Arquitetura para a classificação automática de gêneros musicais utilizando análise de imagens e redes complexas

Architecture for automatic music genre classification using image analysis and complex networks

Rafael Marasca Martins[1], Andrés Eduardo Coca Salazar[2]

RESUMO

Hodiernamente a música é amplamente consumida por meios digitais, como plataformas de *streaming*. Nesse contexto, a classificação de gêneros musicais tem papel vital, pois permite a criação de sistemas de recomendação de música e a organização de grandes bancos de dados musicais. Entretanto, tais tarefas se mostram desafiadoras e tediosas devido ao alto grau de complexidade e repetição. Desta forma, convém o desenvolvimento de sistemas computacionais capazes de realizar a classificação de músicas de acordo com o gênero de forma automática, acurada e eficiente. Diante do exposto, o presente artigo propõe uma arquitetura para a classificação de gêneros musicais usando espectrogramas, texturas de imagens e redes complexas. Ainda, foram desenvolvidos programas para a geração de espectrogramas a partir de arquivos de áudio, bem como das características texturais das imagens obtidas, criação de redes complexas e sua respectiva mineração com medidas topológicas. Além disso, um classificador baseado em redes neurais foi implementado para avaliar a desempenho do sistema, para o qual os resultados demonstraram performance razoável no conjunto de teste.
PALAVRAS-CHAVE: Classificação de Dados; Processamento digital de Sinais; Redes complexas.

ABSTRACT

Nowadays, music is broadly consumed by digital means such as streaming platforms. In this context, music genre classification is vital since it enables the development of music recommendation systems and the organization of large musical databases. However, such tasks can be challenging and tedious due to the high complexity degree and repetition involved. Therefore, it is appropriate to develop computational systems capable of automatic music classification by genre efficiently and accurately. In the face of that, this research proposes an architecture for music genre classification using spectrograms, image textures, and complex networks. Furthermore, scripts to generate spectrograms from audio files, capture textural features of the resultant images, create complex networks, and perform the respective mining with topological measurements have been developed. For evaluating the system's performance, a neural network classifier is applied, for which the results showed reasonable accuracy in the test split.
KEYWORDS: Data classification; Digital signal processing; Complex networks.

INTRODUÇÃO

Os gêneros musicais constituem uma forma sistemática de agrupar músicas de acordo com características em comum, tais como instrumentos utilizados, melodia e estrutura rítmica (TZANETAKIS; ESSL et al., 2001). Entretanto, a classificação manual de obras musicais baseadas no gênero é

[1] Discente no Curso de Engenharia de Computação. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: rafael.030601@alunos.utfpr.edu.br. ID Lattes: 5355687369791576.
[2] Docente no Curso de Engenharia de Computação.Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: andressalazar@utfpr.edu.br. ID Lattes: 9651142799957514.

uma tarefa árdua e, muitas vezes, subjetiva, demandando tempo e um amplo conhecimento técnico na área de teoria musical. Neste sentido, o aprimoramento dos sistemas atuais de classificação automática de músicas baseado no gênero faz-se proveitoso e oportuno. Portanto, o presente trabalho propõe uma arquitetura para um sistema de classificação automática de gêneros musicais baseado na análise de imagens e técnicas de representação e mineração com redes complexas. Este documento está dividido da seguinte forma: primeiramente é introduzido o estado d'arte, após é apresentado o referencial teórico, então é relatada a metodologia utilizada e, por fim, são expostos os resultados, conclusões e trabalhos futuros.

ESTADO D'ARTE

Esforços vêm sendo empregados no desenvolvimento de sistemas de classificação de gêneros musicais, sendo extensas as abordagens que empregam representações visuais do espectro como base. O trabalho de Costa et al. (2011) utiliza espectrogramas, bem como atributos extraídos com matriz de co-ocorrência de cinza (GLCM - *Grey-Level Co-occurrence Matrix*) para realizar a classificação de gêneros musicais, obtendo uma acurácia máxima de 65,7%. Já, Coca (2022a), recorre a padrões locais binários completos (CLBP - *Completed Local Binary Patterns*) para gerar redes complexas tripartites a partir de mel-espectrogramas e gamatonegramas, atingindo acurácia de 53,10%. Por sua vez, Coca (2023b) faz uso de descritores de informação estrutural e níveis de cinza (GLSI - *Gray Level and Structural Information*) para produzir redes complexas, obtendo 56,28% de acurácia. Ainda, destaca-se o trabalho de Coca (2022), que emprega uma estratégia de mineração híbrida e hierárquica de redes complexas, onde são utilizadas como entradas ao classificador medidas texturais obtidas a partir do GLCM, extraídas diretamente da rede complexa que foi gerada com mel-espectrogramas, alcançando uma acurácia de 98,3%.

