Quando não se chora ao Choro: aprendizado de máquina falha em identificar emoções em músicas a partir do gênero musical.

Palavras-Chave: Aprendizado de Máquina, Música, Computação Afetiva

Autores/as:

Leonardo Vilela de Abreu Silva Pereira [FEEC]

Prof. Dr. Tiago Fernandes Tavares [FEEC]

INTRODUÇÃO:

Ao longo dos anos o aprendizado de máquina tem sido cada vez mais estudado, discutido e aplicado em diversas áreas do conhecimento. Entre essas aplicações, existem campos ligados à música e à indústria musical, em especial em sistemas de recomendação. Iremos explorar este território neste trabalho através das lentes da computação afetiva.

Cada pessoa se relaciona de maneiras diferentes com obras musicais, ou seja, uma mesma música pode gerar euforia catártica para alguns enquanto provoca retração e introspecção em outros. Experiências musicais são frequentemente referidas sob a perspectiva de emoções ou sentimentos, mas a música também tem um papel ligado ao pertencimento a grupos sociais e culturais. Esses grupos podem, com o tempo, se diferenciar de outros também através do uso de estéticas musicais específicas, que se refletem no uso de técnicas e instrumentações típicas, gerando novos gêneros musicais[1].

Tanto os sentimentos quanto o pertencimento a grupos e as percepções de gosto musical surgem no indivíduo a partir de suas interações e identificações com grupos sócio-culturais. Deste processo, surge a ideia de, por exemplo, um sorriso representar alegria [2].

Neste trabalho, estudamos a relação entre estes dois aspectos - gênero musical e sentimento - a partir de um ponto de vista de aprendizado de máquina. Mais especificamente, investigamos se emoções podem ser utilizadas para prever gênero musical e vice-versa. O objetivo deste estudo é de entender se gênero musical e sentimentos relacionados a uma música são diferentes perspectivas de um mesmo fenômeno, ou se são informações complementares.

REFERENCIAL TEÓRICO

O nosso aprendizado no que tange a expressão e a externalização de sentimentos se dá na infância [2]. Tal processo está intimamente ligado com o desenvolvimento de capacidades de comunicação que nos permite não somente externalizar estes sentimentos aprendidos inerentemente, como também emitir e absorver ideias básicas dos grupos que habitamos.

Tal qual os sentimentos e suas expressões, os processos de fazer e escutar música também são sociais. As reações e emoções que são evocadas a partir de canções, ritmos e instrumentos específicos são aprendidas ao interagir com a sociedade e cultura ao nosso redor [1] e, com o tempo, estes métodos de expressão que são específicos a certa comunidade podem se espelhar e dar origem a um gênero musical [3].

Tanto o sentimento, quanto o gênero musical podem ser ferramentas importantes no que tange predição de material musical por aprendizado de máquina, possibilitando categorização, análise e pesquisa em uma base de dados. Por essa razão a busca automática por sentimento e gênero musical tem sido altamente pesquisada.

METODOLOGIA:

Neste projeto nós utilizamos o Emotify Dataset [4]. Ele contém 400 músicas reunidas a partir do Magnatune Dataset. Cada música tem associado a si um rótulo de gênero dado pelas gravadoras gerando uma base de dados balanceada entre Rock, Pop, Música Clássica e Eletrônica.

Cada música foi avaliada por usuários que mediam o conteúdo afetivo de cada canção. Essas medidas advindo da Geneva Emotional Music Scale (GEMS) que consiste em valores binários em 9 categorias: Espanto, Solenidade, Ternura, Nostalgia, Calma, Força, Ativação Alegre, Tensão e Tristeza.

Em seguida, procedemos para o pré-processamento da base de dados. Primeiramente

calculamos a $\mu_{k,s}$, o número de usuários que avaliaram a música m como positiva para a emoção k dividido pelo total de avaliações da música m. Segundamente calculamos a mediana de $\mu_{k,s}$ para todas as músicas para a emoção k. Finalmente nós atribuímos o valor 1 para a emoção k na música m se $\mu_{k,s}$ > mediana e 0 caso contrário.

Experimentos:

Nós realizamos diversos experimentos no Emotify dataset. Primeiramente nós utilizamos um classificador para predição de gênero e conteúdo afetivo diretamente do conteúdo de áudio da música. Segundamente utilizamos tanto os resultados anteriores, quanto os rótulos afetivos dados para predição de gênero musical.

