A Comparative Study of Classifiers for Music Genre Classification based on Feature Extractors

Pradeep Kumar D, Sowmya B J, Chetan, K G Srinivasa
Department of CSE
M S Ramaiah Institute of Technology,
Bangalore, India
{pradeepkumard, sowmyabj, chetanshetty, kgsrinivas}@msrit.edu

Abstract—**The objective of this paper is to do a comparative study to detect and classify music files automatically based on its genre by using various classification algorithms. Music genre classification is a popular problem in the domain of Music Information Retrieval (MIR) used in many music streaming platforms such as Pandora which is a automated music recommendation service based on the Music Genome Project, that suggests songs to users based on similarity of songs that the user is interested in. In this paper we have done a comparative study using various machine learning classification algorithms to classify music file based on its genre. We have used both Fast Fourier Transform (FFT) and Mel Frequency Cepstral Coefficients (MFCC) to featurize our data, the latter out of which was recommended in a previous study.**

Keywords— Fast Fourier Transform, mel Frequency Cepstrum, logistic regression, Decision trees, Support vector machine, recurrent neural networks, kth nearest neighbour

I. Introduction

Music genre classification is a popular problem in the domain of Music Information Retrieval (MIR) used in many music streaming platforms such as Pandora[1] which is an automated music recommendation service powered by the Music Genome Project, that suggests songs to users based on similarity of songs that the user is interested in. In this paper we have done a comparative study using various machine learning classification algorithms to classify music file based on its genre. We have used both Fast Fourier Transform (FFT) and Mel Frequency Cepstral Coefficients (MFCC) to characterize our data, the latter out of which was recommended in a previous study [3]. The aim is to automatically classify songs based on genre. In order to do this we build different classifiers so as to compare the accuracy of each of the classifiers. The different genre that we have considered are jazz, metal, rock and pop to build different classifiers namely Logistic, Regression, Kth Nearest Neighbors, Decision trees, Support Vector Machine (SVM),Recurrent Neural Network. The paper compares the accuracies of each of the classifiers and also suggests which method of data preprocessing yields a better result. We have used two techniques namely Fast Fourier Transforms (FFT) and Mel-frequency cepstral coefficients. These techniques can be used by application which would want to suggest songs to a particular user. For example if a user is found to listen to a particular genre more frequently, then songs of similar genre can be suggested to the user.

II. LITERATURE SURVEY

The problem of Music Genre Classification has been approached in various ways. Tzanetakis, et al.[9], Sam Clark, et al.[10] and other papers have approached this problem specific to certain algorithms. John cast, et al.[11], suggest that features coupled with MFCCs improve classification accuracy. Martin, et al.[12], suggests that the temporal modulations of MFCCs are important for classification and its performance is better in comparison with the Standard low-level feature set because of its increased ability to classify background crowd noise and popular music. C H lee, et al.[13] proposed a novel feature set for music genre classification based on cepstral (MFCC) features which achieves higher classification accuracy. Most work on this problem is done by using specific data preprocessing techniques like FFT or MFCC exclusively to convert audio data and the training of classification algorithms is done based on these data values. We believe there is no prior work analyzing how various algorithms perform using MFCC and FFT data values to the extent of our knowledge. Hence this paper is a study on such an analysis. From the analysis it is clear that using MFCC data values gives better results overall than using FFT values. The Simpler algorithms such as Logistic Regression and Kth Nearest Neighbors did fairly well in comparison to superior algorithms such as Recurrent Neural Networks and Support Vector Machines. The highest accuracy reached was 86% using Neural Networks.

III. MUSIC DATASET

The data set used for this study is GTZAN [4] genre collection data. This data set contains 1000 songs each of which is 30 seconds long. These songs are classified into 10 genres namely: Rock, Metal, Classical, Jazz, Pop, Reggae, Classical, Blues, Hip-hop, and Disco. We have selected four genres out of these for our analysis namely: Jazz, Metal, Rock, and Pop. Each of these songs are sampled at a rate of 22050 Hz. Hence total number of data points for each song is 661500. All these songs are available in .au format. We have chosen 320 songs

