UNIVERSIDADE DE SÃO PAULO

FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO

DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

GUSTAVO FERNANDES CARNEIRO DE CASTRO,

gus.castro@usp.br

11369684

TIAGO COSTA CARVALHO,

tiagococarv@usp.br

11315102

Redes neurais artificiais em predição de gênero musical.

Ribeirão Preto

2021

Introdução

Pretende-se utilizar redes neurais artificiais, em específico perceptrons multi-camadas (MLP), para predizer o gênero musical de um conjunto arbitrário predeterminado de gêneros e músicas. Para tal, foram coletadas as médias dos **n** coeficientes *MFC* (mel-frequency cepstrum) gerados a cada tempo, utilizando a biblioteca *librosa*, e analisados utilizando o MLP pela biblioteca *scikit learn*, com auxílio da biblioteca *pandas* e seus DataFrames.

Metodologia

Para montar o DataSet, primeiro foram selecionados arbitrariamente 10 gêneros musicais, sendo eles Clássico (C), Eletrônica (E), Heavy Metal (HM), Jazz (J), Pop (P), Rock (R), Rhythm & Blues/Soul (R&B/ReB), Rap e derivados (RAP), Retro Gaming (RG) e Samba (S). Para cada gênero musical, foi selecionado arbitrariamente um conjunto de aproximadamente 22 músicas, totalizando 205 músicas.

Os coeficientes utilizados foram coletados através de um método de extração de características advindo da biblioteca *librosa*. Esta biblioteca permite carregar arquivos .mp3, possibilitando a utilização de métodos já existentes próprios para se manipular faixas de áudio, coletar características, editar e coletar informações, entre outros tipos de manipulação referentes a áudio.

Utilizando o método *librosa.feature.mfcc()*, é possível extrair os *Mel-frequency cepstral coefficients* (MFCCs) de uma faixa de áudio. Estes coeficientes representam, de uma forma simples, o espectro de força de uma onda sonora em um determinado instante de tempo. Com isso, para cada coeficiente **i**, **i** = {1,...,20}, foi calculado o valor deste em todos os instantes **t** de cada música. Depois, foi calculada a média do coeficiente **i** em todos os instantes **t**, finalizando com um total de **n** = 20 atributos para cada música. Finalmente, com os cálculos finalizados, um DataSet contendo 205 músicas, com 20 atributos e uma classe cada.

Inicializando a análise, primeiro, foi importado o DataSet e separado em conjunto de dados (**X)** e classes (**Y)**. Depois, utilizando do *scikit learn*, o **X** foi normalizado e **Y** foi transformado em numérico {C: 0, E: 1, HM: 2, J: 3, P: 4, R: 5, ReB: 6, RAP: 7, RG: 8, S: 9}.

Depois foi escolhido um conjunto de parâmetros que poderiam alterar os resultados da rede neural. Os parâmetros escolhidos foram a função de ativação: ‘identity’ - função linear, ‘logistic’ - função logística sigmoidal, ‘tanh’ - função da tangente hiperbólica e ‘relu’ - ativação linear retificada; a quantidade de camadas ocultas e neurônios; a taxa de aprendizado inicial; o algoritmo de solução para otimizar os pesos: lbfgs’, ‘sgd’, ‘adam’; o número máximo de iterações; se terá aleatorização das amostras durante a resolução do algoritmo; e o momentum.

Estes parâmetros foram escolhidos com base nos conhecimentos adquiridos na disciplina e, alguns, arbitrariamente. Para decidir os valores desses parâmetros, foi utilizado o método *GridSearchCV()*, da biblioteca *scikit learn,* junto do *param_grid* onde os melhores valores eram escolhidos dentre as opções inseridas no *param_grid.* Após definir os melhores parâmetros, foi aplicado um cross validation com 5 *folds*.

Para realizar a análise dos dados obtidos, foi escolhida a matriz de confusão. Para tal, os resultados preditos foram inseridos em um DataFrame e, através do método *confusion_matrix()* do *scikit learn*, obteve-se a matriz de confusão para o melhor caso escolhido no passo anterior.. A análise foi realizada diversas vezes, e os resultados mais interessantes foram inseridos abaixo.

