UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Angelo Cesar Mendes da Silva

Um Modelo para Classificação de Músicas Voltado à Plataforma de Streaming Utilizando Aprendizado de Métricas e Predição Estruturada

Juiz de Fora

2019

Angelo Cesar Mendes da Silva

Um Modelo para Classificação de Músicas Voltado à Plataforma de Streaming Utilizando Aprendizado de Métricas e Predição Estruturada

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora, na área de concentração em Inteligência Computacional e Otimização, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Raul Fonseca Neto

Juiz de Fora

2019

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Cesar Mendes da Silva, Angelo.
Um Modelo para Classificação de Músicas Voltado à Plataforma de Streaming Utilizando Aprendizado de Métricas e Predição Estruturada / Angelo Cesar Mendes da Silva. – 2019.
82 f. : il.

Orientador: Raul Fonseca Neto
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2019.

1. Aprendizado de métricas. 2. Similaridade de músicas. 3. Predição estruturada. I. Fonseca Neto, Raul, orient. II. Título.

AGRADECIMENTOS

Meus agradecimentos iniciam-se para minha mãe que, mesmo como inúmeras dificuldades, sempre priorizou e incentivou minha educação. Agradeço também aos meus avós, que assistem a conclusão desta etapa de um lugar privilegiado, Dinda, tias, tios e amigos a todos os ensinamentos que me formaram e me preparam para seguir pelo mundo. Um agradecimento especial ao meu orientador, Raul Fonseca Neto, por toda orientação, paciência e conhecimento repassado que me ajudou muito durante todo o mestrado e que me incentiva seguir aprendendo.

Foram tantos amigos e momentos que passei nesta fase que precisaria de outra dissertação para agradecer a todos, porém alguns devem ser destacados: as pessoas do ICE que sempre liberaram entrada e saída nos finais de semana e nas várias madrugadas para rodar testes, preparando seminários...; os amigos dos laboratórios, pelas discussões de ideias e companhia nos estudos; e as companhias nas sextas-santas, lanches e, principalmente, CAFÉS! Parafraseando Drummond, jamais me esquecerei destes acontecimentos nas minhas rotinas tão fadigadas.

Finalizando, um agradecimento à universidade, pois é uma enorme satisfação estudar em uma universidade como a UFJF que exerce um trabalho incrível de representatividade em diversas causas, não se omite e resiste aos incontáveis retrocessos políticos e sociais que estamos vivendo.

“Tudo seria fácil se não fossem as dificuldades.”
Barão de Itararé

RESUMO

No cenário em que uma tomada de decisão pode mudar o comportamento das próximas ações, um algoritmo requer muito poder adaptativo e cognitivo. Por exemplo, pode-se destacar a escolha de uma rota para ir ao trabalho, a compra de um produto via *e-commerce*, o investimento no mercado financeiro e a seleção de uma *playlist* usando em plataformas de streaming. O desenvolvimento de modelos para aprender semelhança entre músicas e extração de características de arquivos de mídia de áudio é uma tarefa cada vez mais importante para a indústria de entretenimento devido às dificuldades de obter informações consistentes de metadados. Esse trabalho propõe um novo modelo de classificação musical baseado em aprendizagem de métricas, extração de características de arquivos de áudio MP3 e redução de dimensionalidade. O processo de aprendizagem de métrica considera o aprendizado de um conjunto de distâncias parametrizadas empregando uma abordagem de predição estruturada a partir de um conjunto de músicas distribuídas sobre vários gêneros musicais. O principal objetivo desse trabalho é possibilitar o aprendizado de uma métrica personalizada para cada cliente que foi atingido por meio de duas abordagens, offline e online. Para a solução offline, atesta-se a validade do modelo realizando um conjunto de experimentos e comparando os resultados de treinamento e teste com os algoritmos de linha de base, como k-means e SVM. Os experimentos mostraram resultados promissores e incentivaram o desenvolvimento de uma versão online do modelo de aprendizagem. Com isso, foi feito a solução online que teve como objetivo o aprendizado em tempo real. Por meio de um conjunto de experimentos, fez-se uma avaliação e comparou-se aos resultados obtidos com a versão offline. Os experimentos mostram que a versão online converge em termos de precisão para a solução offline ideal, minimizando o *regret* médio. O desempenho do modelo de aprendizado possibilita a extensão do trabalho à diversas aplicações além da classificação de músicas.

