UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO
Centro de Ciências e Tecnologia
Faculdade de Engenharia
Departamento de Engenharia Eletrônica e de Telecomunicações

PROJETO DE GRADUAÇÃO VIII-A

Classificação de instrumentos musicais baseada em aprendizado de máquina

Gabriela Siqueira Eduardo

Professor Orientador: Prof. Michel Pompeu Tcheou
Professor da Disciplina: Prof. Alexandre de Oliveira Dal Forno

Maio/2022

Sumário

1. Introdução

Desde a pré-história, a música é um elemento fundamental da cultura humana. Ao decorrer dos anos, houve um grande aumento na disponibilidade de músicas digitais, bem como a facilidade de acesso a elas, o que faz crescer o número de ouvintes.

Com o avanço da tecnologia, também houve uma evolução em como os sinais de áudio podem ser gerados, acessados e consumidos.

Os grandes responsáveis pela facilidade de distribuição e de consumo de músicas da atualidade são as plataformas de *streaming*, como, por exemplo, *Spotify*, *Apple Music*, *Deezer*, entre muitas outras opções. Essas plataformas podem se utilizar de sistemas de classificação para a categorização do catálogo de músicas, bem como de sistemas de recomendação, a fim de aprimorar a experiência dos usuários, sugerindo estilos semelhantes para cada ouvinte.

1.1. Objetivos

No estudo a ser desenvolvido, pretende-se classificar instrumentos musicais presentes em composições de estilos variados, utilizando-se algoritmos computacionais de aprendizado de máquina supervisionado.

Para isso, serão estudadas e utilizadas algumas características espectrais de um sinal de áudio - como largura de banda de frequência, centróide espectral, coeficientes cepstrais de frequência-Mel - extraídas através de algoritmos próprios.

Além disso, buscar-se-á explorar o universo dos algoritmos de aprendizado de máquina escolhidos - *Support Vector Machine* e Redes Neurais -, abordando o seu funcionamento e os seus parâmetros.

1.2. Motivação

O que nos motivou a utilizar, como objeto de análise, algoritmos computacionais de aprendizado de máquina supervisionado na classificação de

instrumentos musicais presentes em composições de estilos variados foi o fato de existir uma grande dificuldade no reconhecimento de cada um dos múltiplos instrumentos em uma canção, devido à sobreposição de tempo e de frequência, à variação de timbres e à falta de dados classificados. Além disso, deve-se levar em consideração que, na realidade, as componentes espectrais de um mesmo instrumento não são constantes, mesmo que se esteja estudando uma mesma nota - o que eleva o grau de dificuldade no seu reconhecimento.

Finalmente, a classificação de áudio de instrumentos, de gêneros, de notas, entre outros, faz-se interessante na automatização de consultas de peças musicais, de criação de catálogo, de transcrição de músicas, bem como na criação de sistemas de recomendação.

2. Revisão preliminar da bibliografia

O estudo de recuperação de informações musicais (MIR) constitui um campo relativamente novo, quando comparado ao estudo de outros tipos de recuperação de informações, como o processamento de fala (*speech proccessing*).

O MIR trata-se de um campo de pesquisa utilizado para categorizar, manipular e até criar música. Suas aplicações são: classificação de gênero musical, dentre gêneros pré-determinados, como *rock* e *jazz,* extraindo-se a sua melodia; classificação de instrumentos (objeto de estudo deste projeto); classificação de artistas; criação de sistemas de recomendação (utilizando técnicas de similaridades entre músicas); transcrição automática de músicas (tanto em partituras, quanto em arquivos MIDI); e geração de música, embora com efetividade limitada.

No ano de 2000, em Massachusetts, nos Estados Unidos, realizou-se o primeiro simpósio de recuperação de informações musicais. O assunto é tão relevante que, em 2008, foi fundada uma sociedade internacional (ISMIR), responsável por realizar anualmente conferências com o objetivo de apresentar estudos sobre análise, processamento e recuperação de informações de sinais de áudio.

Bosch, Janer, Fuhrmann e Herrera (2012) estudaram como a segregação de um sinal de áudio multi-timbral pode afetar o reconhecimento dos instrumentos predominantes nele. Para isso, foi utilizada uma base de dados composta por trechos curtos de músicas ocidentais gravadas profissionalmente. Escolheu-se o estilo ocidental devido à alta relação harmônica entre os instrumentos nesse tipo de música, o que pode resultar no compartilhamento de frequências em suas componentes espectrais.

Esse estudo foi dividido em cinco experimentos. O primeiro foi o mais simples, no qual não foi realizado nenhum tipo de separação de sinal, apenas o sinal de áudio monofônico e aplicado um modelo de classificação único para a base como um todo.