978-1-5090-1623-5/16/$31.00 ©2016 IEEE

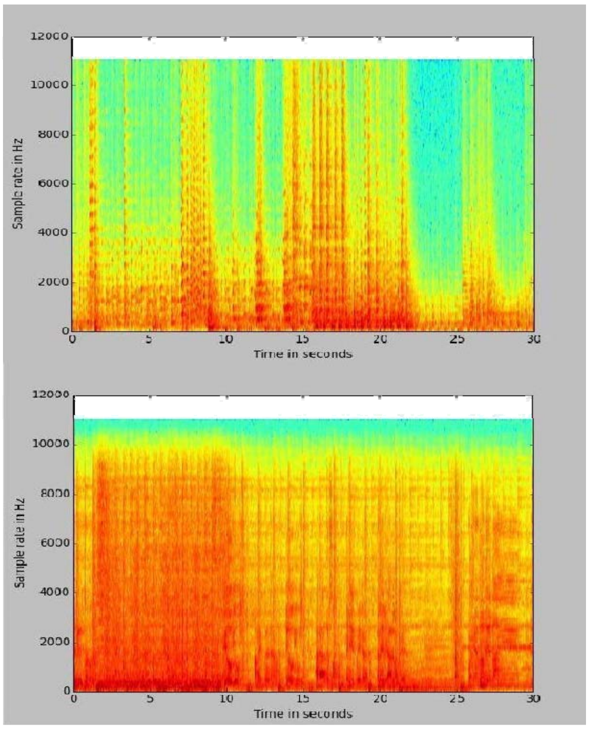

Fig 1: Spectrogram of two sample songs (top: jazz song, bottom: rock song)

(80%) for our training data and 80 songs (20%) for test data for measuring results. Spectrogram of a rock song and a jazz song are shown in Figure 1. It can be observed that there is pattern of how songs of different genre are sampled at different times. This forms the basis of our analysis.

IV. DATA PROCESSING AND NORMALIZATION

To convert our audio data for processing, two popular algorithms used in speech recognition namely: Fast Fourier Transform and Mel Frequency Spectrum (MFCC) were used to extract individual frequency intensities from the raw sample readings. Scipy's FFT [5] algorithm and Scikits toolbox's [6] MFCC algorithms were used to convert the audio data.

A.Fast Fourier Transforms

A FFT is a mathematical method to obtain DFT(Discrete Fourier Transform) for a sequence or the inverse of a sequence. A Fourier analysis is performed to obtain a frequency domain representation of the original domain. Rapid computation of this transform by the factorization of Discrete Fourier Transform Matrix into a sparse factors' product is job done by an FFT. Because of which, the complexity of obtaining a DFT is reduced to O(n log n) from O(n^2) to , where n represents the data size. FFT plot of two sample songs are shown in Figure 2.

Let x_0, ..., $x_{N\text{-}1}$ denote complex numbers. The DFT is obtained by the formula

$X_k = \sum$n to n=1 $xe^{-i2\pi kn/N}$

where k = 0,1,….., N-1

B.Mel Frequency Cepstral Coefficient

Mel Frequency Cepstral coefficients emphasize on obtaining the exact structure of the audio signal to extract linguistic features and discard the background noise. The linear cosine transform of a logarithmic power spectrum on a Mel scale which is non-linear is the basis of its calculation. A collection of Mel Frequency Cepstral Coefficients form a Mel-frequency cepstrum.

The Mel scale is calculated as

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

s(n) be the time domain signal.

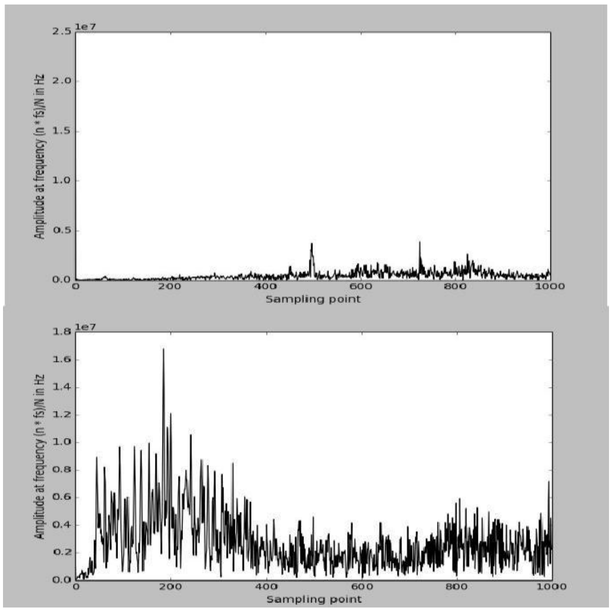

Fig 2: FFT Plot of two songs (top: jazz song, bottom: rock song)

The following computations are made to obtain the Discrete Fourier transform of the frame -

$S_i(k) = \sum$n to n=1 $s(n)h(n)e^{-j2\pi kn/N}$ *where* $1 \leq k \leq K$

h(n) – is the Analysis window for N samples

K- Length of the Discrete Fourier Transform

The power spectral estimate based on the Peridogram for $s_i(n)$, which is the speech frame is specified by

$P_i(k) = |S_i(k)|^2$

This peridogram is further processed to obtain 26 cepstral co-efficients, DCT is applied on 26 log filter bank energies which is called the MFCC.