Experimentos com áudio:

Para a classificação de áudio foi utilizada uma rede neural VGG19 pré-treinada [5]. Os dados gerados pela VGG19 foram em seguida normalizados por um *standard scaler* e, posteriormente, classificados por um classificador *K-Nearest-Neighbors*. Nos nossos experimentos utilizamos k = 5 neighbors e divididos aleatoriamente 70% dos nossos dados para treino e 30% para teste.

A projeção PCA dos nossos dados de teste, mostrados na Figura 1, ilustra como a VGG19 foi capaz de encontrar fatores que separam de forma eficiente cada gênero musical na base de dados. Como é de se esperar nessa situação a nossa precisão alcançada foi de 100%, como mostrado na matriz de confusão da Figura 2.

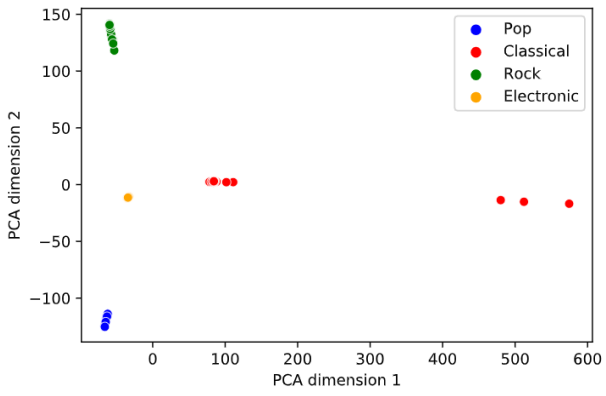

Figura 1: Projeção PCA do espaço de features de áudio colorido pelos gêneros dados.

Utilizando os mesmos aparatos do experimento anterior partimos para a predição de sentimentos da escala GEMS a partir de amostras de áudio. A precisão para cada afeto é mostrada na Tabela 1. Como podemos observar, a precisão é significativamente menor do que o caso de predição de gêneros.

A Figura 3 mostra as matrizes de confusão para cada uma das emoções. As matrizes de confusão confirmam que prever rótulos afetivos é uma tarefa mais árdua para a máquina, quando comparada a predição de gênero musical.

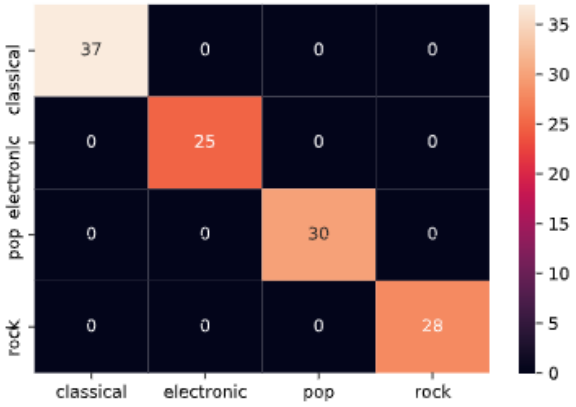

Figura 2: Matriz de confusão para áudio como entrada e gênero como saída.

Apesar de gêneros e rótulos afetivos surgirem de um diálogo entre indivíduos e a sociedade ao redor deles [1,2] e do mesmo conjunto de dados ter sido usado para prever tanto gêneros quanto afetos no passado [6,7], é possível observar que os dados gerados pela VGG19 são mais eficaz para predição de gênero musical do que de emoções.

Emoção	Precisão
Espanto	0.52
Solenidade	0.67
Ternura	0.67
Nostalgia	0.63
Calma	0.57
Força	0.62
Ativação Alegre	0.52
Tensão	0.63
Tristeza	0.43

Tabela 1: Precisão para predição de emoções utilizando áudio como entrada.