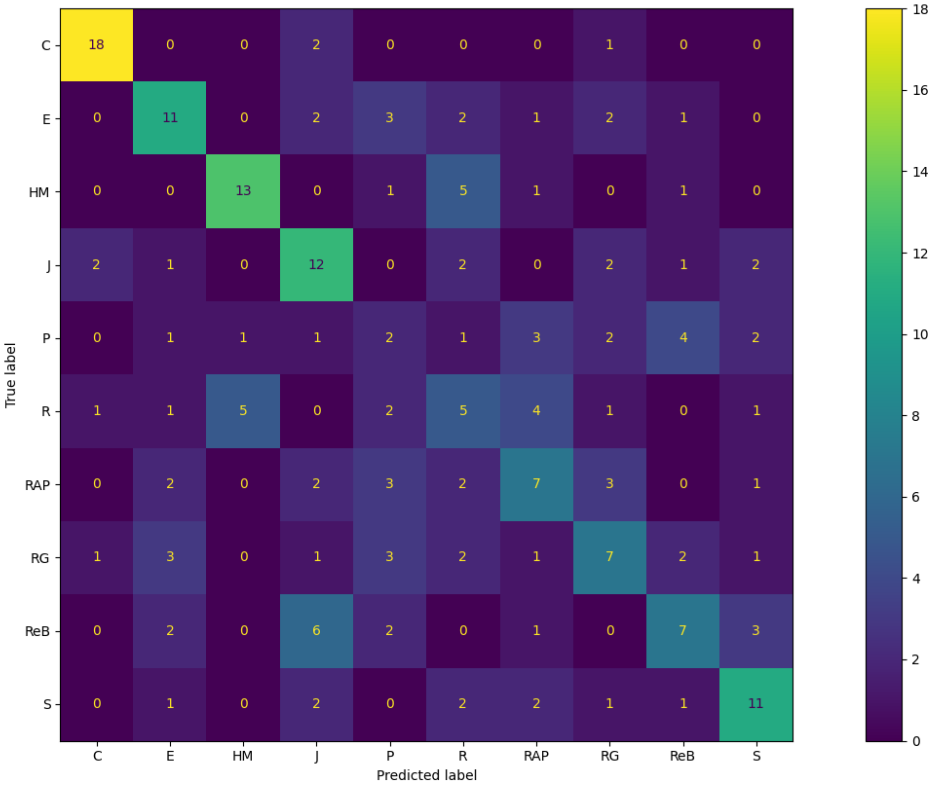
Resultados

Dentre os resultados obtidos, a melhor taxa de acerto de 45,3% foi obtida pelas redes com os seguintes parâmetros:

Rede 1:

- função de ativação: identidade
- duas camadas ocultas com 200 perceptrons cada
- taxa inicial de aprendizado: 0,01
- algoritmo de solução: *adam*
- número máximo de iterações: 200
- aleatorização da ordem das entradas
- *momentum*: 0.9