V. SYSTEM DESIGN

The architecture consists of an input dataset which is first pre processed using two components. The pre processed data is fed into different forms of classifiers to train each of them. The classified data is then tested using the test data or the required music track to be classified. The diagram of the system architecture is shown in the figure 3.

A.Algorithm

1. Convert data into easy readable format.
2. Preprocess the data using FFT and MFCC.
3. Train the classifier using various classification algorithms.
4. Run the classifier on a GPU.
5. Compare performances between CPU and GPU.
6. Predict the genre of the user given data.

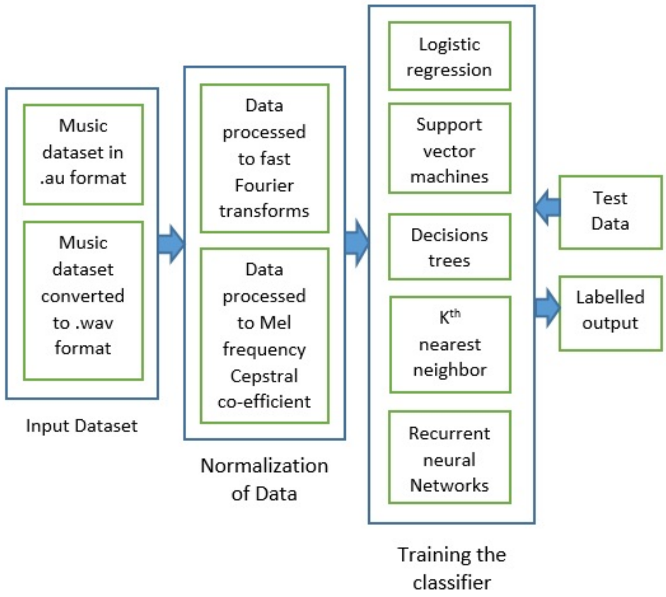
Music dataset in .au format
Music dataset converted to .wav format
Input Dataset
Data processed to fast Fourier transforms
Data processed to Mel frequency Cepstral co-efficient
Normalization of Data
Logistic regression
Support vector machines
Decisions trees
K^th nearest neighbor
Recurrent neural Networks
Training the classifier
Test Data
Labelled output

Fig 3: System design

B.Logistic Regression

In cases, where there is a need to predict the probability of an outcome that only has two values, Logistic Regression is used. The prediction is based on the usage of one or several predictors that may be numerical or categorical. It produces a logistic curve whose values are between 0 and 1. Rather than the probability, logistic curve is plotted using the natural log of odds of the target variable. The predictors need not be normally distributed or have equal variance in each group. We make use of One vs. All model for Logistic Regression in the Implementation as this problem is a multi-classification problem. The logistic regression equation is written in terms of an odds ratio by using a simple transformation as:

$$p/(1-p)=\exp(b_0+b1_x)$$

where b_0 is the Regression constant and b_1 is the slope that defines steepness of the logistic curve.

By taking the natural log on both sides, a linear equation of the predictors is derived which can be written in terms of log-odds (logit):

$$\ln(p/(1-p))=b_0+b1_x$$

C.Kth Nearest Neighbors

K nearest neighbors is an algorithm that stores all the currently classified new cases based on previous available cases by computing similarity measures. A majority vote of a case's neighbors classify it and each of the case being assigned to the class most common among its K nearest neighbors is computed by a distance function. If the K value is equal to 1, then it is simply allocated to the class of its nearest neighbor. All the distance measures are only valid for continuous variables. We make use of

Manhattan Distance measure in our application which is given as:

$$d=\sum_{j=1}^{k}|x_i - y_i|$$

Hamming distance must be used in cases when categorical variables are used .when both numerical and categorical variables are present in the dataset, numerical variables between 0 and 1 must be standardized. Choosing the optimal value for K is done best by first inspecting the data. Generally, a higher value for K is used to reduces the overall noise. Cross-validation is another way to determine a good K value by using an independent dataset to validate the K value. The optimal K value for most of the datasets has been in the range 3-10. A K value of 5 was used in the implementation of this algorithm for our analysis.

D.Support Vector Machine

Support Vector Machine (SVM) classifies by finding the hyperplane which maximizes the margin between the two classes. support vectors are vectors that define the hyperplane. SVM starts by defining an optimal hyperplane, it may have a penalty term for misclassifications and finally the data is mapped to a high dimensional space.

