Classification of Music Genre using Neural Networks with Cross-Entropy Optimization and Soft-Max Output

Dharin Shah
Dept. of Information Technology
K.J.Somaiya College of Engineering

Chirag Sachdev
Dept. of Information Technology
K.J.Somaiya College of Engineering

Bhavik Shah
Dept. of Information Technology
K.J.Somaiya College of Engineering

ABSTRACT
In this paper, an abstract model to predict the genre of a music audio file is proposed (specifically a wave file).The output of the model is the probability distribution along the considered genres. A machine learning approach is employed. The adaptive learning process is modeled by neural networks with back-propagation as its learning algorithm and cross entropy as its optimization function. The emphasis is on feature extractors since the learning paradigm is well known to other applications. Simple Analysis on the Features were performed for appropriate selection.

General Terms
Music Classification, Optimization Problem.

Keywords
Auto-Encoder, Classification, Cross-Entropy, Machine Learning, Music Genre, Neural Networks, Soft-max Output.

1. INTRODUCTION
There are numerous studies that are investigated in the field of digital music and how it would be possible to enhance user's experience. A lot of untagged music files are being archived, while some contain assumed or false tags. However automatic genre classification is not an easy task considering music evolving within short periods. The best approach to modeling such kind of problem is by using statistical learning methods. Instead the whole graphics generation process can be Existing Studies ([4]) are based on classification by melody of a MIDI file. This paper uses a different approach both on feature extractors and learning paradigm while considering digital uncompressed audio files.

Selecting digital wave files instead of MIDI files was a decision targeted to digital archives which is growing each day. Altogether this paper provides a pragmatic software based approach to personalize the feature extractors and generate training set for the same.

2. EXTRACTION AND CLASSIFICATION SYSTEM
Extraction System. The Extraction system primarily uses a GNU toolbox [1] to extract certain features. The toolbox is tweaked for performance enhancements and requirements. Simple Clustering analysis was done for consideration of features along the feature space.

Recursive Feature Elimination. The general approach for dimensionality reduction was 'Recursive Feature Elimination'. It is a 3 step approach.

- Train the model and evaluate performance.
- Remove 10% of the weakest features and retrain on remaining features.
- Continue until drop in Performance.

Features. The Detail information about the features is listed in Table 1.

Classification System. The classification system essentially is a Neural net. The Neural net consists of 3 layers with number of features as the input and 4 output neurons alongside experimentally evaluated number of neurons in the hidden layer since there is no heuristics for the same.

Further study shows that for performance gains, rather than using Mean Squared Error as our optimization function, cross entropy function was used.

$$H(p,q) = \sum p(z).\log\left(q(z)\right)$$

Output. The output would be a probability distribution along considered genres.

$h_j = \frac{e^{z_j}}{\sum e^{z_j}}$ where j is the output neuron number.

Clearly both functions are differentiable.

3. FEATURES

Table 1 : List of Features

	Feature	Brief Description
Dynamics	Rms (Root-Mean-Square)	The global energy of the signal computed by taking root average of the square of the amplitude
	Low Energy	The energy curve can be used to give an assessment of the temporal distribution of energy, in order to see if it remains constant throughout the signal, or if some Frames are more constructive than others. One way to estimate this consists in computing the low energy rate, i.e. the percentage of frames showing less-than-average energy. (number of values below the average rms value)
Pitch	Range	Difference Between Highest and lowest Pitch.
	Number of Notes	Number of Pitches Played at-least once.
Timbre	Roll-Off	Estimate the amount of High frequency energy signals such that certain fraction of total energy (Default 85%) is contained below that frequency (roll-off Frequency).
	Brightness	Measuring energy above certain frequency (Specifying the cutoff frequency).
	Irregularity	Degree of variation of successive peaks of the spectrum.
Rhythm	Dominant tempo	Frequency of highest tempo bin (in bpm)
	Second Dominant Tempo	Frequency of the second highest tempo bin.
	Combined dominant tempos	Combine frequency of the 2 dominant tempos.
	Dominant Tempo Strength Ratio	Ratio of the Frequency of two highest tempo bins.
	Dominant Tempo Ratio	Ratio of the two dominant Tempos.
	Tempos	Number of tempo bins with frequency > 20%, 10%, 25%.
	Average Event Frequency	Number of note onset/second

Generating Training Set. The training set is automatically generated by determining the number of directories which make up the classes. (Directories are considered classes which contain the audio files.)