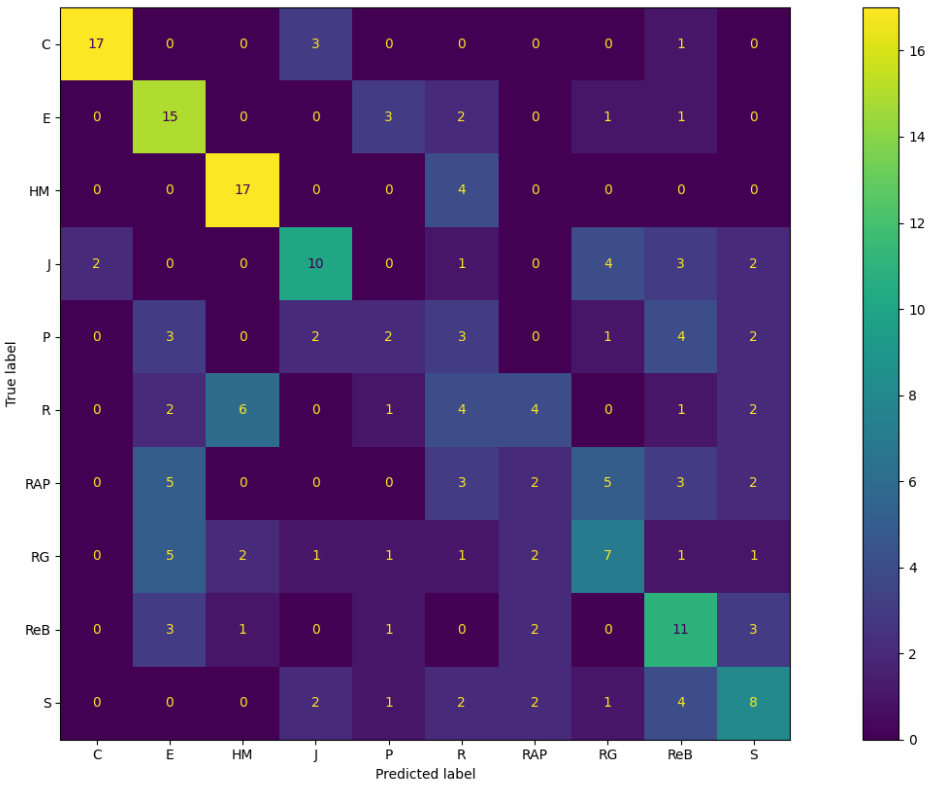
Com a seguinte matriz de confusão:



Rede 2:

- função de ativação: função hiperbólica tangencial
- uma camada ocultas com 100 perceptrons cada
- taxa inicial de aprendizado: 0,05
- algoritmo de solução: *sgd*
- número máximo de iterações: 200
- aleatorização da ordem das entradas
- *momentum*: 0.9

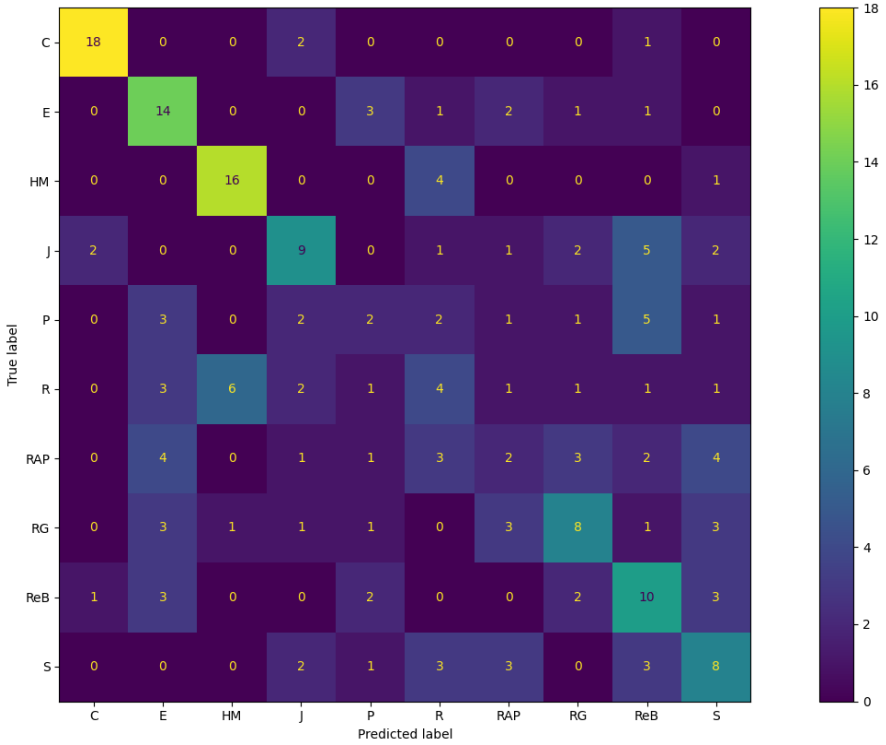
Com a seguinte matriz de confusão:



Outro teste notável, com taxa de acerto de 44,3%, a seguinte rede:

- função de ativação: função hiperbólica tangencial
- uma camada oculta com 9 perceptrons
- taxa inicial de aprendizado: 0,1
- algoritmo de solução: descida gradiente estocástica
- número máximo de iterações: 200
- *momentum*: 0.9

Com a seguinte matriz de confusão:



Vale ressaltar também que a taxa de acerto média quando a base possuía aproximadamente 60 entradas era de aproximadamente 15%.

