Music Classification CS 429/529

Xiaomeng Li and Vanessa Job
11/14/2017

Abtsract

This is the Project 3 report for Machine Learning: Music Classification. The report is about Support Vector Machine (SVM) and Logistic Regression (LR) classifiers' implementations and result discussions. The codes are produced by Python and there will be README file for the code to accurately run. The introduction, results and discussion will be included in this report.

Problem Description

Assuming a scenario where we want to classify a set of music files into the music genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock based on the music file itself. We will use GTZAN dataset, which contains the first thirty seconds of one hundred songs per genre. In the training data file we were given, we have ten genres in total and ninety songs per genre. For the testing data file, we have one hundred unlabeled songs. The tracks are recorded at 22,050 Hz (22,050 readings per second) mono in the WAV format. Using the function from Python, we can derive the samples and sample_rates. Using Matplotlib in Python, we can observe that the representations of frequencies that occur in a song are different clearly from genre to genre.

The goal of this project is to extract individual frequency intensities from the raw sample readings and feed them into a classifier. There are two frequency extraction methods: Fast Fourier Transform (FFT) and Mel Frequency Cepstral Coefficients (MFCC). For FFT we will derive first one thousand components for initial training. For MFCC, we will ignore the first and last ten percent of each song since they are less genre-specific than the middle part and we will do an averaging per coefficient over all the frames, i.e., we will only use 13 features per song for the initial training.

Third Feature Extraction

We implemented FFT (Fast Fourier Transform) and MFCC (Mel Frequency Cepstral Coefficients) feature extraction as per the assignment as well as a custom classifier. The custom classifier used the following features:

- The MFCC features extracted earlier, but with the maximum taken over all the frames, instead of the average.

- The first thirteen of the FFT features extracted earlier.
- The tempo of the sample as calculated by the function from librosa.
- The contrast of the sample as calculated by the function from librosa.

As you can see in Table 1 below, the accuracy for the MFCC features with SVM is 0.40000, much better than the FFT accuracy for 1000 coefficients. This makes sense because the Mel Frequencies model how humans perceive sounds. So we began our extractor design using the MFCC features as our base features. Then we experimented.

In [1], the Deepa and Suresh stated "In the case of mel frequency cepstral coefficients (MFCC), delta MFCC and autocorrelation MFCC, almost all coefficients of a particular frame are near the maximum value of that frame". We decided to try taking the maximum of our MFCC coefficients instead of the mean, since according to Deepa and Suresh, this should not make much difference. With the SVM classifier we found that using the maximum made no difference in the accuracy, while using the maximum with the LR classification increased our accuracy by 8%. So we included taking the maximum over the MFCC coefficients rather than taking the mean. In Table 1, this is denoted as MFCC-max.

Next we tried adding in some FFT coefficients. We experimented with various numbers of them and found that if we used 13, we got the greatest improvement in accuracy. It is likely that adding in this small number of FFT coefficients gave greater weight to the most prominent frequencies in the samples and perhaps provided slightly greater separation between the classes.

Finally, we tried adding tempo and contrast features obtained using functions from LibRosa [2], a library of music signal analysis. We got modest, though slight increases in accuracy from these features.

How to Improve Third Feature Extraction

We would like to try more of the methods outlined in [1], particularly Delta MFCCs, which measure the changes in the MFCCs. We would also like to try every single function in [2]. In doing a search of the literature, we've found that other authors have greater accuracy [4]. We've just barely scratched the surface of what we could try.

Also, though it would not improve the feature extraction, we would like to try a neural network classifier.

Methods introduction

A. Support Vector Machine

Introduction

Support Vector Machine (SVM) is a supervised learning model which is able to implement algorithms to analyze data. It could be used for both regression and classification. With the training examples, SVM will assign new examples to one category or the other for its model to make it more reliable. SVM model represents examples as points and divides them by different gaps. New data for testing will be put into the model and tested by the classifier to see which gap they could fall into to make better prediction.

Why choose SVM?

With the kernel (a class of algorithms which are for pattern analysis) to perform nonlinear classification, we could analyze the music file to map them into different genres based on training music file's individual frequency intensities or power spectrums.

B. Logistic Regression

Introduction

Logistic Regression (LR) is a learning model which is used for classification in training data. The prediction of output will be transformed by logistic function. At the end of the training, the model that is built will produce a function that could best represent the relation between input training data and output training data. New testing data will be applied on these functions to figure out which class they should belong to.

Why choose LR?

LR could measure the relationship between categorical dependent variables (target class: music genre) and one or more independent variables, which is good for using music file's individual frequency intensities or power spectrums as training data features.

