Music Genre Classification

Archit Rathore - 12152
Margaux Dorido - EXY1420

Indian Institute of Technology, Kanpur
Department of Computer Science and Engineering

Abstract. This project was primarily aimed to create an automated system for classification model for music genres. The first step included finding good features that demarcated genre boundaries clearly. A total of five features, namely MFCC vector, chroma frequencies, spectral roll-off, spectral centroid, zero-crossing rate were used for obtaining feature vectors for the classifiers from the GTZAN genre dataset [5]. Many different classifiers were trained and used to classify, each yielding varying degrees of accuracy in prediction. An ensemble classifier based on majority voting was then created to incorporate all of the classifiers into one.

Key words: music, genre, classification, MFCC, chroma, ensemble classifiers, spectral features, GTZAN genre dataset

1 Introduction

Wikipedia states that "music genre is a conventional category that identifies pieces of music as belonging to a shared tradition or set of conventions." The term "genre" is a subject to interpretation and it is often the case that genres may very fuzzy in their definition. Further, genres do not always have sound music theoretic foundations, e.g. - Indian genres are geographically defined, Baroque is classical music genre based on time period. Despite the lack of a standard criteria for defining genres, the classification of music based on genres is one of the broadest and most widely used. Genre usually assumes high weight in music recommender systems. Genre classification, till now, had been done manually by appending it to metadata of audio files or including it in album info. This project however aims at content-based classification, focusing on information within the audio rather than extraneously appended information. The traditional machine learning approach for classification is used - find suitable features of data, train classifier on feature data, make predictions. The novel thing that we have tried is the use of ensemble classifier on fundamentally different classifiers to achieve our end goal.

2 Related Work

The most influential work on genre classification using machine learning techniques was pioneered by Tzanetakis and Cook [5]. The GTZAN dataset was created by them and is to date considered as a standard for genre classification. Scaringella et al.[2] gives a comprehensive survey of both features and classification techniques used in the genre classification. Changsheng Xu et al.[7] have shown how to use support vector machines for this task. Most of the work deals with supervised learning approaches. Riedmiller et al.[6] use unsupervised learning creating a dictionary of features. [4] gives a detailed account of evaluation of previous work on genre classification.

3 Dataset

We have used the GTZAN dataset from the MARYSAS website. This is the dataset used in [5]. It contains 9 music genres, each genre has 100 audio clips in .au format. The genres are - blues, classical, country, disco, pop, jazz, reggae, rock, metal. Each audio clips has a length 30 seconds, are 22050Hz Mono 16-bit files. The dataset incorporates samples from variety of sources like CDs, radios, microphone recordings etc. We split the datset in $0.9 : 0.1$ ratio and used 5-fold cross validation for reporting the results.

4 Preprocessing

The preprocessing part involved converting the audio from .au format to .wav format to make it compatible to python's wave module for reading audio files. The open source SoX[3] utility was used for this conversion.

5 Workflow

To classify our audio clips, we chose 5 features: Mel-Frequency Cepstral Coefficients, Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, Spectral Roll-off. These 5 features are appended to give a 28 length feature vector. Then, we used different multi-class classifiers and an ensemble of these to obtain our results.

6 Methodology

6.1 Extraction of features

5 features inspired by [1] were used to create a single feature vector.