REFERENCIAL TEÓRICO

- As **redes complexas** modelam relações e interações entre dados por meio de um grafo. Desta forma, através das medidas topológicas da rede é possível extrair informações importantes que evidenciam essas relações, as quais podem ser usadas para fazer inferências. As seguinte três medidas topológicas são extensamente utilizadas na literatura (SILVA; ZHAO, 2016):

  Grau médio, $\overline{k}$, representa a quantidade média de conexões dos vértices. Esta medida opera sobre o conjunto de vertices da rede, γ, e sobre ε, que representa o conjunto de arestas do grafo, assim:

$$\overline{k} = \frac{1}{|\gamma|} \sum_{(v,u)\in\gamma^2} 1 \text{ se } (u,v) \in \varepsilon \ . \tag{1}$$

  Coeficiente de agrupamento, C, indica o quanto os vértices tendem a se agrupar. Este parâmetro depende de $|e_i|$, que é o número de arestas compartilhadas entre os vizinhos do vértice *i*, e de k_i, que corresponde o grau do vértice *i*, desta forma:

$$C = \frac{1}{|\gamma|} \sum_{i \in \gamma} \frac{2|e_i|}{k_i(k_i - 1)} \quad . \tag{2}$$

Sortividade, *r*, representa a correlação de graus entre vértices vizinhos e baseia-se em u_e e v_e, que são os graus dos vértices ligados pela aresta e, e em $E = |e|$, que representa o número de arestas da rede, logo:

$$r = \frac{E^{-1} \sum_{e \in \varepsilon} (u_e v_e) - [\frac{E^{-1}}{2} \sum_{e \in \varepsilon} (u_e + v_e)]^2}{\frac{E^{-1}}{2} \sum_{e \in \varepsilon} (u_e^2 + v_e^2) - [\frac{E^{-1}}{2} \sum_{e \in \varepsilon} (u_e + v_e)]^2} \quad . \tag{3}$$

- **Cromagrama** é um meio de representar visualmente a informação espectral agregada de uma música ao longo do tempo (MÜLLER, 2015). Basicamente, o cromagrama consiste em uma imagem, cuja dimensão vertical é dividida em 12 intervalos, correspondentes aos 12 semitons da escala cromática, enquanto a dimensão horizontal corresponde a intervalos discretos de tempo. Assim, para cada intervalo de tempo é computada a energia das componentes de frequência do sinal correspondentes a cada uma das notas.

- **Escalograma** é uma representação do sinal em termos de frequência e tempo, computada através da transformada *wavelet*. Em comparação com os espectrogramas concebidos por técnicas como a transformada de Fourier, o escalograma é capaz de fornecer resolução variável de frequência e tempo, assim sinais mais lentos podem ser analisados com menor resolução temporal e maior resolução espectral, enquanto que sinais mais rápidos podem ser analisados com maior resolução temporal e menor resolução espectral (TZANETAKIS; ESSL et al., 2001).

- **Texturas** descrevem padrões de distribuição espacial de pixels em imagens (HARALICK et al., 1973). Desta maneira, as texturas podem ser utilizadas como fonte de atributos para tarefas de reconhecimento de padrões em imagens (MARTINS et al., 2011). Diante disso, o GLCM é uma ferramenta amplamente empregada na literatura, uma vez que, através dela, pode-se calcular, diretamente, medidas que caracterizam texturas (HARALICK et al., 1973).

MATERIAIS E MÉTODOS

Para a obtenção dos sinais de áudio, foi utilizado o banco de dados GTZAN (TZANETAKIS; COOK, 2002), que é extensivamente explorado na literatura para o desenvolvimento dos sistemas de classificação de gênero musical.