Accuracy for Combinations of Classifiers and Features

Classifier	Features	Kaggle Accuracy
LR	FFT-1000	0.20000
SVM	FFT-1000	0.28000
LR	MFCC 13	0.36000
LR	MFCC-max	0.42000
SVM	MFCC, means taken over frames	0.40000
SVM	MFCC-max, i.e. max taken over frames	0.40000
LR	MFCC, FFT-13, tempo, contrast	0.46000
SVM with linear kernel	MFCC-max, FFT-12, tempo, contrast,	0.38000
SVM with kernel of degree 2	MFCC-max, FFT-13, tempo, contrast	0.48000 Best score

Table 1: Accuracy table from Kaggle with three data features and two classifiers

Here FFT-1000 means taking the first 1000 coefficients of the Fast Fourier transform (Analogously, FFT-13 means taking the first 13.).

K-fold test and accuracy report

In this project, we perform 10-fold test for each of the three data features with SVM and LR. Though the accuracy for each fold of ten is different, we only use the mean value of these ten accuracy values for each data feature testing of SVM and LR.

Topic	SVM with FFT	LR with FFT	SVM with MFCC	LR with MFCC	SVM with Third	LR with Third
Accuracy	0.271	0.362	0.498	0.473	0.482	0.366

Table 2: Accuracy Table for 10-fold experiment

It is easy to prove that when it comes to SVM and LR with all three data features, the 10-fold results from MFCC with SVM give the best performance whereas FFT from SVM gives the worst performance.

Confusion matrix for best Accuracy and discussion

By definition a confusion matrix C is such that Cij is equal to the number of observations known to be in group i but predicted to be in group j. In this 10-fold analysis process, we only use the confusion matrix which comes from the summation of all of the ten confusion matrices after all of the 10-fold analysis.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	21	27	14	1	8	5	5	6	2	1
classical	0	78	1	2	0	9	0	0	0	0
country	1	29	27	12	6	4	1	4	2	4
disco	6	27	29	7	5	4	2	3	5	2
hiphop	5	11	17	6	33	3	1	8	3	3
jazz	0	30	16	3	4	35	2	0	0	0
metal	6	22	20	11	6	4	12	4	5	0
pop	11	6	14	10	17	3	2	20	6	1
reggae	3	17	39	9	3	2	3	2	7	5
rock	10	16	38	4	8	2	1	5	2	4

Table 3: Confusion Matrix of best performance for SVM with FFT

From Table 3, we see that when using FFT and SVM, classical and country are the most often confused with another genre. SVM mistakenly classified disco as country (29 times) and reggae as country (39 times) and rock as country(38 times). Similarly, SVM classifies blues as classical (27 times), country as classical (29 times), disco as classical (27 times) and jazz as classical(30 times). Our guess is that FFT makes SVM mistakenly think classical and country has a lot in common with other features so the classifier made plenty of mistakes on them. In the same case, disco's data feature looks quite undistinguished with classical and country therefore it trips SVM seriously.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	38	5	11	3	5	5	7	7	7	2
classical	1	68	6	4	1	6	3	0	0	1
country	4	2	35	9	4	5	5	5	10	11
disco	4	10	25	13	1	5	6	4	11	11
hiphop	5	3	7	9	20	2	9	6	17	12
jazz	7	7	9	4	1	51	4	0	3	4
metal	7	2	7	6	7	3	30	4	13	11
pop	4	1	9	11	3	9	13	28	8	4
reggae	6	2	21	7	1	4	9	3	16	21
rock	3	2	14	9	4	2	10	6	13	27

Table 4: Confusion Matrix of best performance for LR with FFT

Observing from Table 4, it is easy to find that when it comes to FFT, LR tends to classify reggae as country (21 times) and classify disco as country (25 times). In this case, our guess is that the frequency intensities of reggae and disco should be similar to the country, which is why they confused LR classifier. LR also tends to classify reggae as rock (21 times). We think the reason should be similar. In both cases, reggae shows country and rock also have a lot in common with reggae. They both confused LR 21 times.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	37	2	7	6	3	3	11	15	5	1
classical	1	79	1	0	0	5	0	0	1	3
country	10	1	31	8	9	3	1	13	7	7
disco	3	0	8	35	6	4	7	8	6	13
hiphop	6	0	12	11	13	5	5	15	13	8
jazz	6	18	4	4	0	39	8	5	3	3
metal	0	0	1	6	3	1	70	4	3	2
pop	5	0	4	7	2	1	3	63	3	2
reggae	6	2	18	8	5	2	3	1	40	5
rock	6	0	8	18	1	6	18	5	10	18