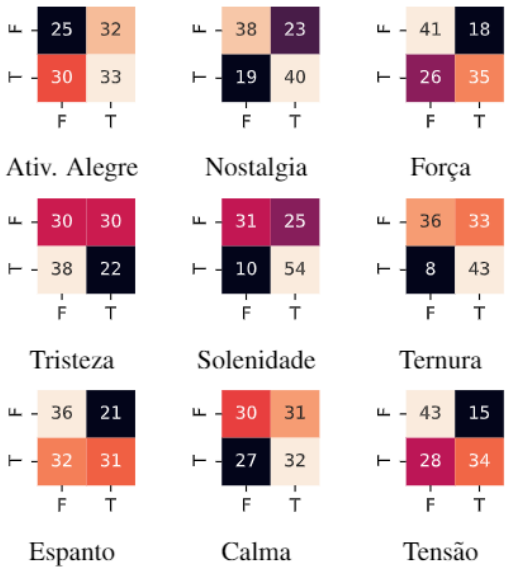

Figura 3: Matrizes de confusão para predição de cada emoção utilizando áudio como entrada.

Experimentos no espaço afetivo:

Os experimentos mostrados nesta seção utilizam as avaliações na escala GEMS para prever gênero musical. Nós realizamos dois experimentos: o primeiro usou avaliações calculadas no estágio de pré-processamento descrito na seção de Experimentos anteriormente e o segundo utilizou as probabilidades estimadas pelo classificador KNN. Estes valores foram em seguida passados para um novo classificador KNN com a tarefa de prever rótulos de gênero musical. Tal qual o último experimento utilizamos k = 5 neighbors e divididos aleatoriamente 70% dos nossos dados para treino e 30% para teste.

As avaliações na escala GEMS providos na base de dados geraram um espaço vetorial interessante, como mostra a Figura 4. Como podemos ver, apesar da especulação inicial de que cada gênero ocuparia uma região diferente do espaço vetorial, existem muitas intersecções entre regiões. É difícil determinar se isso é uma característica dos gêneros musicais por si ou se é um enviesamento da base de dados.

Os resultados da predição, mostrados na Figura 5, indicam uma precisão maior quando tratamos de música Pop. Isso pode estar relacionado à predominância da música Pop no canto inferior esquerdo da projeção PCA da Figura 4. Retornamos ao ponto de que isso pode ser tanto uma característica do gênero, quanto um enviesamento da base de dados.

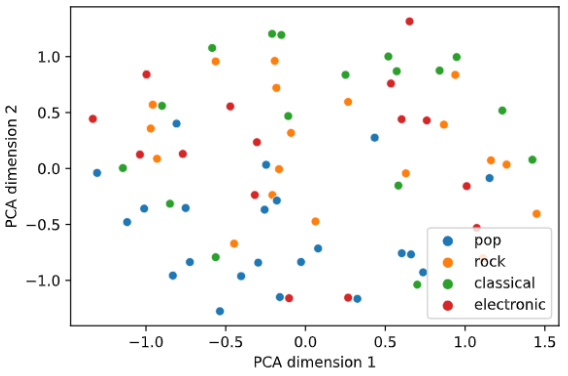

Figura 4: Projeção PCA do espaço vetorial gerado pelas avaliações na GEMS da base de dados.

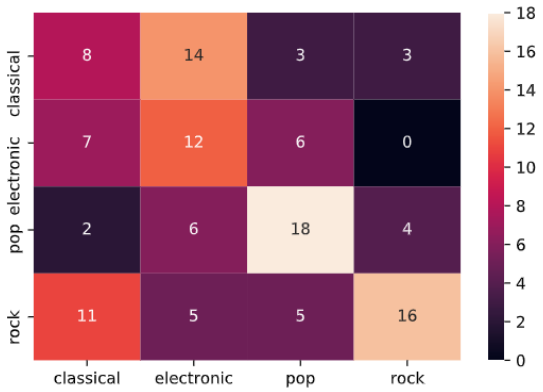

Figura 5: Matriz de confusão para predição de gênero utilizando avaliações na GEMS dados pela base.

Quando geramos o espaço vetorial utilizando as probabilidades estimadas do classificador baseado no áudio, resultados diferentes são alcançados. O espaço gerado por estas predições, como mostrado na Figura 6, indica uma interpolação ainda maior nas regiões ocupadas por cada gênero quando comparado ao resultado anterior. A matriz de confusão para este experimento, mostrada na Figura 7, indica uma grande tendência de polarização para a música clássica.