Palavras-chave: Aprendizado de métricas. Similaridade de músicas. Predição estruturada.

ABSTRACT

In the scenario where decision making can change the behavior of next actions, an algorithm requires a lot of adaptive and cognitive power. For example, we can highlight choosing a route to go to work, buying a product via e-commerce, investing in the financial market, and selecting a playlist using a streaming platform. Developing models for learning similarity between pieces of music and extracting features from audio media files is an increasingly important task for the entertainment industry due to the difficulties of obtaining consistent metadata information. This paper proposes a new model of music classification based on metric learning, feature extraction from MP3 audio files and dimensionality reduction. The metric learning process considers the learning of a set of parameterized distances employing a structured prediction approach from a set of pieces of music distributed over various musical genres. The main objective of this paper is to enable the learning of a custom metric for each client that was reached through two approaches, offline and online. For the offline solution, we attest to the model's validity by conducting a set of experiments and comparing training and test results with baseline algorithms such as K-means and SVM. The experiments showed promising results and encouraged the development of an online version of the learning model. Thereby, we made the online solution that aimed at real-time learning. Through a set of experiments, we made an evaluation and compared to results obtained with the offline version. Experiments show that the online version accuracy converges to the ideal offline solution, minimizing the average regret. The performance of the learning model makes it possible to extend work to a variety of applications beyond the music classification.

Key-words: Metric Learning. Music Similarity. Structured Prediction.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

A música está presente no comportamento da sociedade desde o início de sua formação, destacando os aspectos sobrevivência, cultural, religioso e lazer. O impacto da música na construção sociocultural de vários povos vai além dos relatos em que imitações sonoras eram utilizadas pelo homem da pré-história para atrair suas presas durante uma caça. O continente africano, generalizando em virtude da grande diversidade de povos, possui uma musicalidade bem atrelada à presença de instrumentos de percussão e uma grande variedade de tambores de todos os tipos e tamanhos. Desde cedo, as crianças começam a tocar e cantar e aprendem a inserir a música em seu cotidiano com viés religioso como forma de comunicação com seus ancestrais e com seus deuses, para celebração de casamentos e nascimentos, para curar doenças e para acompanhar o trabalho. No Extremo Oriente, grande parte de sua música era escrita para danças da corte ou para o teatro. Também há músicas de cunho religioso, com instrumentos de percussão usados para aplacar os deuses e afastar os demônios. As ilhas espalhadas pelo Oceano Pacífico possuem uma tradição musical diretamente ligada ao mar, servindo para pedir aos deuses proteção durante a navegação. No passado, os povos que viviam nessa região viajavam de ilha para ilha, levando consigo seus instrumentos como tambores, flautas e apitos, que tiveram importante contribuição durante o povoamento de diversas ilhas. Devido à imensa concentração de povos indígenas no continente americano, os elementos musicais encontrados são comuns em grande parte do continente e em geral a música era cantada e expressava a crença desses povos nos deuses relacionados à natureza e também para suas danças. No Brasil, alguns mitos narram que a música teria sido um presente dos deuses, entristecidos com o silêncio no mundo dos humanos.

Com o passar do tempo, a música ainda possui seu papel de formação sociocultural ao oferecer mais opções para pessoas serem inseridas na sociedade e como elemento de caracterização de uma nação. São inúmeras bandas surgindo, festivais regionais musicais acontecendo, aulas de músicas como instrumento de inserção no meio artístico. Enfim, a música permanece com forte atuação na formação e caracterização de uma nação. Alguns exemplos na história recente, no cenário musical brasileiro, é o trabalho que o Rap está fazendo com suas letras de superação e críticas sociais que mostram a realidade de vários brasileiros fazendo com que vários cantores vão surgindo, ganhando espaço na mídia informativa e servindo como referência para outras pessoas que estão no início de sua carreira. Esse cenário foi visto no início dos anos 2000 com o Sertanejo, em 1980 com o Rock que dominava o *mainstream* e inspirava a formação de banda espalhadas pelo país, na década de 1970 a MPB conquistava o espaço das rádios após o domínio de Bossa Nova durante a década anterior. Vários outros movimentos musicais com enorme importância à nível regional e nacional, em alguns casos, também contribuíram com a formação histórica e cultural brasileira.