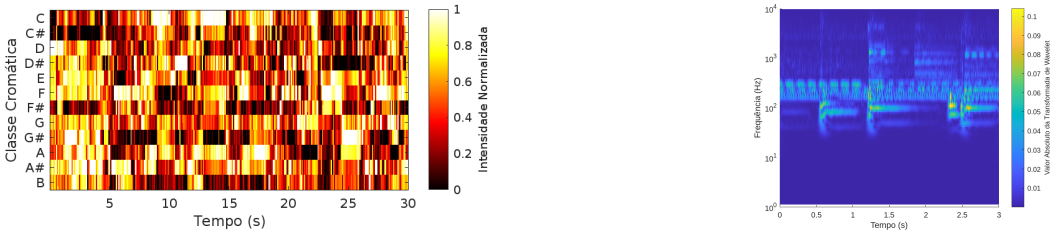
Para um trecho do sinal acústico de entrada foram obtidas duas representações visuais: escalograma e cromagrama, esses espectrogramas foram convertidos para escala de cinza e divididos em trechos de 3 segundos. Utilizando o método de GLCM, com 4 pixels de deslocamento nas direções vertical, horizontal e nas diagonais, obteve-se como saída 16 matrizes (uma para cada deslocamento), a partir das quais, individualmente, foram extraídos os descritores de textura: contraste, correlação, energia e homogeneidade (HARALICK et al., 1973). Desta forma, cada imagem gerou 16 vetores de dimensão 4, que correspondem aos vértices das redes complexas. As arestas foram construídas a partir da conexão de cada vértice com os vértices seus vizinhos mais similares, considerando a

distância Euclidiana (assim, as conexões servem como uma forma de representar a semelhança de duas regiões da imagem na rede complexa). Estas redes foram mineradas a partir das medidas topológicas antes descritas, as quais compõem o vetor de entrada de um algoritmo de classificação multiclasse, que, por sua vez, gerou como saída uma estimativa do gênero musical.

RESULTADOS

A Fig. 1 apresenta os espectrogramas gerados para o arquivo de áudio #100 do banco de dados GTZAN. Na Fig. 1(a) é exposto o cromagrama para o arquivo, enquanto que na Fig. 1(b) verifica-se o escalograma para um trecho contendo os primeiros 3 segundos do áudio.

Figura 1 – Espectrogramas gerados para o arquivo de áudio #100 do banco de dados GTZAN



(a) Cromagrama **(b) Escalograma**

Fonte: Elaborado pelos autores (2022)

A Fig. 2 apresenta as redes complexas geradas para os primeiros 3 segundos do arquivo #100 do banco de dados GTZAN utilizando 5 vizinhos mais próximos, em que, a Fig. 2(a) apresenta a rede complexa derivada do cromagrama e a Fig. 2(b) expõe a rede complexa derivada do escalograma.

Figura 2 – Redes complexas gerada para os primeiros 3 segundos do arquivo #100 do bando de dados GTZAN

(a) Rede gerada a partir do cromagrama **(b) Rede gerada a partir do escalograma**

Fonte: Elaborado pelos autores (2023)

As medidas topológicas extraídas para as redes complexas apresentadas na Fig.2 são apresentadas na Tabela 1.

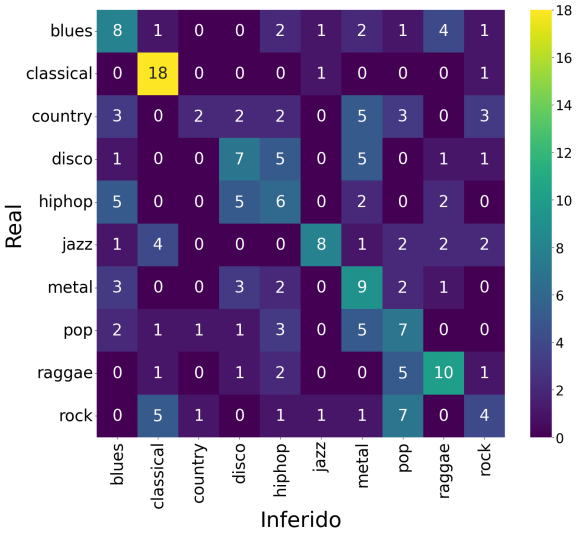
Tabela 1 – Medidas Topológicas extraídas das redes da Fig. 2

Medidas	Rede do cromagrama (Fig. 2(a))	Rede do escaograma (Fig. 2(b))
Coeficiente de agrupamento	0,80	0,82
Grau médio dos vértices	6,00	6,50
Sortividade	-0,14	-0,05

Fonte: Elaborados pelos autores (2023)

Foi desenvolvido um classificador utilizando uma rede neural composta de uma camada de entrada, 3 camadas ocultas de 32, 64 e 32 neurônios, respectivamente, uma camada de rejeição e uma camada de saída de tamanho 10. O conjunto de dados foi dividido em 20% para teste e 80% para treino. Os resultados de classificação são apresentados na Tabela 2, enquanto que a matriz de confusão gerada é apresentada na Fig. 3.