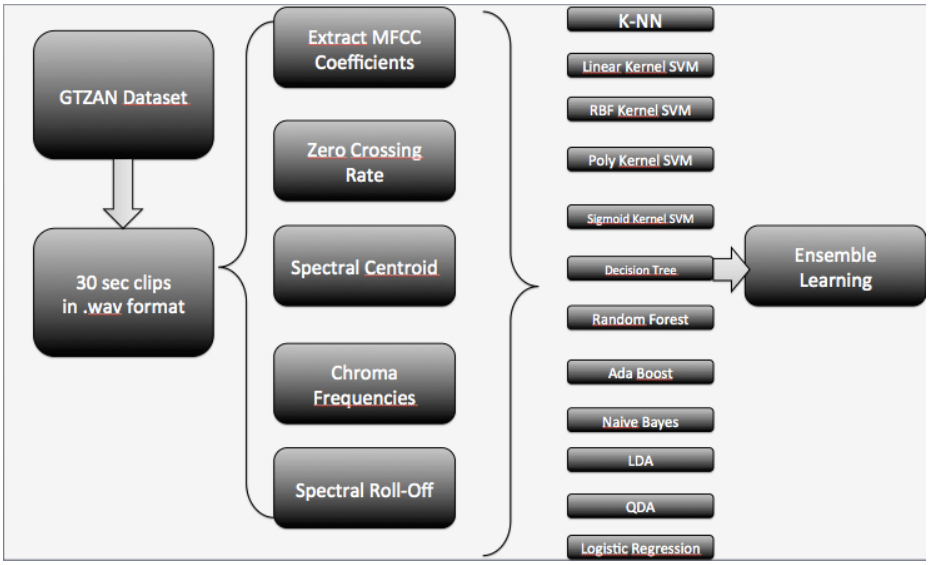

Fig. 1. Diagramtic representation of the method used

Mel-Frequency Cepstral Coefficients MFCC represents a set of short term power spectrum characteristics of the sound and have been used in the state-of-the-art recognition and sound categorisation techniques. It models the characterics of human voice. This features is a large part of the final feature vector (13 coefficients). The method to implement this feature is below :

- Dividing the signal into several short frames. The aim of this step is to keep an audio signal constant.
- For each frame, we calculated the periodogram estimate of the power spectrum. This is to know frequencies present in the short frames.
- Pushing the power spectra into the mel filterbank and collecting the energy in each filter to sum it. We will then know the number of energy existing in the various frequency regions.

$$M(f) = 1125 \ln(1 + f/700)$$

Fig. 2. Formula to work with Mel Scale

- Calculating the logarithm of the filterbank energies in the previous It enables humans to have our features closer to what humans can hear.
- Calculating the Discrete Cosine Transform (DCT) of the result. It decorrelates the filterbank energies with each others
- Keep first 13 DCT coefficients. We remove the higher DCT coefficients which can introduce errors by representing changing in the filterbank energies

From PracticalCryptography.com tutorial

Chroma Frequencies Chroma frequency vector discretizes the spectrum into chromatic keys, and represents the presence of each key. We take the histogram of present notes on a 12-note scale as a 12 length feature vector. The chroma frequency have a music theory interpretation. The histogram over the 12-note scale actually is sufficient to describe the chord played in that window. It provides a robust way to describe a similarity measure between music pieces.

Spectral Centroid It describes where the "centre of mass" for sound is. It essentially is the weighted mean of the frequencies present in the sound. Consider two songs, one from blues and one from metal. A blues song is generally consistent throughout it length while a metal song usually has more frequencies accumulated towards the end part. So spectral centroid for blues song will lie somewhere near the middle of its spectrum while that for a metal song would usually be towards its end.

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)\, x(n)}{\sum_{n=0}^{N-1} x(n)}$$

Fig. 3. Formula to calculate the Spectral Centroid

$x(n)$ is the weighted frequency value of bin number n
$f(n)$ is the center frequency of that bin

Zero Crossing Rate It represents the number of times the waveform crosses 0. It usually has higher values for highly percussive sounds like those in metal and rock.

$$zcr = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\}$$

Fig. 4. Formula to calculate the Zero Crossing Rate

s_t is the signal of length t
$II\{X\}$ is the indicator function (=1 if X true, else =0)

Spectral Roll-off It is a measure of the shape of the signal. It represents the frequency at which high frequencies decline to 0. To obtain it, we have to calculate the fraction of bins in the power spectrum where 85% of its power is at lower frequencies.

6.2 Classification

Once the feature vectors are obtained, we train different classifiers on the training set of feature vectors. Following are the different classifiers that were used -

- K Nearest Neighbours
- Linear Kernel SVM
- Radial Basis Function (RBF) Kernel SVM
- Polynomial Kernel SVM
- Sigmoid Kernel SVM
- Decision Tree
- Random Forest
- Ada Boost
- Naives Bayes
- Linear Discriminant Analysis (LDA) classifier
- Quadratic Discriminant Analysis (QDA) classifier
- Logic Regression

The parameters used for various classifiers were obtained by manual tuning. It was observed that any single classifier did not classify all the genres well. For example in the SVM with polynomial kernel worked well for most genres except blues and rock (See fig 5). This could have been due to the fact that many other genres are derived from blues.