Obviamente que a música possui um papel de destaque, porém vários fatores relacionados compuseram o período de maior sucesso de cada um dos gêneros comentados e um dos principais segue tendo forte atuação também nos dias de hoje, os meios de comunicação. Durante a época em que o rádio era o principal e mais popular meio de comunicação o consumo musical das pessoas era construído e influenciado diretamente por interesses comerciais das estações de rádio. Com a evolução do rádio para a televisão, surgiram as imagens que personificaram as vozes que ilustravam a imaginação das pessoas e além da música a questão estética também influenciava no conteúdo que chegara aos consumidores. Assim como no rádio, a televisão também tornou-se um meio de comunicação muito popular, atualmente mais que o rádio, porém igualmente moldados por interesses comerciais para propagação de conteúdo, principalmente musical. Assim, com a popularização da internet como meio de comunicação inerente ao surgimento de redes sociais e plataformas para divulgação de mídias digitais, os meios de divulgação de conteúdo ficaram muito mais democráticos com a possibilidade de qualquer pessoa mostrar seu trabalho e ser conhecida por qualquer outra pessoa com auxílio de alguns mecanismos de busca e recomendação.

O surgimento de plataformas como YouTube e Spotify está relacionado à essa democratização de acesso à informação que tornou o consumidor o principal influenciador na disseminação de conteúdo que chegará até ele. Modelos de recomendação são submetidos a constante aprimoramento com intuito de direcionar conteúdo que irá ao encontro do gosto do consumidor baseado, principalmente, em seu histórico de buscas. Pela primeira vez, durante o processo de evolução dos meios de comunicação, o perfil de consumo do cliente é quem determinará o que lhe será apresentado. Ele que decidirá se algo te atrai ou não, e isso é importantíssimo para quem vender algum produto, pois devido à grande concorrência atrelada às opções de compras que existem atualmente, reter um cliente se tornou uma tarefa supervalorizada e de enorme valia para o mercado.

Ser capaz de definir as preferências de um cliente cria uma oportunidade para oferecer produtos de modo personalizado e, consequentemente, o poder para influenciar diretamente no processo de compra de um produto, ou serviço. É um grande desafio que profissionais de marketing e publicidade enfrentam, pois suas campanhas não podem se tratar apenas de apresentar um produto ao público, e sim apresentar algo que vá ao encontro da necessidade do cliente e faça-o adquirir o produto. No milionário mercado da música, especialmente em plataformas de streaming, conhecer o perfil de consumo do cliente é crucial para manutenção da fidelidade e conquistar novos clientes. Entre as táticas para retenção dos clientes usadas por estas plataformas, está a recomendação de músicas de forma unitária ou em organizadas em playlists.

A composição do perfil de consumo musical pode ser influenciada por preferências de gêneros, estilos, artistas, tradições regionais, presença de instrumentos entre outras

características. É muito comum de ocorrer divergências entre pessoas sobre a inserção de uma música em um determinado gênero ou estilo, uma música regravada por outro artista pode agradar à algumas pessoas que não gostaram da versão original, porém também pode gerar rejeição em quem prefira a primeira versão. A grande variedade de características que carecem de interpretação para serem utilizadas para rotular uma música torna esta tarefa complexa devido à subjetividade na avaliação das pessoas, causando um impacto negativo diretamente ao construir um perfil de consumo inconsistente.

1.1 MOTIVAÇÃO

A divisão em gêneros também é contestada assim como as definições de música porque cada composição ou execução pode se enquadrar em mais de um gênero ou estilo e muitos consideram que esta é uma forma artificial de classificação que não respeita a diversidade e a pluralidade musical. Há abordagens para classificação de músicas que estão inseridas em um contexto onde especialistas definem *tags* que serão adicionadas às músicas de forma a padronizar o processo de rotulação. A escolha de especialistas acarreta na diminuição da inconsistência das categorias que uma música pode ser inserida, porém a subjetividade das avaliações ainda está presente.

