samples. We try different percentages of test and training size, but show the results with training percentage 0.8. The remaining 20 percentage of songs are used for cross validation and testing.

3.1: K-Nearest Neighborhood

For 10 types of music, KNN shows a poor performance-the error rate is pretty high(error rate is 50% or higher) no matter what kind of distance we use. But for 4 types of music, we have a relatively good estimation(around 30% of the error rate). For different choices of k, it does not show many differences for smaller k,but if we keep increasing k which is greater than 50, the error rate will be higher.



(a) Euclidean Distance

(b) City Block Distance

Figure 6: Error Rate vs K

3.2: K-Means Clustering

The error rate for K-means is vastly high (over 50%) in 10 types of music case and its very hard to assign the correct labels to the clusters(for example, one cluster only contains two data, which is impossible to use conditional probability to determine its corresponding label). When the number of types of music reduced to 4, the error rate achieves around 37% which is improved significantly but still very high. The overlap data and similarity of feature are considered to be the reasons.

3.3: Multi-Class Support Vector Machine

The Multi-Class SVM doesn't work well for 10 types(the error rate is around 40%), which does't show much priority comparing with the KNN and K-means. But as for 4 types, the estimation is more accurate, the average error rate is about 15%.

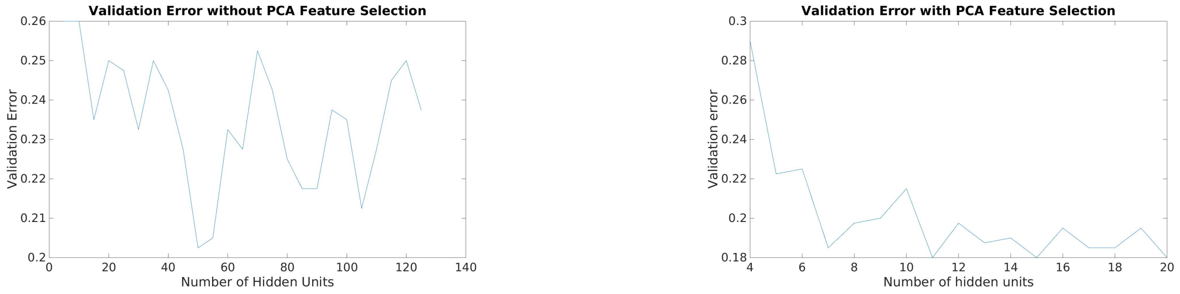
3.4: Random Forest

The results from random forest are much better than K-means. The error rate for 10 types of music is 23.5% and it reduced to 2.5% for the 4 types of music case. The reason why random forest performs better is that it do variable selection naturally and uses bootstrap sampling to avoid overfitting. Whats more, random forest handles the problem that the number of features is close to the sample size, because every time when split the classification trees, it randomly picks features, which number is optimized.

3.5: Neural Network

We applied normalization to all the features in the training set and validation set before we feed the data into neural network. We test neural network algorithm on the music training dataset with pre-processing of PCA and without pre-processing of PCA. In addition, we test the algorithm separately on 10 types of music case versus 4 types as mentioned in subsection 3.2.

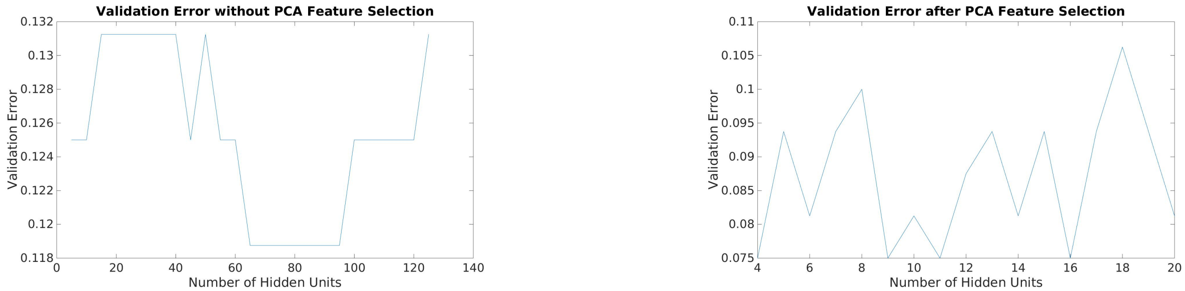
We evaluate the validation error of neural network with different hidden units. We adjusted the number of hidden units in the neural network between the dimension of input features and output units. The test validation error with different hidden units is shown in figure7 to figure8. When the algorithm is applied to the whole dataset to classify 10 music genres, we found that the optimal validation error for input features without PCA feature selection is 20.25%. On the other hand, if we applied PCA before we feed the training data into neural network, the optimal validation error reduces to 18%. In addition, when we reduce the music style into 4 major types, the validation error for dataset without PCA is 11.87% and 7.5% for dataset with PCA pre-processing.