RESULTADOS E DISCUSSÃO:

Neste trabalho nós fizemos experimentos com a predição de gênero e emoção a partir de áudio. Em seguida nós tentamos usar rótulos ligados a emoção para gerar um espaço de features que permitisse a predição de gênero musical. Nossos experimentos foram conduzidos no Emotify dataset que contém tanto gênero musical, quanto rótulos ligados à emoção.

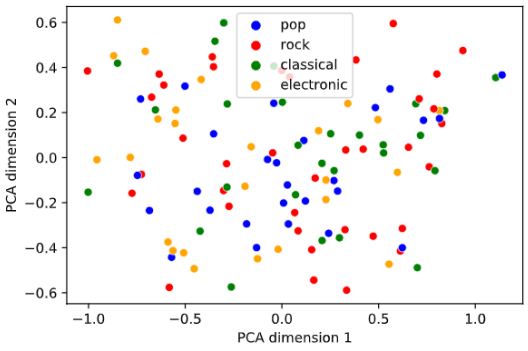

Figura 6: Projeção PCA do espaço vetorial gerado pelas avaliações na GEMS preditos do áudio.

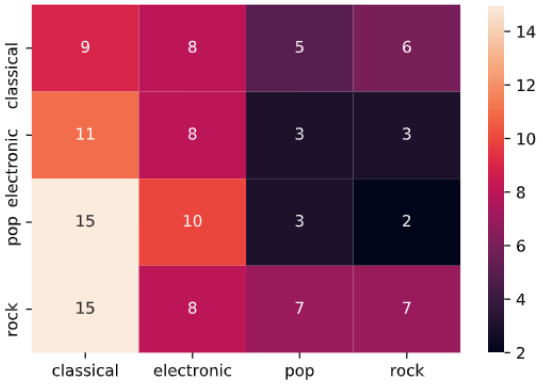

Figura 7: Matriz de confusão para predição de gênero utilizando avaliações na GEMS preditas do áudio.

Nós utilizamos a ideia que tanto gêneros e sentimentos ligados a música surgem de um diálogo entre o indivíduos e a sociedade ao seu redor. Gêneros musicais são relacionados intimamente com a sociedade e seus afetos que emergem entre os seus indivíduos e emoções são, também, relacionadas a identidade que um indivíduo divide com sigo mesmo a sociedade ao seu redor. Essa ideia é apoiada por estudos psicológicos [1] e neurocientíficos [2].

Apesar disso, nós observamos que a precisão da predição de gêneros musicais é consideravelmente maior que a mesma para emoção. Isso significa que, pelo menos para esta base de dados, gênero e emoção não são diferentes perspectivas de um mesmo fenômeno. Ao invés disso, cada um deles deriva de um diferente processo subjetivo.

Uma característica importante da base de dados é a rotulação dos gêneros musicais, dada por gravadoras, ou seja elas têm um significado que condiz com sua comercialização. Os rótulos de emoção, por outro lado, foram dados por humanos, logo eles estão relacionados a uma subjetividade inerente à definição própria de sentimento. Além disso, um modelo afetivo específico foi utilizado - a escala GEMS - podendo causar também ruído na predição.

Tal precisão, apesar de baixa quando comparada com a predição de gênero, não significa que o experimento falhou. Pelo contrário, tais resultados podem ser fruto da experiência subjetiva emocional relacionada ao fato de ouvirmos música. Esta subjetividade causa, inclusive, que as pessoas discordem entre si o significado das emoções na GEMS, logo o aprendizado de máquina não seria capaz de atingir uma precisão elevada, visto que isso significaria uma habilidade super-humana de predição de emoções, que são um fenômeno inerentemente humano, e assim teríamos, evidentemente, um enviesamento do algoritmo para a base de dados que usamos.

Em seguida, nós utilizamos avaliações de emoções para a predição de gênero musical. Este experimento teve uma precisão visivelmente menor quando comparada a predição de gênero utilizando áudio. Isso pode ser atribuído a uma série de elementos.

Uma possibilidade é de que gênero e emoção são inerentemente descorrelacionados. Apesar disso, é comum observarmos que certos gêneros musicais tendem a ser associados a emoções específicas - por exemplo, o axé é comumente associado a um sentimento alegre, enquanto a bossa nova é associada à melancolia. No entanto, podemos especular que o processo de produção dos gêneros e dos rótulos afetivos desta base de dados não tem relação com o significado sociocultural normalmente atribuído a eles.