4. OVERVIEW OF THE SYSTEM

Beat Spectrum. The Images show the beat-spectrum of a four audio files for their genres.

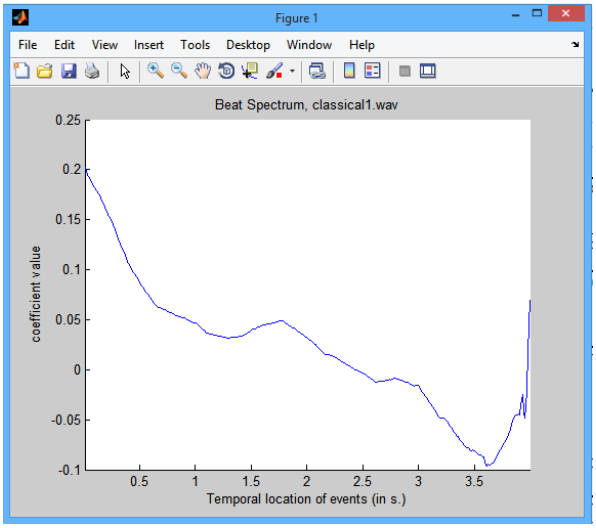

Figure 1 : Classical

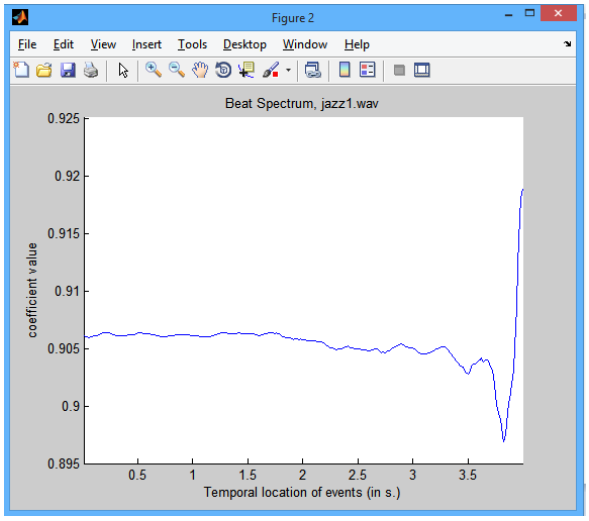

Figure 2 : Jazz

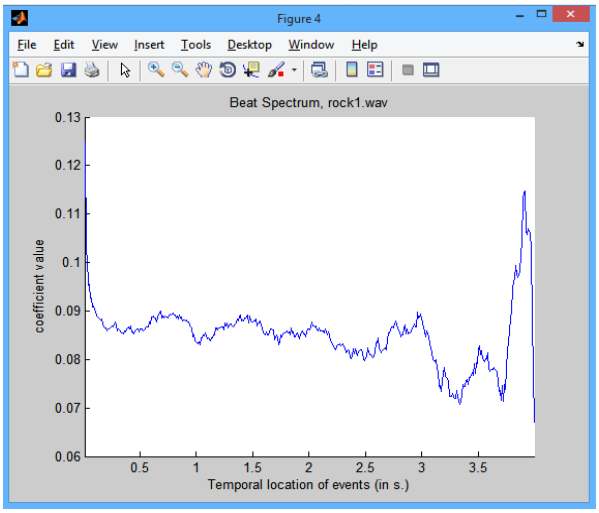

Figure 3 : Rock

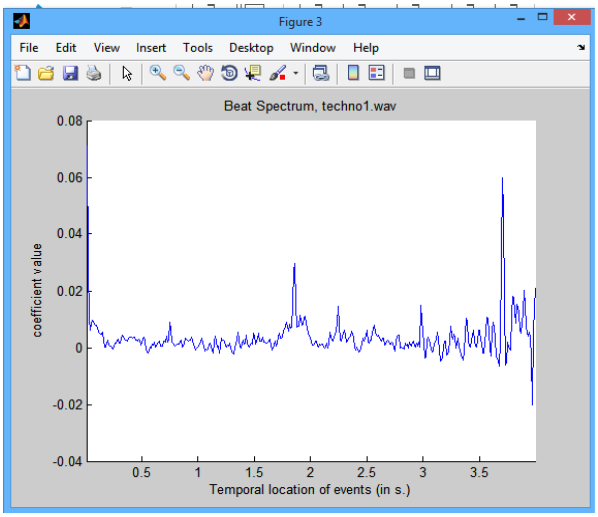

Figure 4 : Techno

Onset Curve. Using the onset curve, the tempo can be determined. As we can see, there is a high number of note onsets in the techno file while relatively few in the classical. This and other features help determining the Genre.

