Análise e Classificação de Gêneros Musicais com Base em Letras de Músicas

Barbara P. J. M. Ferreira[1], **Daniel H. Dalip**[2], **Ismael S. Silva**[3]

[1]Departamento de Computação – CEFET-MG
Centro Federal de Educação Tecnológica
de Minas Gerais (CEFET-MG)
Belo Horizonte, Minas Gerais, Brasil.

bpolianajaber@gmail.com[1], hasan@cefetmg.br[2], ismaelsantana@cefetmg.br[3]

Abstract. *This article presents a methodology for classifying music genres based on song lyrics. Leveraging methods including SVM, Random Forest, and Naive Bayes, we investigate two approaches for lyrics representation: one employing attributes specifically tailored for this study, and another utilizing the 'Bag of Words' technique. Furthermore, we conduct a comparative analysis between hierarchical classification, taking into account the hierarchy of music genres, and the original classification from the dataset. The results enables genre detection with valuable insights into genre-specific characteristics in the dataset, and contribute to the field of music information retrieval.*

Resumo. *Este artigo apresenta uma metodologia para classificar gêneros musicais com base nas letras das músicas. Utilizando métodos como o SVM, Random Forest e Naive Bayes, investigamos duas abordagens de representação das letras: uma empregando atributos específicos desenvolvidos para este estudo e outra utilizando a técnica "Bag of Words". Além disso, realizamos uma análise comparativa entre a classificação hierárquica, levando em consideração a hierarquia dos gêneros musicais, e a classificação original do conjunto de dados. Os resultados possibilitam a detecção de gêneros e compreensão de elementos específicos dos gêneros no conjunto de dados, além de contribuir para o campo da recuperação de informações musicais.*

1. Introdução

A análise de diferentes estilos musicais, combinada com a disponibilidade de letras em diversas plataformas, torna a classificação de gêneros musicais uma tarefa de grande importância tanto para o campo de Recuperação de Informação de Música (RIM) (ou *Music Information Retrieval, MIR*, em inglês) quanto para estudos de aspectos culturais de gêneros musicais [McKinney et al. 2003].

Além do contexto Recuperação de Informação de Música, a classificação de gêneros de música por meio de suas letras, é um tipo de problema de Processamento de Linguagem Natural (PLN), dado processamento dos textos e construção de atributos. Combinado a esse escopo de problemas, existe também o problema de classificação de múltiplas classes [Oh 2017], que é por sua vez comumente abordado em problemas de Classificação Hierárquica, como na de classificação de gêneros música a partir de áudio de músicas. [Parmezan et al. 2020].

Diante disso e da relevância da Classificação Hierárquica para problemas como esse, utilizamos três métodos de aprendizado de máquina: Support Vector Machine (SVM), Random Forest e Nayve Bayes, e comparamos em diferentes experimentos seus desempenhos com a Classificação Tradicional e uma Classificação Hierárquica proposta nesse trabalho. Para isso, foram usadas diferentes representação de dados para compará-los: a representação por *Bag of Words* (BOW) comumente usada na literatura de PLN, e um conjunto da representação de dados que contêm 17 atributos distintos, elaborados a partir de aspectos diferentes das letras de música, explicados na Seção 3. Os resultados dos experimentos realizados demonstraram que a Classificação Hierárquica é melhor que a Classificação Tradicional, tanto para a representação de dados com 17 atributos criados nesse trabalho, quanto para BOW. Para os 17 atributos, o melhor resultado global de Macro-F1 foi do Random Forest, com melhoria percentual de 3,79% da Classificação Hierárquica em relação à Tradicional. Além disso, foi possível observar que os atributos apresentam significativamente melhores resultados para os 17 atributos em comparação ao BOW.

Com o objetivo final de estabelecer um comparativo entre as metodologias e das representações na classificação dos gêneros musicais, este trabalho teve como contribuição os atributos das músicas elaborados nele e o comparativo final da Classificação Hierárquica e Classificação Tradicional. A análise e comparação mais detalhadas dos resultados obtidos é apresentada na Seção 4.

2. Trabalhos Relacionados

Pesquisas anteriores demonstraram a eficácia do uso dos métodos clássicos de aprendizado de máquina para classificação de gêneros de música para o conjunto de representação de dados de Bag of Words [Mayer et al. 2008]. Outras pesquisas também demonstraram que ter diferentes tipos de atributos, além de Bag of Words contribuem para melhor acurácia de classificação dos gêneros musicais, como por exemplo o uso de representação de dados com áudio, que facilita a criação de atributos específicos ricos para a classificação [Mayer and Rauber 2011], resultando numa acurácia de 74% para 10 gêneros musicais.