Conclusão

Durante a etapa de pesquisa, foram ponderados diversos *features* que poderiam ser utilizados para a análise dos gêneros musicais. Dentre eles a frequência. Porém, foi escolhido o MFCC pois este já é muito utilizado na área de reconhecimento de fala, e poderia dar resultados mais satisfatórios para esta pesquisa.

Durante a etapa de desenvolvimento do código, foi possível observar algumas interações e ocorrências interessantes. Dentre elas, durante a fase de teste do código, foi utilizada uma base de dados com menos de 70 músicas, e com uma péssima distribuição. Os resultados obtidos de acerto beiravam os 15%, enquanto esse número aumentou para

aproximadamente 30%, apenas com um aumento considerável na quantidade de músicas. Isso mostra que um dos pontos mais importantes em uma análise desse tipo é a quantidade e qualidade dos dados obtidos, além de uma boa distribuição dos tipos e classes.

Uma das observações intrigantes realizadas está relacionada ao número de neurônios na camada oculta. Durante os testes, a maioria dos melhores resultados possuíam um número alto de neurônios nas camadas ocultas. Porém, ao utilizar a heurística de escolha de neurônios: aproximadamente um para cada classe, o resultado foi surpreendentemente positivo. Com uma taxa de 44,3% de acerto, em comparação a 45,3%, mostrando que a divisão em 9 hiperplanos é o bastante para separar as 10 classes de forma distinta.

Uma observação interessante sobre as músicas analisadas é o agrupamento consistente de alguns tipos de música consideradas "próximas". Rock com Heavy Metal, sendo este um subgênero daquele; Clássico com Jazz, com ambos utilizando de instrumentos de sopro e piano frequentemente, e muitos sendo apenas instrumentais, mesmo sendo fundamentalmente diferentes; e Eletrônica, Pop e R&B, gêneros "primos", mais recentes, com muita intersecção entre eles e diversas músicas em seus intermediários. Isso mostra que é possível, com uma polida e afinação dos algoritmos e uma melhora no Dataset, aumentar a taxa de acerto do algoritmo para número elevados.

Outro ponto interessante são os gêneros que o algoritmo teve mais dificuldade de classificar. Estes sendo principalmente o Rock e o Pop, frequentemente o Retrô Gaming e o Rap, e menos frequente o Samba e o R&B. O Pop é conhecido por ter um apanhado gigantesco de estilos e não ter uma característica marcante. O mesmo pode ser dito do R&B, e do Retrô Gaming, que engloba diversos gêneros utilizados em diversos jogos diferentes, com a diferença de a maioria das músicas escolhidas estarem comprimidas para rodarem e dispositivos mais antigos, facilitando seu discernimento de outros gêneros. O Rock, por sua vez, junto do Rap e do Samba, englobam diversos subgêneros e estilos únicos para cada banda ou artista, dificultando a classificação deles como algo único. O Rap, em específico, teve músicas em inglês e português, fazendo com que, além das diferenças marcantes dentro do próprio gênero musical, também haja diferenças nos estilos culturais do Brasil e dos EUA.

Em conclusão, o pontos que melhor garantem uma taxa de acerto alta em um classificador do tipo MLP é a qualidade do Dataset, levando em conta quantidade, distribuição e escolha da amostra; e a qualidade do *feature* selecionado para análise, sendo

que este deve ser impactante para a classificação de um dado. Caso estes dois requisitos não sejam atendidos, o MLP não conseguirá construir bons hiperplanos e classificar corretamente os dados obtidos. Porém, mesmo com poucos dados, é possível observar alguns padrões e direcionar a análise corretamente, ajustando-a com o tempo e com a melhora nos dados e *features*. Além disso, foi possível observar a dificuldade que é definir e ajustar os valores dos atributos utilizados e que qualquer alteração nestes pode acarretar em uma melhora ou piora significativa no resultado final.

Referências

https://www.kaggle.com/ashishpatel26/feature-extraction-from-audio

https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

https://scikit-learn.org/