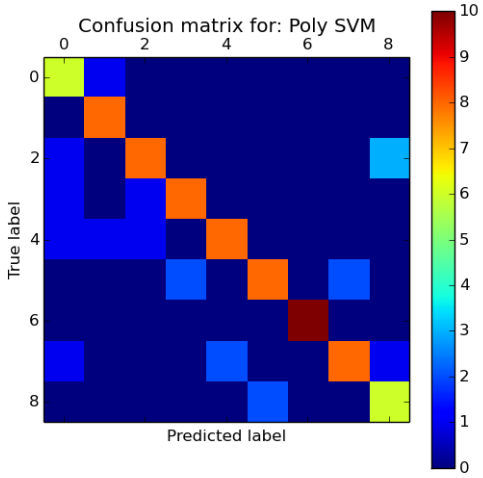

Fig. 5. Confusion Matrix for SVM classifier with polynomial kernel

To improve upon these shortcomings, we created an ensemble of the above clas-

sifiers. This ensemble classifier used the prediction of each classifier and run a majority voting heuristic to obtain the optimal class label for given test input. The weights used to average the classifiers were proportional to the accuracy of the classifiers. The ensemble seemed to give a better overall precision but the accuracy dropped by a small amount.

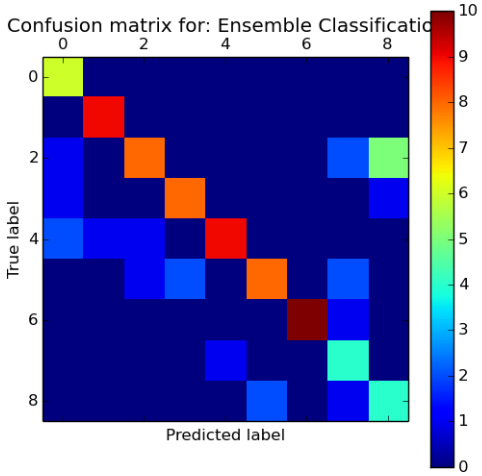

Fig. 6. Confusion Matrix for Ensemble classifier with polynomial kernel

7 Results

Following gives the classification results from current state-of-the-art[].

Table 1. Mean accuracies in GTZAN for each system[4]

System	System Configuration	Mean accuracy
AdaBFFs	Decision stumps	0.776
	Two-nodes tree	0.800
SRCAM	Normalized features	0.835
	Standardized features	0.802
MAPsCAT	Class dependent covariances	0.754
	Total covariance	0.835

We started by classifying a subset of 6 genres (country, reggae, metal, pop, rock,hiphop) using a SVM classifier with polynomial kernel. It gave an accuracy

of 82%. However when we tried extend this classifier to 9 classes the accuracy of classification dropped to 51%. In the next attempt, we tuned the hyperparameters for the classifiers and created an ensemble classifier out of them. Fig 7 shows accuracy values for different classifiers.