[Tsaptsinos 2017] usa a abordagem hierárquica com redes neurais de atenção, com o uso apenas de letras de música para predição, e para um dataset de quase 500 mil músicas com mais de 100 gêneros, obteve a acurácia de 68%. Já [Parmezan et al. 2020] demonstrou em sua abordagem diferentes tipos de classificação hierárquicas de gêneros musicais, utilizando apenas o método de Random Forest, e destaca os resultados da combinação da abordagem híbrida de hierarquia como a com melhores resultados.

O trabalho de [Yang and Lee 2010] utilizou apenas de letras de música para uma classificação multi-classe de emoções, cujas categorias de emoções foram originárias do estudo de psicologia de [Clark and Tellegen 1999], e propôs um processamento estatístico do texto das músicas que foram transformadas num vetor de 182 características psicológicas. O resultado final do trabalho de [Clark and Tellegen 1999] foi o banco de dados gerado com todas essas caractaterísticas.

3. Metodologia

Nessa Seção é apresentada a Metodologia desenvolvida, que foi organizada nas seguintes etapas:

1. Representação de Dados
    a. Extração de atributos
    b. Geração da representação po Bag Of Words
2. Modelagem dos algoritmos de classificação
3. Treinamentos
    a. Treinamento simples ou Classificação Tradicional
    b. Treinamento com Classificação Hierárquica
4. Avaliação dos Resultados

3.1. Representação dos dados

A base de dados escolhida foi retirada do [Kag 2018], fruto da coleta de 362237 letras de músicas, todas em inglês, da plataforma [met 2020] em 2018, com 10 gêneros musicais: Country, Folk, Hip-Hop, Indie, R&B, Jazz, Pop, Eletrônico, Rock e Metal. A quantidade de músicas consideradas para o avaliação dos algoritmos foi reduzida para 18000 devido em função do balanceamento por agrupamento de gêneros feito para Classificação Hierárquica, explicada na Seção 3.2. Os resultados dos atributos coletados nesse estudo estão disponíveis em [git 2020]

O primeiro passo foi um pré-processamento das letras para limpar o texto, para isso foram removidos todos os caracteres que não fossem letras ou números e foram tratadas todas as letras para serem minúsculas. O segundo passo foi a criação da representação de atributos, explicada a seguir.

Considere-se um conjunto de músicas X = $\{x_1, x_2, ..., x_n\}$. Cada música é representada por um conjunto de m atributos A = $\{a_1,..., a_m\}$, de tal forma que, x_i = $\{a_{i1}, a_{i2}, ..., a_{im}\}$ é o vetor que representa x_i, em que cada a_{ij} é o valor de um atributo A_j na música x_i. Neste trabalho, as métricas definidas medem estatísticas diferentes das letras música, tendo sido elaboradas com base no trabalho de [Mayer and Rauber 2011] e de ideias sobre o valor da informação contida nas letras de música dada sua estrutura de texto. Para a compreensão dos atributos criados, considera-se o conceito de *stopwords* como palavras que possuem pouco significado do ponto de vista semântico, tais como preposições, artigos, conjunções e outros.

Nesta proposta, assume-se que o acesso aos dados de treinamento é na forma $\{(x_1,g_1), (x_2,g_2), ..., (x_n,g_n)\}$, em que cada par (x_i,g_i) representa a letra da música e seu gênero correspondente. As descrições dos atributos extraídos das músicas seguem abaixo: A seguir estão os atributos utilizados neste trabalho para a representação de músicas:

1. Quantidade termos únicos da música, sem *stopwords*;
2. Quantidade de termos únicos da música, considerando as *stopwords*;
3. Ruído de informação (do inglês, *Information to Noise* 1 e 2): Divisão entre as relações encontradas para o Atributo 1 e o Atributo 2;
4. Densidade de palavras únicas sem *stopwords* na música: Divisão entre a quantidade de termos únicos sem *stopwords* pelo número total de termos da música com *stopwords*;
5. Quantidade de termos únicos total normalizada: Total de termos únicos da música com

stopwords, normalizado em relação ao número de termos únicos;
6. Quantidade de versos únicos total normalizada: Total de versos únicos da música, normalizado conforme a Equação 1:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