5. RESULTS

Auto-Encoder. We generated 16 features for each audio file of each Genre. And then collected 400 samples for our training set (100 for each genre) and used Auto-Encoders for dimensionality reduction. Experimentally selected 10 neurons in the hidden layer for the auto-encoder.

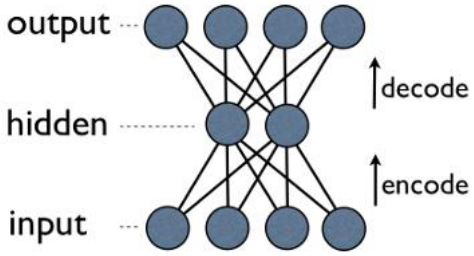

Figure 5 : Auto-Encoder

The Results after training with auto-encoder is pretty underwhelming.

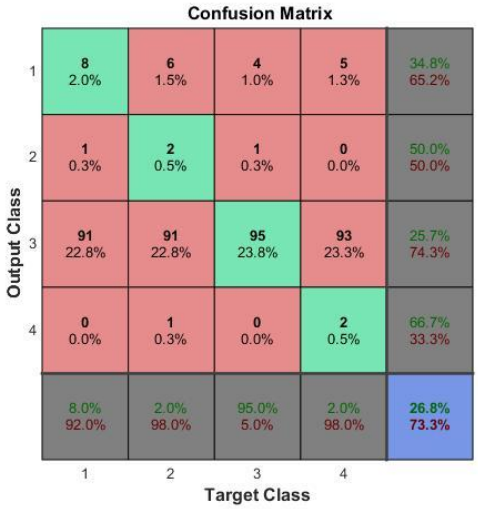

Figure 6 : Auto-Encoder Training Results

We can therefore infer that some of relatively high correlated features were removed and thus we get high error rate.

PCA. We selected principal features after performing *Principal Component Analysis* on the Data but the results were unsatisfactory with high average error.

Two Layer Network. Results improved after the use of simple feed-forward network without auto-encoder training.

The training Samples were divided randomly for each attempt at training. The average error percentage for training, validation and testing is listed in Figure 7.