E.Decision tree

A decision tree constructs regression or classification models as a tree structure. Dataset is divided into smaller subsets and simultaneously a decision tree is developed incrementally. A decision node has two or more branches. The leaf node in the tree represents a decision or classification. The topmost decision node is called as the root node and corresponds to the predictor. The main algorithm used for building decision trees called ID3 and makes use of a top-down, greedy search through the space of possible branches with no backtracking. ID3 makes use of *Entropy* and *Information Gain* to build a decision tree[11].A decision tree is constructed top-down from a root node and entails partitioning of data into subsets containing instances with similar values. Samples homogeneity is calculated by ID3 algorithm using entropy. The entropy of the sample will be 0, if it is totally homogeneous and entropy will be one if it's equally divided.

Entropy using frequency table for a single variable can be calculated as follows

$E(S)=\sum^{c}_{j=1}|x_i - y_i|$

Entropy using frequency table for two variables

$E(T,X) = \sum\sum_{c€x} (c\ )E(c\ )$

F.Recurrent Neural Network

In Recurrent neural networks, a directed cycle is formed in the connections that exist between units. An internal state of the neural network is created permitting it to express a dynamic temporal behavior. In contrast to feed-forward networks, the internal memory of the recurrent neural networks can used for processing random input sequences. Handwriting recognition or speech recognition make use of RNNs mainly because of the same reason. A schematic representation of RNN [8] is shown in Figure 4.

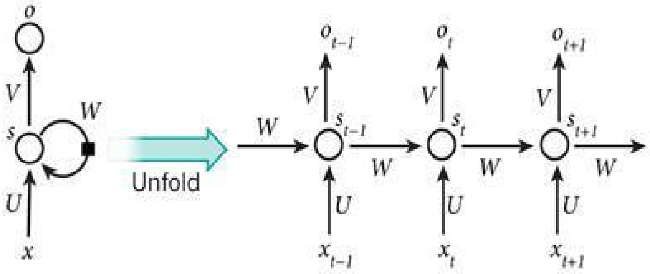

Fig 4: Unfolding Recurrent Neural Network.

The Long Short Term Memory Model (LSTM)[2] RNN has been implemented in this analysis as it is superior to other RNN models in predicting time-series data[2]. The general model of an LSTM RNN is shown in Figure 5.

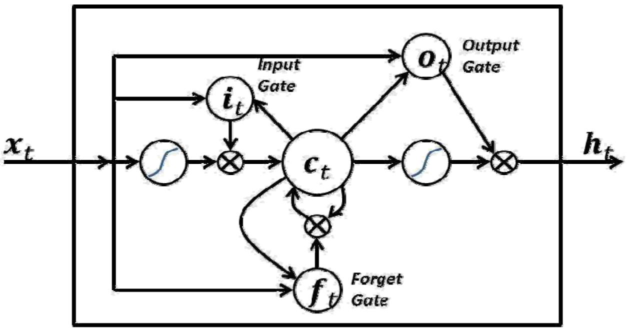

Fig 5: A simple LSTM gate with only input, output, and forget gates. (Source: Wikipedia LSTM)

If the weights are parameterized as W_i , W_f , W_o , W_c , U_i , U_f , U_o and the bias-vectors b_i , b_f , b_c, b_o , the input gate i_t , forget gate f_t and output gate o_t are computed as shown below:

i_t = sigmoid ($W_i x_t + U_i h_{t-1} + b_i$) f_t
= sigmoid ($W_f x_t + U_f h_{t-1} + b_f$) o_t
= sigmoid ($W_o x_t + U_o h_{t-1} + b_o$)

VI. RESULTS

The classification accuracy varied for different machine learning classification algorithms. Comparative analysis has been done on FFT data and MFCC data and results of these algorithms on the aforementioned data are tabulated in the TABLE I.

TABLE I: ACCURACY TABLE

Classification algorithm	Accuracy (%)	
	FFT	MFCC
Logistic Regression	72.25	67.5
Kth Nearest Neighbors (KNN) (n = 5)	65	67.5
Support Vector Machine (SVM)	41.5	82.55
Decision tree (max depth = 5)	57.5	77.5

The comparison of accuracy of classification algorithms used for both types of data is shown in the graph in Figure 6.