7. Média de termos por estrofe na música: Média de termos únicos por estrofe na música, considerando *stopwords*, onde cada estrofe é um segmento de 4 versos;
8. Média de termos únicos por versos, com *stopwords*;
9. Desvio padrão de termos únicos por versos, com *stopwords*;
10. Quantidade máxima de termos únicos por versos normalizada, com *stopwords*;
11. Quantidade mínima de termos únicos por versos normalizada, com *stopwords*;
12. Média de termos únicos por versos, sem *stopwords*;
13. Desvio padrão de termos únicos por versos, sem *stopwords*;
14. Quantidade máxima de termos únicos por versos normalizada, sem *stopwords*;
15. Quantidade mínima de termos únicos por versos normalizada, sem *stopwords*;
16. Índice de positividade da música: Divisão entre a frequência de termos considerados palavras positivas em relação ao número total de termos sentimentais da música.
17. Índice de negatividade da música: Divisão entre a frequência de termos considerados palavras negativas em relação ao número total de termos sentimentais da música.

Tabela 1. Vocabulário para índices de positividade e negatividade

Índice de	Descrição
$positividade$	$baby, confidence, confident, cute, dear, faith, good, honey, hope, hot, life, like, love, peace, right, safe, solitude, truth$
$negatividade$	$bad, broken, cold, cry, dead, death, fear, guilty, hate, helpless, hopeless, hurt, lie, lies, isolation, lonely, pain, sad, sadness, tears, war, wrong$

Diante do exposto, a representação de cada um desses atributos é referenciada neste trabalho como Atributo $1, 2, 3, .., 17$ um conjunto de todos eles juntos é referenciado como Atributos. Já a representação com Bag of Words é referenciada apenas como Bag of Words. Os resultados dos atributos gerados pode ser acessado no [git 2020].

3.2. Classificação Hierárquica

Para elaborar estratégias de Classificação Hierárquica é possível utilizar de diferentes agrupamentos e formas de agrupamento, conforme é demonstrado por [Nakano et al. 2017]. Neste trabalho, foi elaborada a proposta de agrupar alguns gêneros musicais a fim de gerar uma Classificação Hierárquica dos mesmos, com base na similaridade entre os gêneros musicais demonstrada no trabalho de [Parmezan et al. 2020].

Para cada agrupamento $g \in G$, em que G é o conjunto de agrupamentos musicais feitos, g é um grupo de um ou dois gênero agrupados. Dessa forma, existem apenas dois níveis na hierarquia de classificação. No primeiro nível dela, classifica-se a música como parte de um agrupamento g, e no segundo e último nível, classifica-se a música

como parte de um único gênero musical. Esses agrupamentos são explicitados na Tabela 2 assim como os gêneros que permaneceram sem agrupamento.

Os gêneros Hip-Hop e Indie são os gêneros que não foram agrupados pois, dados os 10 gêneros musicais da base de dados, não haviam gêneros similares a eles, em relação às características mostradas por [Parmezan et al. 2020] e às impressões de estilo. Ainda assim, esses gêneros passam pelo processo de verificação de qual grupo pertencem, durante a Classificação Hierárquica, assim como os agrupados. Esse processo é explicado na SubSeção 3.3.

A Tabela 2 apresenta a quantidade de músicas para cada gênero após o balanceamento da base de dados feito para que os grupos tivessem a mesma quantidade de músicas entre si. Como na base de dados original o *Indie* possuía apenas 3140 músicas, reduziu-se o tamanho dos grupos para 3000 músicas para cada um.

Tabela 2. Agrupamento de gêneros e quantidade de músicas por gênero e grupo

Gênero original	Gênero agrupado	Grupo	N	N por grupo
Country	Folk	1	2622	3000
Folk	Country	1	378	3000
Hip-Hop	-	2	3000	3000
Indie	-	3	3000	3000
R&B	Jazz	4	882	3000
Jazz	R&B	4	2118	3000
Pop	Eletrônico	5	2530	3000
Eletrônico	Pop	5	480	3000
Rock	Metal	6	2461	3000
Metal	Rock	6	539	3000

3.3. O Algoritmo de Classificação Hierárquica

A forma que foi desenvolvida a Classificação Hierárquica neste trabalho está ilustrada no Algoritmo 1 que mostra como é feita a filtragem das bases de treino e de teste para a predição do gênero real da música. Para compreender o Algoritmo 1, na Tabela 3 são descritos os significados de cada variável. Antes do ciclo de repetição iniciar, é feita a classificação no primeiro nível. Inicialmente, antes do primeiro laço de repetição, cria-se um modelo M_1 para a predição com base no grupo de gênero. Nesse momento, $\hat{y}^{tr}$ possui as predições do primeiro nível.