Table 5: Confusion Matrix of best performance for LR with MFCC

From Table 5, it shows that when it comes to MFCC through LR classifier, it is easy to observe that LR tends to classify rock as disco (18 times) or as metal (18 times). Our guess is that when it comes to the data fitting of LR, disco and metal have a lot in common with rock so the classifier is confused. Also, LR tends to classify reggae as country (18 times), classify jazz as classical (18 times). We suppose classical genre's power spectrum is similar to jazz and country's power spectrum is similar to reggae for LR.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	39	1	4	12	8	2	5	7	11	1
classical	1	77	1	0	1	7	0	0	0	3
country	9	0	31	12	14	1	0	7	9	7
disco	5	0	10	30	10	2	1	6	7	19
hiphop	5	0	13	15	27	2	3	11	6	6
jazz	2	10	4	4	2	58	0	3	2	5
metal	2	0	1	3	6	0	66	2	2	8
pop	3	0	9	8	6	2	0	55	3	4
reggae	8	0	13	9	11	2	2	1	40	4
rock	5	0	7	24	7	3	6	4	10	24

Table 6: Confusion Matrix of best performance for SVM with MFCC

From Table 6, When MFCC is implemented through SVM, the confusion matrix shows that SVM tends to classify rock as disco (24 times), classify disco as rock (19 times). We think that the MFCC file of rock and disco may have a lot in common so that it confuses SVM when classifying both of them.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	38	6	15	2	2	6	6	3	8	4
classical	3	69	5	2	0	6	1	0	0	4
country	3	3	32	11	2	6	3	5	13	12
disco	3	12	19	17	1	3	8	4	11	12
hiphop	5	3	6	9	18	4	12	8	13	12
jazz	4	8	6	7	1	52	4	0	2	6
metal	5	0	9	7	5	7	31	5	11	10
pop	5	2	8	9	3	9	14	27	8	5
reggae	8	1	19	6	3	7	10	0	18	18
rock	2	2	16	9	4	1	9	6	14	27

Table 7: Confusion Matrix of best performance for LR with the third data feature

Table 7 clearly shows that after LR was applied on the third data feature, country tends to distract disco (19 times). reggae(19 times) and rock(16 times). So the country genre music files seemed to distract music file more no matter which feature is applied. Similarly rock tends to confuse reggae (18 times), and rock also tends to mislead other classifiers in other data features as before.

Genre	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	36	2	10	6	3	1	3	8	15	6
classical	1	70	2	1	0	14	0	1	0	1
country	7	1	48	9	5	1	3	4	8	4
disco	11	0	13	28	5	2	2	6	7	16
hiphop	4	0	3	11	36	1	3	9	16	5
jazz	3	18	1	0	1	56	2	2	2	5
metal	4	0	6	2	0	1	66	1	4	6
pop	9	1	6	10	11	2	2	41	6	2
reggae	17	3	7	6	9	1	4	7	32	4
rock	4	4	10	16	5	8	14	1	8	20

Table 8:: Confusion Matrix of best performance for SV with the third data feature

From Table 8, it shows that reggae is easy to be confused by blues (17 times), jazz is easy to be recognized as classical (18 times), rock is easy to be confused by disco (16 times), hiphop is easy to be judged as reggae and disco is easy to be classified as rock. Our guess is that compared with the data features before, new features tend to be misclassified as reggae. But what is not changing is that the new features also tend to be classified as country, which is similar to before.

Conclusion

From the results of several confusion matrices, we think the classes classical, country and rock influence the judging of classification most frequently since a lot of files from reggae, hiphop or disco are in fact misclassified into their genres. Although the highest accuracy for our training from Kaggle is not high, there are still a lot of possible data features and other algorithms we have not tried.

References

[1] P. L. Deepa and K. Suresh. “An optimized feature set for music genre classification based on Support Vector Machine,” In *Recent Advances in Intelligent Computational Systems (RAICS),* IEEE, 2011, pp. 610-614

[2] B McFee, C Raffel, D Liang, DPW Ellis, M McVicar, E Battenberg, O Nieto. “librosa: Audio and music signal analysis in python: *Proceedings of the 14th Python in Science Conference,* pp. 18-25

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), pp. 2825-2830.

[4] J.Yoon, H. Lim, and D.-W. Kim “Music Genre Classification Using Feature Subset Search”, International Journal of Machine Learning and Computing, Vol. 6, No. 2, April 2016, pp. 134-138.