Outro aspecto importante é que o espaço de features de áudio gerado pela VGG19, está produzindo vetores com aproximadamente10^7elementos, enquanto o espaço emocional usa 9 dimensões, logo a VGG19 está projetando trechos de áudio em um espaço com muitas dimensões, o que torna evidente a separação dos gêneros musicais. Apesar disso sabemos que a VGG19 foi treinada para classificação de imagens, usando assim técnicas de *transfer-learning*.

Isso pode nos indicar que os features da VGG19 não são relacionados a gêneros musicais específicos, mas sim estão explorando características do áudio que advém de técnicas típicas de mixagem e masterização em cada gênero. Isso significa que a alta precisão que alcançamos para classificação de gênero musical pode ser atribuída aos espaços com muitas dimensões e algumas coincidências na base de dados. Essas condições, evidentemente, desapareceriam quando partimos para um espaço de dimensão menor (9) para representar as faixas, levando assim a uma baixa precisão.

Este raciocínio mostra a importância das projeções PCA. Na Figura 4, por exemplo, podemos

ver que os gêneros musicais se interpolam no espaço vetorial gerado pelas avaliações na GEMS. Essa projeção indica que a nossa precisão não foi prejudicada pela falta de dados, mas sim devido a uma correlação inerente dos gêneros, algo que pode dificultar a separação dos mesmos.

Podemos ver no entanto pela Figura 4 que a interpolação de gêneros não se dá ao longo de todo espaço emotivo. Além disso, cada gênero aparenta ter medianas diferentes no espaço. Consequentemente especulamos que apesar de gêneros musicais estarem de alguma forma relacionados a valores afetivos, existe espaço suficiente em cada gênero musical para a existência de músicas com valores afetivos diferentes.

CONCLUSÕES:

Neste artigo nós comparamos a predição de gênero e emoções a partir de áudio utilizando técnicas de *transfer-learning*. Nós também investigamos a predição de gênero musical a partir de um espaço vetorial gerado por avaliações relativas à emoção.

Nós obtivemos uma alta precisão quando prevendo gênero musical de um espaço de features de áudio. Nós acreditamos que este resultado se dá pela alta dimensionalidade do espaço e enviesamentos de produção (técnicas de mixagem e masterização) na base de dados.

Além disso, nós identificamos que gêneros musicais são localizados em regiões com interpolação no espaço emocional gerado. Isso significa que, apesar do tamanho da base de dados, podem existir músicas de diferentes gêneros que estão relacionadas às mesmas emoções.

Por fim, nós destacamos que esse trabalho não teve a pretensão de atingir uma alta precisão de predição, mas sim uma análise do porquê os métodos de predição têm o desempenho que têm. Provemos esta explicação através da ideia de que gênero e emoções são algo, de fato, subjetivo.

Existem muitas coisas que podemos melhorar no que tange a nossa habilidade de explicar o comportamento de classificadores no domínio do áudio. Nós supomos que utilizando conceitos mais concretos tais quais tom ou tempo como suporte pode levar a resultados mais relacionados a conceitos musicológicos. Tal ideia impõem um vasto campo a ser explorado em trabalhos futuros.

BIBLIOGRAFIA:

[1] B. Ilari, “Música, comportamento social e relações interpessoais",Psicologia em Estudo, vol. 11, no. 1, pp. 191–198, Apr. 2006.

[2] A. Damasio and G. B. Carvalho, “The nature of feelings: evolutionary and neurobiological origins,”NatureReviews Neuroscience, vol. 14, no. 2, pp. 143–152, Feb.2013. [Online]. Available: http://www.nature.com/articles/nrn3403

[3] J. C. Lena,Banding Together: How Communities CreateGenres in Popular Music.Princeton University Press,2012.

[4] A. Aljanaki, F. Wiering, and R. Veltkamp, “Collecting an-notations for induced musical emotion via online game with a purpose emotify,” UU BETA ICS Department In-formatica, Tech. Rep., 01 2014.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[6] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,”IEEE Transactions on Speech and AudioProcessing, vol. 10, no. 5, pp. 293–302, 2002.

[7] Li and M. Ogihara, “Content-based music similarity search and emotion detection,” in 2004 IEEE InternationalConference on Acoustics, Speech, and Signal Processing,vol. 5, 2004, pp. V–705.