Para iniciar as predições no segundo nível, considera-se g um agrupamento do conjunto de agrupamentos G, no laço de maior escopo. No primeiro laço dentro desse, é preparada a base de treino para o segundo nível, a partir da filtragem dos gêneros preditos na classificação do primeiro nível, ou seja, os grupos de gênero. A seguir, para preparar a base de teste, para cada instância de treino $x_i \in X$, se o vetor de predições na instância do grupo i pertencer ao grupo g, então ele é filtrado para base de testes. Finalizada a preparação do treino e teste, é criado o modelo M_2 que é treinado com a matriz de treino completa do grupo g, $X^{tr[g]}$ e o vetor de classe alvo de treino do grupo g, $y^{[g]}$. Depois é feita a predição da base de testes do grupo g, $X^{[g]}$, com o vetor de predições de teste $\hat{y}^t$, e o resultado é inserido no vetor de predições de teste do grupo g, $\hat{y}^{t[g]}$. Por fim, o resultado das predições é dado por $\hat{y}^f$.

Tabela 3. Descrição das variáveis do Algoritmo

Termo	Descrição
y^{tr}	Vetor de classe alvo do treino
$y^{tr[i]}$	Vetor de classe alvo, considerando o i-ésimo grupo de gênero
$\hat{y^{tr}}$	Vetor das predições do teste, do primeiro nível
$\hat{y}_i$	Vetor das predições do teste, considerando o i-ésimo grupo de gênero.
X^{tr}	Matriz do treino completa
$X^{tr[i]}$	Matriz do treino completa, considerando apenas intâncias do i-ésimo grupo de gêneros
X	Atributos do teste completa
$X^{[i]}$	Atributos do teste completa, considerando apenas intâncias do i-ésimo grupo de gêneros
M_1	Modelo primeiro nível
M_2	Modelo segundo nível
G	Conjunto de grupos
$\hat{y}^f$	Vetor de predições final

Algorithm 1 Classificação Hierárquica

for all g em G **do**
  prepara o treino, considerando apenas instâncias em que o y_i^{tr} **pertence ao grupo** g
  for all $x_i \in X^{tr}$ **do**
    if $y_i^{tr} = g$ **then**
      $X^{tr[g]} \leftarrow X^{tr[g]} \cup \{x_i\}$
      $y^{tr[g]} \leftarrow y^{tr[g]} \cup \{y_i^{tr[g]}\}$
    end if
  end for
  prepara o teste, considerando apenas instâncias em que o $y^{tr[i]}$ **pertence ao grupo** g
  for all $x_i \in X$ **do**
    if $\hat{y}_i = g$ **then**
      $X^{[g]} \leftarrow X^{[g]} \cup \{x_i\}$
    end if
  end for
  $M_2 \leftarrow$ cria modelo ($X^{tr[g]}$, $y^{tr[g]}$)
  $\hat{y}^{[g]} \leftarrow$ prediz($X^{[g]}$)
  $\hat{y}^f \leftarrow \hat{y}^f \cup \hat{y}^{[g]}$
end for

3.4. Metodologia de Avaliação

A metodologia de avaliação escolhida foi a Validação Cruzada *K-fold* para avaliar o desempenho do modelo de classificação. O conjunto de dados em K subconjuntos de tamanho igual, consistindo em conjuntos de treinamento, teste e validação [Baeza-Yates and Ribeiro-neto 1999]. O valor de K é selecionado com base no *trade-off* entre viés e variância, sendo comumente adotados os valores 5 e 10, que demonstraram ser os melhores *trade-offs* [Baeza-Yates and Ribeiro-neto 1999]. Neste trabalho, foi utilizado $K = 5$. Desses 5 *folds*, 3 *folds* são utilizados para treinamento, 1 *fold* para teste e 1 *fold* para validação em cada partição. O *fold* de validação é empregado na estimação dos parâmetros nos *folds* de treinamento e aplicado ao *fold* de teste. Essa abordagem permite treinar o modelo e avaliá-lo com diferentes hiperparâmetros.

Para estudar avaliar o desempenho dos métodos utilizados, foram utilizadas diferentes métricas de avaliação: precisão, revocação, F1 e Macro-F1, comumente usadas [Russell and Norvig 2002]. Além disso, utilizou-se de matrizes de confusão para representação da acurácia dos modelos de classificação. As matrizes de confusão desse trabalho são demonstradas na Seção a seguir.

