Project 3

Audio categorization

Anupkumar Nagaraj Joshi

101880602

Aimee Ciane Nyambo

101880199

1. Methods of data representation:

    a. MFCC: The Mel frequency cepstral coefficients (MFCCs) of a signal is a set of features (10-20). This helps to describe the shape of a spectral envelope. This can be used to model human voice.

    b. Spectrogram: A spectrogram describes loudness of the sound. One can observe the different energy level given a time. It is similar to the heat map i.e, loudness can be equated to color at that point (ranges from dim to bright).

    c. Power spectrum: It is representation of the sound where vibration at each frequency is shown. It is presented as graph of pressure as a function of frequency. It is measured in Hz i.e, number of vibrations per second.

2. Classifiers used and Reason:

    1. ANN: Artificial Neural Networks (ANN) are multi-layer fully-connected neural nets. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. The network is made deeper by increasing the number of hidden layers.A given node takes the weighted sum of its inputs, and passes it through a non-linear activation function. This is the output of the node, which then becomes the input of another node in the next layer. The signal flows from left to right, and the final output is calculated by performing this procedure for all the nodes. Training involves assigning appropriate weights using back propagation method.

        ANN are flexible but strong deep learning models. They are known to be universal function approximators ie, they can used to model any complex functions. And making networks deeper make them more beneficial making them able to do more non-linear transformations of the input and drawing a more complex decision boundary.

    2. CNN: Its machine learning algorithm mainly used to deal with images and video categorization. It is made up of convolution layers. The main aim of the layer is to extract various features. The deeper the network, more the features identified. Further, Pooling layer helps in obtaining matrix having main features. At last fully connected layers, the image is flattened to get one column vector. The output is fed to the further network and using back propagation, features are divided into dominant and non-significant ones and then classified.

        CNN are suitable because it deals with images. Music is more effectively represented using images like spectrogram. Hence CNN are worth a shot.

3. SVM: It is the method where features are extracted and plotted. The main aim is to divide data such that margin is wider. The hyperplane is term used. And near by data which helps to increase the hyperplane are called support vector. It involves fair division of the data.

   It is being used because this algorithm helps clustering of the data. Since our problem involves categorizing the music, SVM is more appropriate. SVM is powerful because it can be used for classification or regression problem. And also, for linear and non linear division which makes it more flexible.

3. Rationale behind the selection for data representation and classifiers:

ANN is used along with features extracted in csv files. Features include MFCC, spectroid, zero crossing, RMSE and centroid. ANN is used since it is strong deep network. All the features are included in input layer and hidden layers involve assigning appropriate weights so that activation function is activated. This assignment of weight and biases is obtained by backpropagation. ANN is more elegant and powerful algorithm considering our problem. The data is represented in several features in csv file. 26 features are considered in which MFCC comprises of a set of 20. The digital representation of the data is faster to access.

SVM is used along with features extracted in csv files. SVM is better option for our problem because it involves dividing the data/features such that margin is more. Hence it helps better in classification problems. Features like MFCC, centroid,RMSE stored in csv is faster and elegant method of data representation.

CNN is mainly uses convoluted layers which help in identifying different features. CNN mainly deals with images and videos for input layer. Hence spectrogram is used. Spectrogram provides better picture of sound by defining by its loudness. Deeper the network, more important features are identified. Hence CNN is more suitable for music categorization.

4. Accuracies for different classifiers along with confusion matrix

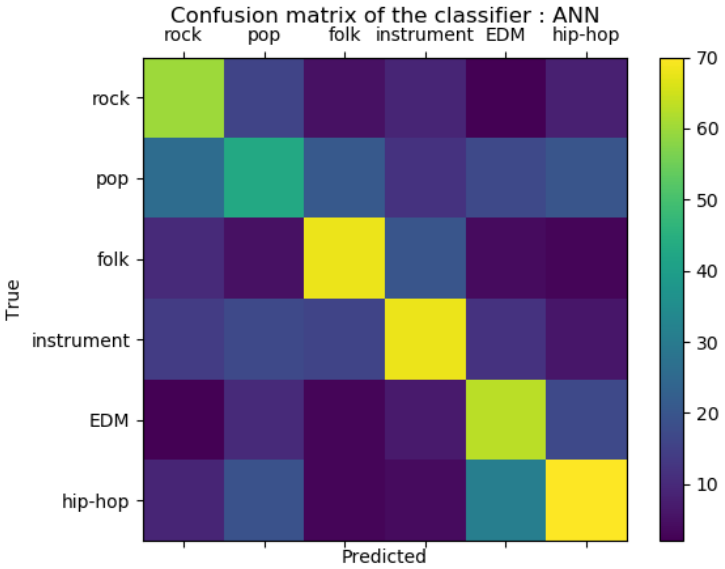
1. ANN:

   Accuracy obtained: 54.6%

   Confusion Matrix:

```
[[60  16  5   9   2   8]
 [26  43  21  12  17  20]
 [10  5   68  20  4   3]
```

[14 17 16 68 12 6]
[2 10 3 7 63 17]
[9 19 3 4 31 70]]



Confidence Interval: (0.39,0 .65)

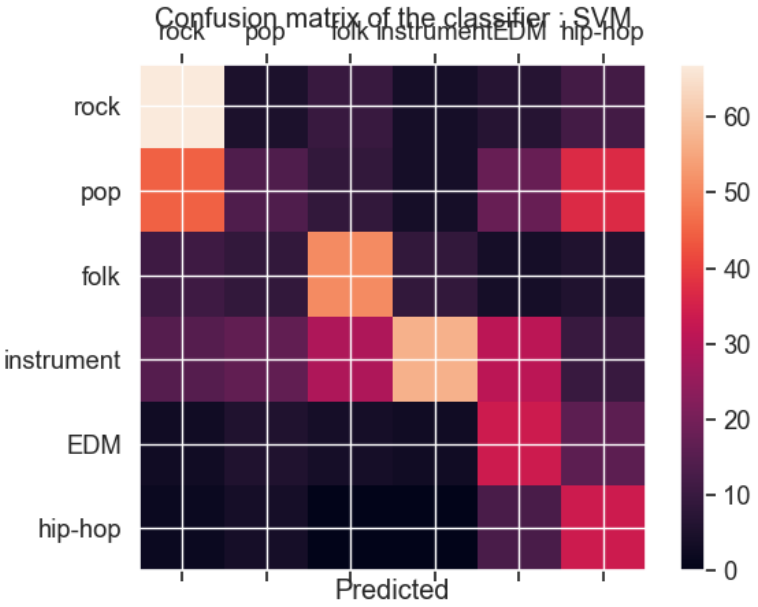
2. SVM:

   Accuracy obtained: 49.8%

   Confusion Matrix:

[[67 5 10 4 7 12]
[45 14 9 4 18 37]
[11 9 51 9 4 6]
[15 17 29 57 31 10]
[3 6 4 3 34 16]
[2 4 0 0 13 34]]

   Confidence interval: (0.4 ,0.6)

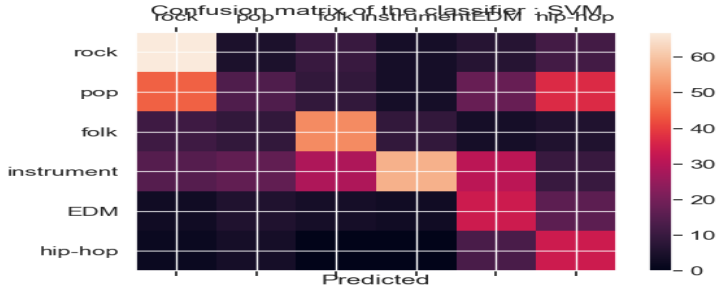


CNN:

Accuracy obtained: 35%

Confusion Matrix:

[[63 15 11 4 7 9]
[35 14 19 4 10 27]
[10 9 51 9 4 2]
[15 7 25 49 31 5]
[7 6 12 13 34 16]
[5 14 8 5 13 34]]



Confidence Interval: (0.3,0.4)

5. Biases observed:

In confusion matrices, we can see that accuracy is pretty good but there are few frequent biases. There are more number of songs which are misconstrued as pop. The reason being pop music follows the pattern which are more likely to inherit from other categories. Pop follows the popular music style which shares commonality in other categories. And also we can observe from the figure that there are more folk, pop and electronic that is being misconstrued as instrumental. The reason might be that folk and pop songs share more common properties with that of instrumental. Hence its harder to discriminate them. Therefore, we can make out that pop and instrumental are comparably harder to distinguish. The reason being they have less innate quality that make them stand out.

6. Way to improve classification:

This task can be better accomplished by using a ML algorithm called Automatic Music genre classification. This approach uses multiple feature vectors and a pattern recognition approach, according to space and time decomposition schemes. Despite being music genre classification a multi-class problem, accomplish the task using a set of binary classifiers, whose results are merged in order to produce the final music genre label (space decomposition). Music segments are also decomposed according to time segments obtained from the beginning, middle and end parts of the original music signal (time-decomposition). The final classification is obtained from the set of individual results, according to a combination procedure.[c]

References:

1. https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-65002008000300006
2. https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6
3. https://medium.com/@sukantkhurana/music-classification-using-artificial-intelligence-3d21c59c5cb2
4. https://machinelearningmastery.com/confidence-intervals-for-machine-learning/
5. https://machinelearningmastery.com/k-fold-cross-validation/