- Training
- Validation
- Testing

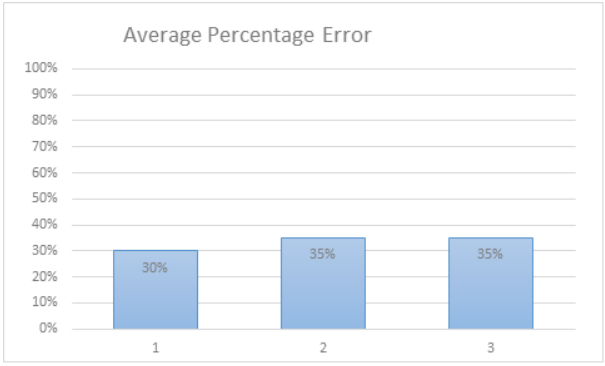

Figure 7 : Average Percentage Error

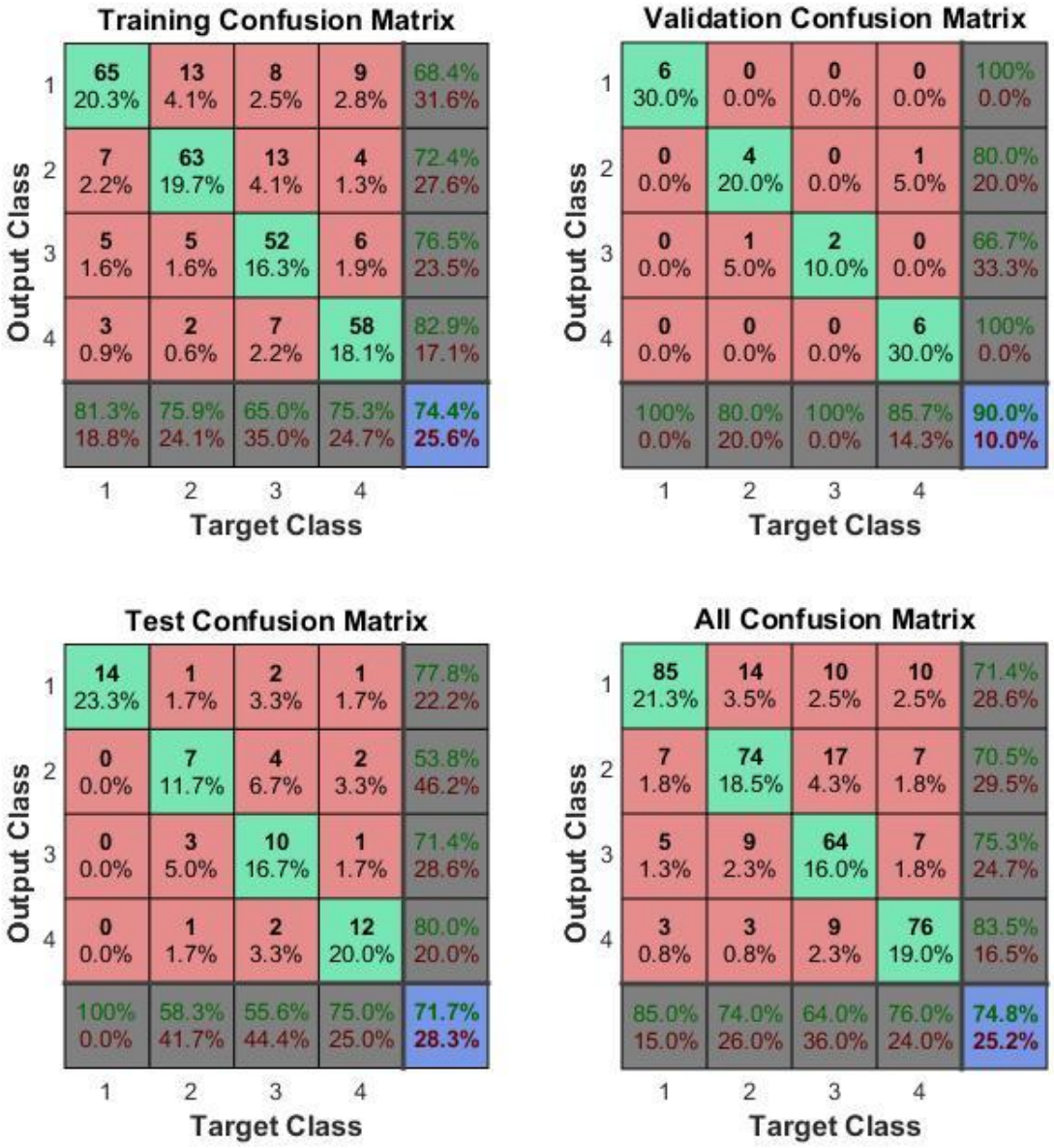

Figure 8 : Final Training, Validation and Testing Results

4.1 Inference

We analyzed the experiments and the results of the aforementioned studies, and two prominent inferences were drawn.

The first is Neural Network proves to be an excellent Classification model for such kind of application in which data is clearly not linearly separable and with few changes prove to be efficient too.

Secondly, number of features apart from Rhythm contribute to the optimization of weights.

6. CONCLUSION

Enhancements in the Model. The Classification Model presented in this paper does not take sub-genres into considerations. A model by [3] can be integrated in our model which says output of the activated neuron will be connected to another neural network to identify the sub-genre.

Future Steps in Research. The next step would be to increase the dataset as more the data, better the results will be. Also increase in the number of genre categories will be considered. Generating the Training set requires a large amount of memory, therefore steps to efficiently compute the training data will be taken.

Feature Selection Approaches. Experimenting with new features and performing statistical test for reduction is a great way to deal with high dimension data, but does not guarantee better results. Therefore the iterative approach works for now until better dimensionality reduction techniques are produced.

7. REFERENCES

[1] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction From Audio," *International Conference on Digital Audio Effects*, 2007.

[2] M. A. Nielsen, Neural Networks and Deep Learning, Determination Press, 2015.

[3] C. Mckay, "Using Neural Networks for Music Genre Classification," *Faculty of Music.*

[4] M. K. Shan and F. F. Kuo, "Music Style mining and classification by melody," *IEICE Transactions on Information and Systems*, Vols. E86- D(3).655-659..

[5] D. Cereghetti, O. Lartillot, K. Eliard, W. J. Trost, M. A. Rappaz and D. Grandjean, "Estimating tempo and metrical features by tracking the whole metrical hierarchy," *3rd International Conference on Music and Emotion*, 2013.

[6] O. Lartillot, "Computational analysis of maqam music: From audio transcription to musicological analysis, everything is tightly intertwined," in *Acoustics 2012 Hong Kong Conference*, Hong Kong.