4. Resultados

Os experimentos para a representação dos Atributos foram feitos com Classificação Hierárquica e Classificação Tradicional, para cada método de Aprendizado de Máquina com a representação de Atributos, com as implementações são da biblioteca de Python, Scikit-learn de [Pedregosa et al. 2011]. Para variar os hiperparâmetros dos métodos foi utilizado o *Tree-Structured Parzen Estimator (TPE)* é uma abordagem que escolhe hiperparâmetros de forma aleatória porém enviesado. Ele é enviesado porque existe uma probabilidade maior de que o TPE escolha valores para hiperpârametros em regiões do plano que ainda não foram exploradas ou regiões do plano que já foram exploradas mas possuem um bom resultado [Baeza-Yates and Ribeiro-neto 1999].

Para cada experimento, considera-se a configuração de Validação Cruzada de 5 *folds*. Sendo que para cada *fold*, variou-se 100 vezes os hiperparâmetros de cada método pelo TPE. Cada vez que se varia um hiperpâmetro, trata-se de um ensaio cuja implementação foi feita com a biblioteca Optuna proposta por [Akiba et al. 2019]. Ou seja, para cada *fold* de um dado experimento, são feitos 100 ensaios, com valores de hiperpâmetros variados pelo TPE. Os valores utilizados pelo TPE para variação de cada método para essa configuração de experimento estão demonstrados na Tabela 4.

Tabela 4. Variação de hiperparâmetros com representação de Atributos

Método	**Variação de hiperparâmetro**
SVM	*Custo C: 2^-5 a 2^15*
Random Forest	*Número mínimo de instâncias para poda: 10% a 50%,* *Número máximo de atributos: 10% a 90%* *Número de árvores: 5 a 100*
Naive Bayes	Nenhum

Os hiperparâmetros variados tem significados diferentes para cada método. O Custo C é variado no SVM, conforme a Tabela 4, aumentando o expoente de 2 em 2, como proposto por [Hsu et al. 2003]. Esse Custo C, como explicado no Capítulo 2 determina o comportamento da margem do algoritmo, por isso sua variação determina a classificação do mesmo. Para o Random Forest é variado o *Número mínimo de instâncias para poda* que, supondo um valor de 10%, significa que se o tivermos menos de 10% de instâncias para um nodo, ele vira uma folha, ou seja, acontece uma poda, o que reduz o *overfiting* [Russell and Norvig 2002]. O *Número máximo de atributos*, supondo um valor de 10%, significa que as árvores do Random Forest serão criadas com apenas 10% dos atributos das músicas. E o *Número de árvores* de árvores criadas é o que o nome do hiperparâmetro diz por si só. Por fim, para o Naive Bayes não foi feita variação de parâmetros, e foi utilizado o modelo gaussiano do método sem variação de parâmetros, o que é adequado para o tipo de classificação proposta, segundo [Russell and Norvig 2002].

Nos experimentos com a representação por Bag of Words não foram variados todos os hiperparâmetros dos métodos de Aprendizado de Máquina. Enquanto o SVM seguiu o padrão de variação descrito na Tabela 4, o número de árvores no Random Forest foi mantido constante em 10. Na avaliação dos resultados, a melhor combinação de parâmetros foi aplicada nos conjuntos de teste dos folds em cada partição da Validação

Cruzada, e a performance resultante foi medida. Dessa forma, o resultado final reportado de cada experimento é a média dos resultados das 5 partições da Validação Cruzada. Já para a representação de Atributos, foram realizados 100 ensaios, enquanto para a representação Bag of Words foi executado apenas um ensaio, exceto para o SVM, que foi executado em 2 ensaios. A diferença no número de ensaios entre Bag of Words e Atributos pode estar relacionada às diferenças nos resultados obtidos. No entanto, é importante destacar que o Bag of Words naturalmente demanda mais tempo de execução e memória. Por esse motivo, não foi considerado viável executar um número maior de ensaios para essa representação.

4.1. Análise de performance dos métodos

A métrica utilizada para avaliar os métodos foi a Macro-F1, que é especialmente indicada para lidar com bases de dados desbalanceadas, conforme destacado por [Baeza-Yates and Ribeiro-neto 1999]. Dado que os gêneros musicais não agrupados eram desbalanceados, a escolha da Macro-F1 mostrou-se apropriada para esse contexto. A Tabela 5 apresenta os resultados dos experimentos conduzidos neste trabalho, considerando as representações por Atributos e Bag of Words, tanto para as abordagens de Classificação Hierárquica quanto para a Classificação Tradicional.