(a) With PCA

(b) Without PCA

Figure 7: Validation error for dataset with 10 classes with/without PCA



(a) With PCA

(b) Without PCA

Figure 8: Validation error for dataset with 4 classes with/without PCA

3.6:Performance Comparison

Here we have the table which contains the error rate for different algorithms. The training percentage is 0.8 for all algorithms.

Average Error Rate(%)		Different Algorithms				
		KNN	K-means	SVM	Random Forest	Neural Network
Num of Music Types	10	52.7	61.2	41.2	23.5	19.1
	4	28.9	34.2	15.5	2.5	9.6

Table 1: Comparison with Different Algorithms

Conclusion

4.1: Discussion

According to the test results we have form different classification stages, all algorithms performed fairly well when classifying 4 types. But with respect to 10 music types, only Random Forest and Neural Network show great estimations. Simple K-means and KNN approaches are generally less accurate than SVM,Neural Network and Random Forest. Without surprise, the most sophisticated classification method is Random Forest.

For all 10 types of music, the most frequent misclassification happens among Reggae ,Rock, Hiphop and Disco, which are within our expectation since even in real life, sometimes it's not easy to classify them by us.

4.2: Future Work

In regards to the preprocessing, the total size of data(containing training and test data) is only 1000, which is not enough for accurate analysis. Besides, we only take fractional parts of the whole song at sampling rate of 28kHz, which could be a potential factor of inaccurate input for the classifiers. Thus trying different fractions in a song and different sampling rate is a good choice for more accurate error analysis. The feature extraction throughout the whole project is based on the single feature(MFCC), So exploring different feature representations of music will help us compare different algorithms in machine learning.

As for the design classification states, Multi-Class SVM fails to classify all 10 music types accurately, which came as a surprise for us. It doesn't show the performance as we expected and we don't have a reasonable answer why this happens. So future work on alternative SVM classification is highly recommended. In addition, we can further investigate deep neural network algorithm on the music dataset to see if deep learning can achieve better performance than normal neural network.

References

[1] http://cs229.stanford.edu/proj2009/RajaniEkkizogloy

[2] http://www.speech.cs.cmu.edu/15-492/slides/mfcc

[3] $http://cs229.stanford.edu/proj2011/HaggbladeHongKao-MusicGenreClassification$

[4] Deshpande, Hrishikesh, Rohit Singh, and Unjung Nam. ”Classification of music signals in the visual domain.” Proceedings of the COST-G6 Conference on Digital Audio Effects. sn, 2001.

[5] Cataltepe, Zehra, and Berna Altinel. ”Music recommendation based on adaptive feature and user grouping.” Computer and information sciences, 2007. iscis 2007. 22nd international symposium on. IEEE, 2007.

[6] Tzanetakis, George, and Perry Cook. ”Musical genre classification of audio signals.” Speech and Audio Processing, IEEE transactions on 10.5 (2002): 293-302.

[7] Saari, Pasi, Tuomas Eerola, and Olivier Lartillot. ”Generalizability and simplicity as criteria in feature selection: Application to mood classification in music.” Audio, Speech, and Language Processing, IEEE Transactions on 19.6 (2011): 1802-1812. [8] Celma, Oscar. Music recommendation. Springer Berlin Heidelberg, 2010.

[9] Mandel, Michael I., and Daniel PW Ellis. ”Song-level features and support vector machines for music classification.” ISMIR 2005: 6th International Conference on Music Information Retrieval, 11-15 September, 2005. Queen Mary, University of London, 2005.

[10] Saari, Pasi, Tuomas Eerola, and Olivier Lartillot. ”Generalizability and simplicity as criteria in feature selection: Application to mood classification in music.” Audio, Speech, and Language Processing, IEEE Transactions on 19.6 (2011): 1802-1812.

[11] L. Breiman. Random forests. Mach. Learning, 2001. 4

[12] http://mirlab.org/jang/books/audioSignalProcessing/goTutorial.html

[13] https://www.sec.in.tum.de/assets/lehre/ws0910/ml/slideslecture9.pdf

[14] A. Liaw and M. Wiener. Classification and Regression by randomForest. R News, $Vol.2/3$, December 2002