Já no segundo, o áudio foi separado de acordo com os seus canais estéreo, sendo eles Esquerdo/Direito/Meio/Lado (*Left/Right/Right-Mid/Side*, LRMS), sendo Meio = Esquerdo + Direito e Lado = Esquerdo - Direito. Para tais divisões, foi utilizado um único modelo classificador.

No terceiro experimento, utilizou-se a estrutura flexível de separação de fonte de áudio (FASST), em que se dividiu o sinal em quatro partes: “bateria”, “baixo”, “melodia” e “outros”. Essa separação considera que idealmente a parte da melodia seja composta pelos instrumentos predominantes que devem ser reconhecidos. Assim como no primeiro e no segundo, foi utilizado um mesmo classificador para essas partes.

No quarto experimento, o FASST foi aplicado da mesma forma que no terceiro, porém empregou-se um classificador diferente para cada saída dele. Nele, percebeu-se que instrumentos com frequências mais baixas são mais facilmente reconhecidos.

Finalmente, o quinto experimento consistiu de uma tentativa de aprimoramento do quarto, utilizando-se técnica de grau de sobreposição na combinação de instrumentos.

A realização desses experimentos mostrou que o desempenho da classificação não é proporcional à complexidade do experimento, visto que aquele

menos complexo (o primeiro) obteve a melhor precisão. O mais complexo, além de ser o que exige mais capacidade computacional, teve como melhor métrica, dentre todos os outros experimentos, a sensibilidade e o micro-F1. Portanto, para decidir qual o melhor experimento, é necessário levar-se em conta os objetivos e as condições físicas para a obtenção dessas metas.

Racharla, Kumar, Jayant, Khairkar e Harish (2020) utilizaram componentes espectrais do sinal de áudio para a classificação dos instrumentos, já que, segundo os autores, as mesmas notas musicais geram espectrogramas diferentes, quando considerados instrumentos distintos.

Dessa forma, os estudiosos entenderam que os coeficientes cepstrais de frequência-Mel, que representam a transformada discreta de cosseno (DCT) do espectro logarítmico de um sinal, analisado na escala Mel, que é mais próxima do sistema auditivo humano, seriam fundamentais para o objetivo, já que esses números representam a forma que um tom é descrito por uma frequência. Também foram necessárias outras componentes espectrais, como a largura de banda, centróide, entre outras, para uma classificação mais acurada.

Sobre o modelo de classificação, foram testados modelos de aprendizado de máquina supervisionados e não supervisionados, obtendo-se um resultado melhor para o modelo supervisionado, visto que o modelo não supervisionado não conseguiu agrupar os instrumentos de forma apropriada.

Diferentemente dos estudos apresentados anteriormente, nos quais os sinais estudados constavam de gravações profissionais com três segundos de duração e que possuíam mais de um instrumento tocando simultaneamente, Sell, Mysore e Chon (2006) criaram sinais de música polifônica de forma artificial.

A criação dos sinais de áudio foi feita a partir da gravação de performances solo de cinco instrumentos diferentes, sendo que cada amostra consistia de apenas um instrumento. Em seguida, foram realizadas combinações de instrumentos, sobrepondo-se os múltiplos sinais de cada. Dessa forma, a base de dados criada é constituída de áudios com dois, três e cinco instrumentos distintos em cada amostra.

Em seguida, foram selecionados três tipos de preditores diferentes, escolhidos através de um conhecimento prévio dos instrumentos disponíveis, sendo eles a magnitude da transformada discreta de Fourier (DFT), os coeficientes cepstrais de frequência-Mel (MFCCs) e a mudança de energia entre os quadros das amostras.

A magnitude DFT foi escolhida com o desejo de conseguir explorar algumas características de frequência que podem diferenciar os instrumentos presentes, como os picos de frequência do tom de uma nota, a largura de banda e o alcance (variação) da frequência.

Os MFCCs foram selecionados com o objetivo de representar os espectros diferentes de cada sinal, além do que já foi apresentado anteriormente.

A mudança de energia entre quadros foi incluída devido à forma que um som se projeta ao longo do tempo. Essa forma consiste no ataque, decaimento, sustentação e liberação (ADSR) de, no caso do estudo, cada instrumento.

Após a extração dos preditores, foram utilizados três modelos de classificação supervisionados para avaliar a presença de cada instrumento. Com esses resultados,provou-se que, quanto mais instrumentos em uma música, mais difícil é diferenciá-los, visto que as classificações foram mais corretas nas amostras com dois instrumentos diferentes.

3. Roteiro preliminar

Para a obtenção do objetivo, o projeto pode ser estruturado da seguinte forma.