	Atributos		**Bag of Words**	
Método	*Hierárquico*	*Tradicional*	*Hierárquico*	*Tradicional*
SVM	43,73%	42,11%	32,18%	35,78%
RF	**58,67%**	56,53%	22,42%	14,53%
NB	46,06%	38,33%	22,85%	23,51%

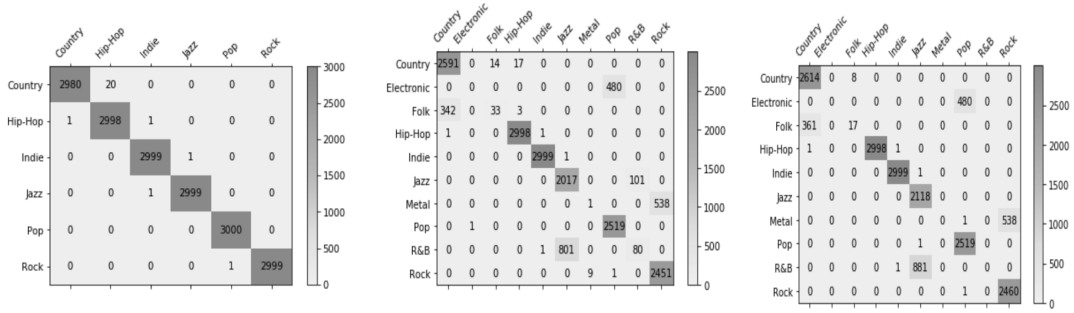
Tabela 5. Macro-F1 dos Experimentos

Conforme demonstrado na Tabela 5, para a representação por Bag of Words, com Classificação Tradicional, o método com melhor resultado é o SVM com apenas $35,78\%$ de Macro-F1, contra $14,53\%$ e $23,51\%$ do Random Forest e Naive Bayes, respectivamente. O SVM foi avaliado em dois ensaios diferentes, nos quais houve variação de hiperparâmetros, evidenciando a influência positiva desse ajuste na obtenção do resultado final. Por outro lado, o método Naive Bayes, não teve seus hiperparâmetros variados em nenhum de seus experimentos e seu resultado com a representação de Atributos para Classificação Tradicional é de $38,33\%$, que é maior que seu resultado $23,51\%$, para Bag of Words.

Outra característica comum dos experimentos com Naive Bayes é que como para este nenhum parâmetro é variado, seus ensaios tem sempre os mesmos parâmetros, o que qualitativamente é encarado como um ensaio apenas, mesmo quando são feitos 100 ensaios para o mesmo experimento. Logo, pode-se concluir que a representação dos Atributos é, em média, melhor que a representação por Bag of Words. Num comparativo geral, analisando a média geral da Classificação Hierárquica entre os resultados para os Atributos, Ao comparar as metodologias de classificação, levando em consideração apenas a representação de Atributos, percebe-se que os resultados da Macro-F1 para o modelo hierárquico, são todos melhores que os do tradicional. O que por fim, valida a ideia inicial deste trabalho de que a Classificação Hierárquica tem um desempenho melhor para

o problema de classificação em questão. O resultado baixo do Bag of Words pode ser atribuído função da menor variação dos hiperparâmetros. Mesmo assim, considerando o tempo de execução de ambas as representações, a de Atributos é mais vantajosa do que o Bag of Words. Só que não se descarta a possibilidade de, em trabalhos futuros, executar os experimentos de Bag of Words com variações maiores para obtenção de melhores resultados. Com o objetivo de analisar a classificação por classe predita, o experimento de melhor resultado para representação de Atributos, é analisado na subSeção a seguir. E na Seção 5 é demonstrada a relevância dos Atributos propostos neste trabalho.

O método Random Forest, foi o método que teve melhor desempenho entre os experimentos da representação de Atributos, com $58,67\%$ de Macro-F1 para Classificação Hierárquica e $56,53\%$ para Tradicional, com uma melhoria percentual de 3,79%. Por isso, ele foi o método escolhido para análise por classe de predição, que é feita a seguir pela explicação de suas matrizes de confusão demonstradas na Figura 1. Em que o eixo x representa a classe (gênero) predita pelo modelo e o eixo y a classe (gênero) real das músicas.