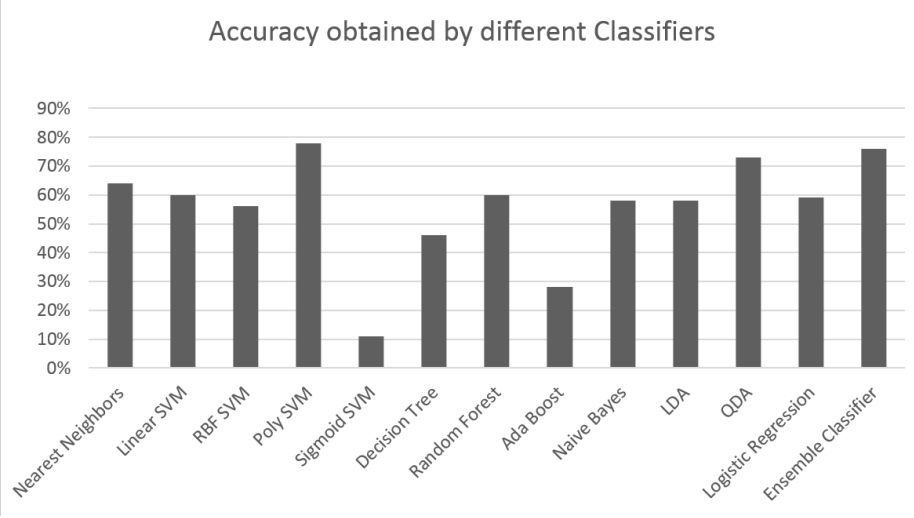

Fig. 7. Accuracy obtained by different classifiers

Classifier	Mean Accuracy	Mean Precision	Mean Recall
K-NN	0.64	0.70	0.64
Linear Kernel SVM	0.60	0.68	0.60
RBF Kernel SVM	0.55	0.78	0.56
Poly Kernel SVM	0.78	0.79	0.78
Decision tree	0.45	0.48	0.46
Random Forest	0.54	0.59	0.54
Adaboost	0.28	0.28	0.28
Naive Bayes	0.57	0.65	0.58
LDA	0.58	0.68	0.58
QDA	0.73	0.77	0.73
Logistic Regression	0.58	0.68	0.59
Ensemble Classifier	0.76	0.81	0.78

Table 2. Statistics for different classifiers

8 Conclusion

The classifier that works best is SVM with Polynomial Kernel. Some classifier are very efficient for some specific genres (like SVM with RBF Kernel for"country"). The highest verified accuracy on the GTZAN dataset is reported at approximately 84%[4]. The ensemble method improves upon Polynomial SVM classifier by reducing large errors in classifying specific genres. The genre-wise precision, recall and f-1 scores are given in the text file here.

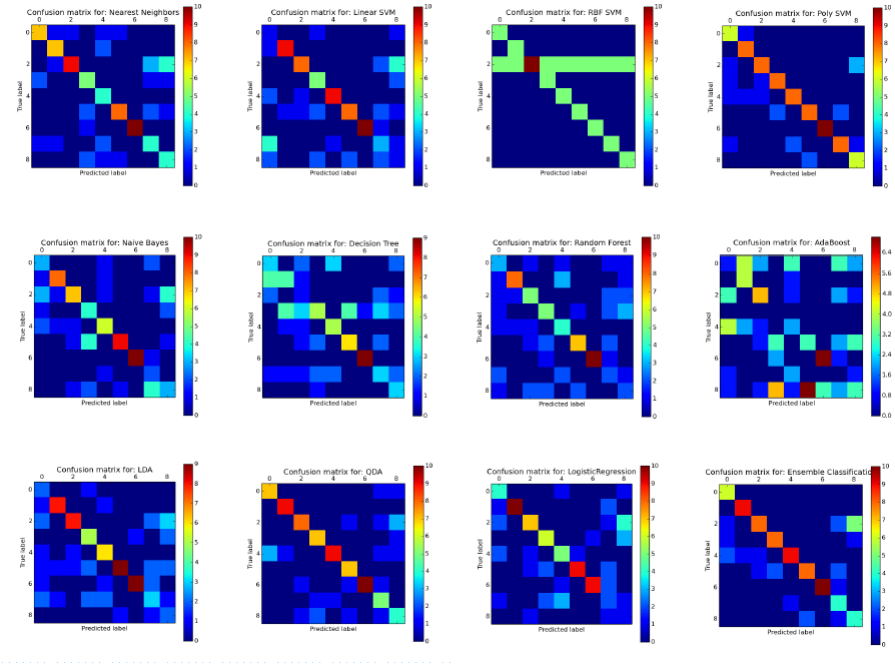

Fig. 8. Confusion matrices for various classifiers

References

1. Omar Diab, Anthony Manero, and Reid Watson. *Musical Genre Tag Classification With Curated and Crowdsourced Datasets*. Stanford University, Computer Science, 1 edition, 2012.
2. N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Process. Mag.*, 23(2):133–141, 2006.
3. Sox.sourceforge.net. Sox - sound exchange — homepage, 2015.
4. Bob L. Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.
5. G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.
6. Jan Wülfing and Martin Riedmiller. Unsupervised learning of local features for music classification. In *ISMIR*, pages 139–144, 2012.
7. Changsheng Xu, MC Maddage, Xi Shao, Fang Cao, and Qi Tian. Musical genre classification using support vector machines. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 5, pages V–429. IEEE, 2003.