3.1. Base de dados

Essa base consiste de 9579 amostras de áudio, de 3 segundos de duração, do tipo .*wav* (*Waveform Audio File Format*), com 2 canais de reprodução e são separados de acordo com o instrumento predominante ali presente. Como um todo, a base apresenta 11 classificações diferentes de instrumentos, sendo eles:

violoncelo, clarinete, flauta, violão acústico, guitarra elétrica, órgão, piano, saxofone, trompete, violino e voz humana.

A base será dividida entre treino e teste, sendo 70% para treino (6705 arquivos) e 30% para teste (2874 arquivos). Todas as amostras são acompanhadas de um arquivo de texto, que, na base de teste, pode apresentar mais de um instrumento identificado, enquanto na base de treino apresenta apenas um, o predominante.

3.2. Pré-processamento dos dados

Para tratamento da base e da extração das *features* (preditores), a serem aplicadas no modelo de classificação, utilizar-se-á a biblioteca *Librosa* do *python*.

Primeiramente o sinal de áudio com dois canais de reprodução (estéreo) será reduzido para apenas um (monofônico) e, em seguida, serão obtidas informações do sinal, como o valor quadrático médio (RMS), o valor do centróide espectral (SC), a largura de banda espectral (SB), a frequência de *roll-off* (SR), a taxa de cruzamento do zero (*zero-crossing rate*, ZCR) e os coeficientes cepstrais de frequência-Mel (MFCCs).

3.3. Modelo de classificação

Com os dados já processados e prontos para entrar no modelo de classificação, serão desenvolvidos dois modelos de classificação de aprendizado de máquina supervisionado - o *Support Vector Machine* (SVM), da biblioteca do *Python Scikit Learn,* e as Redes Neurais Artificiais, da biblioteca *Keras*.

4. Custo do projeto[1]

Recursos materiais

Notebook	US$ 1476,00
Acessos à Internet	US$ 590,00
Custo de material total	US$2.066,00

Recursos humanos

Custo homem/hora	US$ 7,60
Carga horária diária	6 horas
Total de semanas	40 semanas
Nº de dias da semana	7 dias
Custo de RH total	US$12.768,00

Custo total

Custo de material	US$ 2.066,00
Custo de RH	US$ 12.768,00
Custo total	US$ 14.834,00

[1] Valores adaptados para dólar, cotado a R$5,08.

5. Cronograma de atividades

	Mês/Ano							
Especificações	02/22	03/22	04/22	05/22	06/22	07/22	08/22	09/22
Preparação do projeto com definição de tema e orientador								
Revisão bibliográfica inicial								
Revisão bibliográfica								
Criação do código do projeto								
Pesquisa sobre processamento de áudio								
Pesquisa sobre aprendizado de máquina								
Apresentação ao orientador								
Discussão da redação preliminar								
Redação final								
Entrega de cópias para a banca								
Defesa								
Redação definitiva								

6. Referências bibliográficas

BOSH, J. *et alii.* "*A Comparison of sound segregation techniques for predominant instrument recognition in musical audio signals"*. *In:* ISMIR: 13th International Society for Music Information Retrieval Conference. Porto, 2012. (pp. 559-564)

DESHPANDE, H., NAM, U. & SINGH, R. "C*lassification of music signals in the visual domain*". *In: Proceedings of the COST G-6 Conference on Digital Audio Effects .* Limerick, Irlanda, 2001.

GURURANI, S., SHARMA, M. &. LERCH, A. "*An attention mechanism for musical Instrument recognition*". *In:* ISMIR: Proceedings of the 20th International Society for Music Information Retrieval Conference. Delft, Países Baixos: 2019. (pp. 83–90)

ISMIR. ISMIR: *International Society for Music Information Retrieva*l, 2022. Página inicial. Disponível em: <https://www.ismir.net/>

LIVSHIN, A. & RODET, X. "Musical instrument identification in continuous recordings". *In*: 7*th International Conference on Digital Audio Effects* (DAFX-4), Nápoles, Itália, 2004.

MAZARAKIS, G, TZEVELEKOS, P. & KOUROUPETROGLOU, G. "*Musical Instrument recognition and classification using time encoded signal processing and fast artificial neural networks*". *In*: ANTONIOU, G. *et alii* (org.). Advances in *artificial intelligence. Lecture notes in computer science,* vol 3955. Berlim, Heidelberg: Springer, 2006. https://doi.org/10.1007/11752912_26

MÜLLER, M. *Fundamentals of music processing: audio, analysis, algorithms*, applications. Alemanha: Springer, 2001.

RACHARLA, K., *et alii.* "*Predominant musical instrument classification based on pectral features*" . *In: 7th International Conference on Signal Processing and Integrated Networks (SPIN),* 2020. (pp. 617-622)

SELL, G. *et alii*. "*Musical instrument detection: detecting instrumentation in polyphonic musical signals on a frame-by-frame basis*". In: *Center for computer research in music and acoustics.* Stanford, 2006.