(a) Cl . Hierárquica - Primeiro nível **(b) Cl. Hierárquica - Segundo nível** **(c) Cl. Tradicional**

Figura 1. Matrizes de Confusão

As Figuras 1a e 1b representam as predições da Classificação Hierárquica no primeiro e segundo nível, respectivamente. Nota-se que na Figura 1a, no primeiro nível, praticamente todas os instâncias são preditas como de sua classe corretamente. Existe um erro, mas ele é pequeno. Já a Figura 1b, o segundo nível do mesmo experimento da Figura 1a, tem maior semelhança com a Figura 1c, porque ambas correspondem aos resultados que predizem as classes de gêneros reais, não agrupados. Ao analisar detalhadamente as Figuras 1b e 1c percebe-se que os gêneros não agrupados, Hip-Hop e Indie, são bem classificados nos dois casos e tem o mesmo erro. Além disso, é possível observar a predição dos gêneros Pop e Eletrônico, somente como Pop, e ambos fazem parte do grupo Pop.

Outra situação visível na matriz da Figura 1b é a do gênero R&B em que pelo menos 80 instâncias que são R&B originalmente, foram classificadas como tal, enquanto que na Classificação Tradicional na Figura 1c, todas as instâncias R&B foram classificadas como Jazz, que é o gênero do agrupamento de R&B e Jazz. A mesma situação ocorre com o gênero o Folk, que teve 33 instâncias preditas corretamente no modelo hierárquico enquanto que no tradicional, teve apenas 17.

Diante do exposto, fica evidenciado que a Classificação Hierárquica possui um erro menor que a Classificação Tradicional nas predições. E por isso, conclui-se que

obteve um melhor resultado. Mas esse resultado não se deve apenas à Classificação Hierárquica, pois como demonstrado na Tabela 5 a representação de Atributos foi a melhor representação, contribuindo para essa conclusões.

5. Relevância dos Atributos

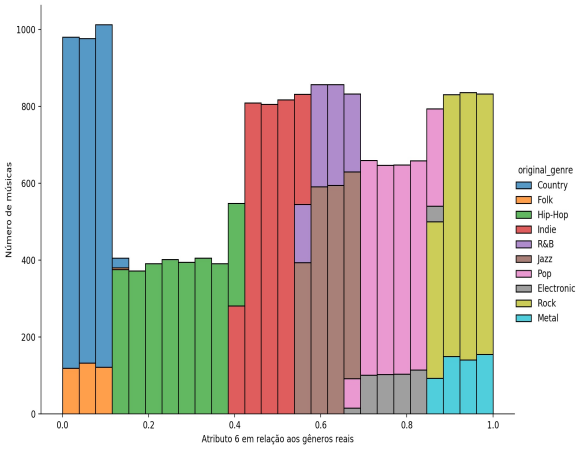
A relevância dos Atributos propostos neste trabalho é discutida com base no *ranking* do Ganho de Informação (do inglês, *info gain*) dos Atributos, demonstrado na Tabela 2a. Esse cálculo foi feito com o auxílio da ferramenta Weka, proposta por [Hall et al. 2009]. O Ganho de Informação é uma medida baseada em impureza, o qual usa entropia como medida de impureza [Russell and Norvig 2002]. Conforme a Tabela 2a fica evidente que ganho de informação do Atributo 6, que é a relação da *Quantidade de versos únicos* de cada música, tem a maior relevância diante dos demais. Por isso, foi feito um gráfico que toma como referência a *Quantidade de versos únicos* e os gêneros reais das músicas. Esse gráfico está representado na Figura 2b.

Analisando o gráfico da Figura 2b, percebe-se que existe a distribuição da *Quantidade de versos únicos* das músicas em relação a número de músicas, para cada um dos gêneros reais. Por exemplo, o gênero Country, representado pela cor azul escuro se encontra no início do gráfico, com uma quantidade versos relativamente menor que os demais gêneros, junto ao gênero Folk. O Hip-Hop, que engloba gêneros como Rap, Trap-Hop, dentre outros, possui quantidade de versos maior em relação ao gênero Country. No entanto, possui uma distribuição maior em relação aos demais gêneros, possivelmente por englobar gêneros que possuem menos versos, de um modo geral.

(a) Ranking de Ganho de Informação dos Atributos

Atributo	Ganho de Informação
6	*2,585*
5	0,3187
2	0,3177
1	0,3163
4	0,2597
14	0,1192
10	0,1192
11	0,1192
7	0,1169
8	0,105
12	0,105
9	0,0877
13	0,0877
15	0,0824
16	0,0643
17	0,0606
3	0

(b) Valor do Atributo 6 para os gêneros originais



Dessa forma, nota-se um padrão ao longo do gráfico. Os gêneros que foram agrupados por terem semelhanças demonstradas no trabalho proposto por [Parmezan et al. 2020], se encontram próximos na disposição do gráfico da Figura 2b. O Metal, por sua vez, se encontra dividindo a posição de maior Ganho de Informação com o Rock, o que pode ser explicado em função de o Metal ter comumente um vocabulário maior, que outros gêneros musicais [Martín-Gómez and Navarro Cáceres 2018].

