An Investigation on Music Genre Classification by Timbral Feature

by

Wenchao Du Youke Shen Zihao Wu

CS 486 project report

1 Introduction

Music genres are labels people used to categorize music. The boundaries between different genres can be defined from various aspects, for example, instrumentation, harmony, singing style, and the literal content. The purpose of this project is to generate insights into the classification of music by non-content based features, as content based classification would require expertise in the theory of music composition.

2 Data Set and Features

We are using "Million Songs Dataset" from http://labrosa.ee.columbia.edu/millionsong/. The data set provides labels from a crowd-sourced music encyclopedia (musicbrains.org). The data set also comes with segmental timbral features and relative dominance of pitches. In general, information on pitch is deeply related with the harmony of music, which is out of the scope of this project, so we are only using timbral features for classification. We chose about 600 songs from three genres which has highly distinguishable instrumentation: metal, electronic, and folk. Otherwise the classifier trained on timbral features would perform poorly. In all the classification tasks, we used 2-fold cross validation.

3 Feature Engineering

Timbral features are extracted from consecutive segments of equal lengths of a song instead of the entire song. We came up with several ideas on how to use the segmental features, some of which made different assumptions on the nature of segments:

(i) A song is a "sum" of segments.

(ii) A song is a "product" of segments, i.e., the relative order of segment

A method called "voting" is based on assumption (i). We train the classifier on segments labelled same as the song to which they belong, and predict a song to be the genre that has the most of its segments.
For assumption (ii), we are faced the problem that songs have different number of segments in the data set. So we partition the songs into certain number of sections, and each section

is the average of equal number of consecutive segments, and we train the classifier on the averaged features of each section of the song. When the number of partition is 1, this method is equivalent to taking the average of each feature over all segments.

4 Evaluation

We used Gaussian naive Bayes model, logistic regression model, and feedforward neural network for the classifier. The FNN contains one hidden layer computed by sigmoid, and the output is produced by multinomial logistic regression (softmax), iterating 10 times with learning rate 0.01. Compare the results (All accuracies are average of 100 times experiments):

	Voting	1 Section	5 Sections	50 Sections
Gaussian NB	65.3%	73.4%	73.9%	73.5%
Logistic	72.0%	77.7%	70.6%	69.0 %
FNN	50.8%	53.3%	56.7%	56.7%

It can be seen that voting was outperformed the others. So assumption (ii) is more reasonable than assumption (i). Indeed music is not a simple addition of its subintervals; it has inherent logic. For Gaussian naive Bayes, we can see that it suffered from voting even more. This is because each song has different number of segments, and voting method creates an unrealistic statistics of the frequency of genres, and naive Bayes biased frequencies of classes.

Both accuracies of naive Bayes and logistic regression model decreases when the section number increase. This is probably because more sections gives more features to learn from, and this requires more data. But the classifiers have only about 300 data to learn from.

We feel that neural network can do better, so we increase the numer of iteration to 100, and did some parameter tuning.

Increasing number of sections without dropout:

	10 Sections	50 Sections	100 Sections	200 Sections
Accuracy	65.5%	68.4%	70.1%	72.5%

Tuning dropout rate of hidden layer with 10 sections:

Dropout	0.1	0.2	0.3
Accuracy	69.3%	66.5%	65.5%

It seems that enabling dropout can increase accuracy (reduce overfitting), but increasing dropout rate will require a larger set of features to learn from to ensure accuracy.

5 Conclusion

(i) Music should be treated as a sequence of data instead of a collection.

(i) Bayesian learning algorithms perform worse with lower data-feature ratio, whereas neural network can benefit from more features available.

(i) Among the model we experimented, logistic regression is the one that is pretty accurate and fast to train.

6 Recommendation

If time permitted, we would try one more model: HMM-FNN. The idea is to encode each segment (or section) to a state by its feature, and learn the probability of its transition, and train a neural network on these probabilities observed.

References

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.