Figura 3 – Matriz de Confusão



Fonte: Elaborado pelos autores (2023)

Tabela 2 – Medidas de Desempenho

Classe	Precisão	Sensibilidade	F_1-*score*
Blues	0,35	0,40	0,37
Clássica	0,60	0,90	0,72
Country	0,50	0,10	0,17
Disco	0,37	0,35	0,36
Hiphop	0,26	0,30	0,28
Jazz	0,73	0,40	0,52
Metal	0,30	0,45	0,36
Pop	0,26	0,35	0,30
Raggae	0,50	0,50	0,50
Rock	0,31	0,20	0,24
Média	0,42	0,40	0,38

Fonte: Elaborado pelos autores (2023)

Diante disso, verifica-se uma precisão de 42%, sendo a música Clássica a classe com melhor performance no sistema de classificação, com precisão de 60%, sensibilidade de 90% e F_1-*score* de 72%. Já, Hiphop e Pop, foram os gêneros com pior precisão. Ainda, observa-se que a classe Country, apesar de possuir uma precisão de 50% tem sensibilidade e F_1-*score* relativamente baixos, indicando baixa confiabilidade.

CONCLUSÃO

Neste trabalho foi apresentada uma arquitetura, para sistemas de classificação de gêneros musicais, que consiste nas etapas de geração de cromagramas e escalogramas, extração de descritores texturais obtidos a partir de GLCM, geração de redes complexas e extração de medidas topológicas. Os resultados foram avaliados a partir de medidas de desempenho, obtendo-se precisão média de, aproximadamente, 42%. Em trabalhos futuros, aspira-se a melhora do desempenho, adicionando mais medidas de texturais e topológicas, bem como mediante a utilização de outros classificadores.

Disponibilidade de Código

Os códigos estão disponíveis em: https://github.com/RafaelMarasca/SICITE-2023

Agradecimentos

Os autores agradecem à UTFPR-TD por todo o suporte durante o desenvolvimento do presente projeto de pesquisa através do Programa Institucional de Voluntariado em Iniciação Científica e Tecnológica (PIVIC) EDITAL PROPPG – 05/2022.

Conflito de interesse

Não há conflito de interesses.

REFERÊNCIAS

COCA, A. CLBP Texture Descriptor in Multipartite Complex Network Configuration for Music Genre Classification. **Procedia Computer Science**, v. 222, p. 331–340, 2023b. International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023). DOI: https://doi.org/10.1016/j.procs.2023.08.172.

COCA, A. GLSI Texture Descriptor Based on Complex Networks for Music Genre Classification. In: Proc. International Joint Conference on Neural Networks (IJCNN). Gold Coast, Australia: IEEE, 2023a. p. 1-8. DOI: 10.1109/IJCNN54540.2023.10191818.

COCA, A. Hierarchical mining with complex networks for music genre classification. **Digital Signal Processing**, v. 127, p. 103559, 2022. DOI: https://doi.org/10.1016/j.dsp.2022.103559.

COSTA, Y.; OLIVEIRA, L.; KOERICB, A.; GOUYON, F. Music genre recognition using spectrograms. In: Proc. International Conference on Systems, Signals and Image Processing (IWSSIP). Sarajevo, Bosnia e Herzegovina: [s.n.], 2011. p. 1-4.

HARALICK, R.; SHANMUGAM, K.; DINSTEIN, I. Textural Features for Image Classification. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-3, n. 6, p. 610–621, 1973. DOI: 10.1109/TSMC.1973.4309314.

MARTINS, J.; MALDONADO, Y.; COSTA, G.; BERTOLINI, D.; EDUARDO, L.; OLIVEIRA, L. Uso de descritores de textura extraídos de GLCM para o reconhecimento de padrões em diferentes domínios de aplicação. In: Conferência Latinoamericana de Informática (CLEI). Quito, Ecuador: [s.n.], 2011. p. 637-652.

MÜLLER, M. **Fundamentals of Music Processing**. Cham: Springer International Publishing, 2015. DOI: 10.1007/978-3-319-21945-5.

SILVA, T.; ZHAO, L. **Machine Learning in Complex Networks**. Cham: Springer International Publishing, 2016. DOI: 10.1007/978-3-319-17290-3.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. **IEEE Transactions on Speech and Audio Processing**, v. 10, n. 5, p. 293–302, 2002. DOI: 10.1109/TSA.2002.800560.

TZANETAKIS, G.; ESSL, G.; COOK, P. Audio Analysis using the Discrete Wavelet Transform. In: Proc. Conference in Acoustics and Music Theory Applications (WSEAS). Cairns, Australia: [s.n.], dez. 2001. p. 318-323.