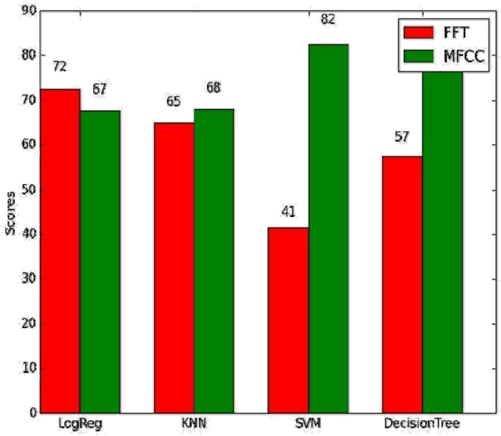

Fig 6: Comparitive projection of classification algorithms against FFT and MFCC data

It can be observed that these classification algorithms performed better with MFCC data, with only Logistic Regression algorithm being an exception, falling short by a small factor. Hence an LSTM model of RNN was trained only with MFCC for 100 epochs on a GPU (NVIDIA GTX Titan). The accuracy graph can be seen in Figure 7. The total time for training this model was approximately 50 minutes.

VII. CONCLUSIONS AND FUTURE WORK

From the above analysis it is clear that using MFCC data values gives better results overall than using FFT values.

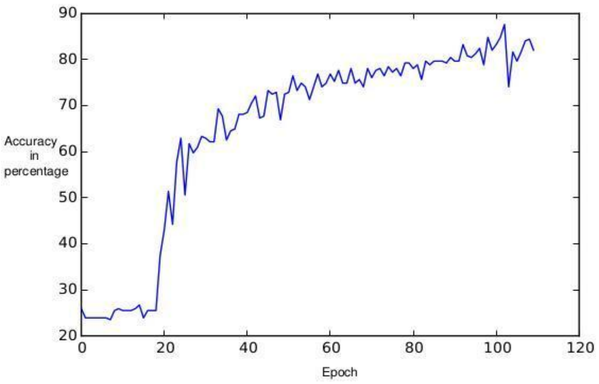

Fig 7: Accuracy plot for RNN

The Simpler algorithms such as Logistic Regression and K Nearest Neighbors did fairly well compared to superior algorithms such as Recurrent Neural Networks and Support Vector Machines. The highest accuracy reached was 86% using Neural Networks.

The analysis uses four genres out of a total of 10 genres in the GTZAN dataset. We have also used four sharply contrasting genres for obtaining better results. This analysis could be extended to train and classify all 10 genres.

Our analysis did not include testing how well each classification algorithm predicts individual genres. This could be done by plotting Receiver Operator Characteristic Curves (ROC) and/or building a confusion matrix. This will help better analyze the performance of various algorithms specific to each genre. The Kth Nearest Neighbors algorithm was implemented using Manhattan Distance. This analysis could be extended by replacing the existing Manhattan Distance with Kullback-Lieber (KL) Divergence while implementing KNN.

Another interesting analysis could be to implement unsupervised algorithms such as K-Means to cluster the songs based on genres and the results could be compared with the supervised algorithms

REFERENCES

[1] Pandora Internet Radio : https://en.wikipedia.org/wiki/Pandora_Radio

[2] Long Short Term Memory RNN Wikipedia: https://en.wikipedia.org/wiki/Long_short-term_memory.

[3] Fu, A., Lu, G., Ting, K.M., Zhang, D. “A Survey of Audio-Based Music Classification and Annotation” IEEE Transactions on Multimedia.

[4] GTZAN Dataset: "Musical genre classification of audio signals" by G. Tzanetakis and P. Cook in IEEE Transactions on Audio and Speech, Processing 2002. http://marsyasweb.appspot.com/download/data_sets.

[5] Spicy: Discrete Fourier Transform Pack: http://docs.spicy.org/doc/spicy/refernce/fftpack.html

[6] Talkbox: A set of python modules for signal/speech processing: https://scikits.appspot.com/talkbox.

[7] A list of Classification Machine Learning Algorithms: http://www.saedsayad.com/classification.htm.

[8] Recurrent Neural Networks Wipedia: https://en.wikipedia.org/wiki/Recurrent_neural_network.

[9] George Tzanetakis, Georg Essl, Perry Cook. “Automatic Musical Genre Classification Of Audio Signals”, ISMIR, 2001.

[10] Sam Clark, Danny Park, Adrien Guerard. “Music Genre Classification Using Machine Learning Techniques”

[11] Cast, John, Chris Schulze, and Ali Fauci. "Music Genre Classification."

[12] McKinney, Martin F., and Jeroen Breebaart. "Features for audio and music classification." ISMIR. Vol. 3. 2003.

[13] Lee, Chang-Hsing, et al. "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features." IEEE Transactions on Multimedia 11.4 (2009): 670-682.