MIREX 2014

Audio Similarity : An Ensemble Approach

Abhinava Srivastava
New Delhi, India
abhinva.srivastava1@gmail.com

Mohit Sinha
New Delhi, India
sinhamohit10@gmail.com

I. Abstract

In this submission we have built an ensemble using multiple feature extraction techniques and distance measures. The idea behind an ensemble is to reduce the variance of the model. In this submission we have used stacking to combine multiple models.

II. Introduction

This submission extends the ideas from previous submissions and uses a stack of different features and distance measures. Ensemble [1] uses multiple models for a particular task and makes the overall model more stable across multiple data sets. The entire process is implemented in Java, for feature extraction we've used the "*Audio Feature Extraction Toolkit from Vienna University of Technology*"[2], "*jAudio*"[3] and "*CoMIRVA*"[4]. We have implemented the different distance measures and the underlying feature summarizations in Java.

III. Features

We have used both timbral and rhythm features.

A. MFCC Single Vector

We have summarized 20 MFCC features over a texture window of 40 analysis frames by taking running mean and standard deviation to get a single vector representation. Euclidean L2 Norm has been used as the similarity measure [5] [6]

B. MFCC Multivariate Gaussian Distribution

We have approximated each track as a multivariate Gaussian distribution [7]. To describe the distribution we have used the mean vector and covariance matrix of 20 MFCC features computed for each analysis window of size 512 samples, sampled at 22050Hz. Jensen – Shannon Divergence has been used as the similarity measure.[8]

C. Statistical Spectrum Descriptor (SSD)

Statistical Spectrum Descriptor [9] - The spectrum is transformed into Bark scale. The audio track is described by the following statistical moments on the values of each of the 24 critical bands: mean median, variance, skewness, kurtosis, min-value and max-value. Euclidean L1 Norm has been used as the similarity measure

D. Fluctuation Pattern (FP)

Fluctuation Pattern [10] describes the amplitude modulation of loudness per frequency band and describes the characteristics of the audio which are not described by spectral similarity measure. Euclidean L1 Norm has been used as the similarity measure

E. Spectral Pattern

Spectral Pattern [11] characterizes a song's timbre via modeling those frequency components that are simultaneously active. It is a block level feature. Euclidean L1 Norm has been used as the similarity measure

IV. Distance Measures

We have used Euclidean Norms of vector representations and Jensen-Shannon Divergence for multivariate Gaussian representation.

A. Euclidean L1 Norm

It measures sum of the absolute distances in each dimension in feature space, also known as Manhattan Norm.

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|.$$

B. Euclidean L2 Norm

It is the length of line segment connecting two points in feature space.

$$\|\boldsymbol{x}\| := \sqrt{x_1^2 + \cdots + x_n^2}.$$

C. Jensen – Shannon Divergence

Is used to estimate the similarity between two Gaussians, is based on the Kullback-Leibler Divergence. We have used left-type Kullback-Leibler Centroid [12] of the two Gaussian distributions as an approximation.

$$\mu_m = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2$$
$$\Sigma_m = \frac{1}{2}\left(\Sigma_1 + \mu_1{\mu_1}^T\right) + \frac{1}{2}\left(\Sigma_2 + \mu_2{\mu_2}^T\right) - \mu_m{\mu_m}^T.$$

These are used to approximate the Centroid and the final Jensen Shannon divergence is computed using

$$JS(X_1, X_2) = \frac{1}{2}\log|\Sigma_m| - \frac{1}{4}\log|\Sigma_1| - \frac{1}{4}\log|\Sigma_2|.$$

V. Ensemble

All the individual models of similarity are built and then a linear combination [1] of these is tuned to maximize k-NN genre classification accuracy [13] at k = {1, 5, 10, 50, 100} and minimize variance using n-fold cross validation [14] where n = {5, 10} on the following data sets –

- GTZAN Genre Collection [15]
- A Benchmark Dataset for Audio Classification and Clustering [16]
- Custom data set of personal music collection

We shortlisted top 5 weight combinations for each of the datasets and then took the average of the weights. Following are the final contributing models and their respective weights.

Feature	Similarity Metric	Weight
MFCC Single Vector	L2 Norm	0.130434783
MFCC Multivariate Gaussian	Jensen-Shannon Divergence	0.347826087
Statistical Spectrum Descriptor	L1 Norm	0.086956522
Fluctuation Pattern	L1 Norm	0.260869565
Spectral Pattern	L1 Norm	0.173913043

VI. Conclusion

We manually looked for the optimal weights combination for each datasets (due to lack to time and resources) probably leading to the arrival at the local optima. This can be improved upon. The ensemble however was stable and performed as expected across all the three data sets.

References

[1] Ensemble Learning , http://en.wikipedia.org/wiki/Ensemble_learning

[2] Audio Feature Extraction Toolkit from Vienna University of Technology, http://www.ifs.tuwien.ac.at/mir/audiofeatureextraction.html

[3] jAudio, http://jaudio.sourceforge.net/

[4] CoMIRVA, www.cp.jku.at/comirva/

[5] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5):293 – 302, July 2002

[6] G. Tzanetakis, Music Information Retrieval, http://marsyas.cs.uvic.ca/mirBook

[7] M. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In Proceedings of the 6th International Conference on Music Information Retrieval ISMIR'05, 2005

[8] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On rhythm and general music similarity. In Proceedings of the 10th International Conference on Music Information Retrieval. ISMIR'09, 2009

[9] T. Lidy, A. Rauber. Evaluation Of Feature Extractors And Psycho-acoustic Transformations For Music Genre Classification

[10] E. Pampalk. Evaluation Of Feature Extractors And Psycho-acoustic Transformations For Music Genre Classification

[11] K. Seyerlehner .Content-Based Music Recommender Systems: Beyond simple Frame-Level Audio Similarity

[12] D. Schnitzer. Indexing Content-Based Music Similarity Models for Fast Retrieval in Massive Databases

[13] T. Pohle. Automatic Characterization of Music for Intuitive Retrieval. PhD thesis, Johannes Kepler University Linz, 2010.

[14] Cross-validation http://en.wikipedia.org/wiki/Cross-Validation_(statistics)

[15] G. Tzanetakis. GTZAN Genre Collection. http://marsyas.info/download/data_sets/

[16] Homburg, Helge and Mierswa,Ingo and Moller, Bulent and Morik, Katharina and Wurst, Michael. A Benchmark Dataset for Audio Classification and Clustering. In Joshua D. Reiss and Geraint A. Wiggins (editors), Proc. of the International Symposium on Music Information Retrieval 2005, pages 528--531, London, UK, Queen Mary University, 2005.