Ainda sobre esse gráfico, é possível analisar o gênero Pop, que não tem um histórico de versos muito elaborados, no entanto, ele apresenta uma quantidade de versos representada pela faixa entre 0.6 e 0.83, aproximadamente. Isso pode ter acontecido pois o Atributo 6 contabiliza termos como *"yeah", "oh"* entre outros desses tipos, que pode tornar um verso único em relação ao outro simplesmente pela diferença do número de *"yeah"* presentes em cada um. Assim, pode-se ter um número maior de versos únicos, mesmo que eles sejam semânticamente idênticos. Essa mesma análise se aplica ao gênero Eletrônico. Conforme a visualização da distribuição das quantidades de versos por gêneros reais, é possível observar que os gêneros que dividem as mesmas faixas de quantidade, como Rock e Metal, com os valores de 0.84 a 1.0, são gêneros que foram agrupados, confirmando assim, a validade dos agrupamentos por similaridade propostos neste trabalho.

6. Conclusão

Este trabalho identificou que a Classificação Hierárquica apresenta um desempenho superior em relação à Classificação Tradicional para a representação de atributos criados. A combinação dessa metodologia com representações relevantes, como a quantidade de versos, pode resultar em um desempenho ainda melhor, sendo uma possível direção para futuras pesquisas.

Ao comparar as representações de dados utilizadas nos experimentos deste trabalho, observamos que a média dos resultados para o método Bag of Words foi menor do que a média dos atributos propostos neste trabalho. Demonstrando a relevância dos atributos criados. Por outro lado, é importante considerar que realizamos menos ensaios para o Bag of Words, por sua execução ser mais lenta. Por isso, para trabalhos futuros, podemos explorar uma maior variação de hiperparâmetros para Bag of Words, mas também propor outras formas de Classificação Hierárquica, com mais níveis, e outras combinações possíveis de agrupamentos, a fim de estudar as diferenças de agrupamentos e os seus impactos na classificação. Além disso, é possível aplicar a metodologia desenvolvida para outros bancos de dados de músicas, para realizar um comparativo com massa de dados maiores e outros métodos de aprendizado de máquina.

Referências

(2018). Kaggle. `https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics/data`. Acessado em: 2019-10-30.

(2020). Metrolyrics. `https://www.metrolyrics.com/`. Acessado em: 2019-10-30.

(2020). Repositório dos atributos coletados. `https://github.com/bpoliana/music-genre-analysis/blob/master/18k_features_1_17_normalized.csv`. Acessado em: 2023-05-26.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Baeza-Yates, R. and Ribeiro-neto, B. (1999). Modern information retrieval.

Clark, L. and Tellegen, A. (1999). On the dimensional and hierarchical structure of affect. *Lee Anna Clark*, 10.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Hsu, C.-w., Chang, C.-c., and Lin, C.-J. (2003). A practical guide to support vector classification chih-wei hsu, chih-chung chang, and chih-jen lin.

Martín-Gómez, L. and Navarro Cáceres, M. (2018). Applying data mining for sentiment analysis in music. pages 198–205.

Mayer, R., Neumayer, R., and Rauber, A. (2008). Rhyme and style features for musical genre classification by song lyrics. pages 337–342.

Mayer, R. and Rauber, A. (2011). Music genre classification by ensembles of audio and lyrics features. In *ISMIR*.

McKinney, M., Breebaart, J., and (wy, P. (2003). Features for audio and music classification.

Nakano, F. K., Pinto, W., Pappa, G., and Cerri, R. (2017). Top-down strategies for hierarchical classification of transposable elements with neural networks. pages 2539–2546.

Oh, S. (2017). Top-k hierarchical classification. In *AAAI Conference on Artificial Intelligence*.

Parmezan, A., Silva, D., and Batista, G. (2020). A combination of local approaches for hierarchical music genre classification.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Russell, S. J. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall.

Tsaptsinos, A. (2017). Lyrics-based music genre classification using a hierarchical attention network.

Yang, D. and Lee, W.-S. (2010). Music emotion identification from lyrics. pages 624 – 629.