Uma outra proposta largamente adotada para classificação de música, são os conhecidos filtros colaborativos. Filtragem colaborativa é um método popular para aplicações de recomendações multimídia no qual, os dados (sons, artistas, livros, filmes...) são construídos e comparados em termos de pessoas que os usam. Eles são semelhantes a um banco de dados compostos por avaliações de forma implícita coletando dados automaticamente, ou explícita ao apresentar um questionário para o usuário responder baseando-se em categorias predefinidas. O grande volume de avaliações junto às categorias permite reduzir ainda mais a subjetividade e possivelmente construir um modelo de classificação mais consistente o que auxilia na definição do perfil de consumo e consequentemente beneficia uma recomendação.

Dois graves problemas impactam negativamente sobre a filtragem colaborativa: partida fria e cauda longa. No problema de partida fria, toda música recém-lançada não possui nenhuma avaliação, assim não será classificada como similar a nenhuma outra e, consequentemente, não será recomendada para ninguém. O problema cauda longa é semelhante ao princípio de Pareto, em que poucas músicas são avaliadas por um número grande de pessoas e muitas músicas são avaliadas por um número pequeno de pessoas. Esta desproporcionalidade no número de avaliações causa uma esparsidade e torna a métrica que avalia a similaridade inconsistente, e também uma preocupação com a escalabilidade devido ao grande número de avaliações que podem existir, pensando nos milhões de usuários de um plataforma de música.

Como solução aos problemas anteriores, pode-se adotar uma proposta de classifica-

ção capaz de “escutar” o conteúdo de um arquivo de áudio e calcular a similaridade a partir de descritores semânticos. São inúmeras características que consegue-se extrair e, enfim, propor um cenário consistente para associar músicas às classes. O problema em trabalhar com informações acústicas está na alta dimensionalidade do vetor de características que é construído. Manipular um arquivo de áudio com 30 segundos no domínio do tempo ou, com a aplicação de uma *Fast Fourier Transform* (FFT), resulta em um vetor com milhares de posições e essa enorme dimensionalidade é um complicador para avaliação da similaridade.

O modo como consome-se música está diretamente relacionado com a evolução dos meios de comunicação. Atualmente, a aquisição de músicas contidas em mídias de armazenamento como fitas cassete, discos de vinil e CD foi superada pelo o acesso dinâmico às músicas disponibilizadas em plataformas conectadas à internet. Essa dinamicidade de acesso transformou a forma de avaliar o perfil de consumo dos clientes, pois agora torna-se possível acompanhar um cliente em tempo real, pode-se saber o que ele está ouvindo em qualquer momento, avaliar seu comportamento durante um intervalo de tempo e, possivelmente, prever o que ele gostaria de ouvir por meio de uma recomendação. Nota-se que o mercado possui as informações do cliente e potencial para ofertar exatamente o que está em seu interesse.

Visando um mercado em que as principais plataformas de streaming Spotify, Apple Music e Deezer possuem aproximadamente 207 milhões, 75 milhões e 14 milhões de usuários e 30 milhões, 30 milhões e 56 milhões de músicas respectivamente, um modelo de aprendizado deve ser capaz de acompanhar a volatilidade do consumo dos clientes, ele precisa trabalhar com as “infinitas” músicas existentes, disponibilizadas via streaming, com suas variações de ordem de reprodução que surgem devido aos recursos *on demand* que permitem ao cliente decidir o que ouvir em sua própria ordem. O modelo deve ter o cliente como ator principal do sistema, e aprender com suas mudanças de comportamento. Mais prioritário que rótulos predefinidos de gêneros, artistas ou qualquer outro metadado, são as decisões do cliente que são passíveis de mudanças em qualquer instante. Entender este processo é ir ao encontro de toda informação disponibilizada, extrair o conhecimento necessário para conquistar o cliente e oferecer a melhor experiência do produto a ele.

Em muitas tarefas de aprendizado de máquina, bom desempenho depende de uma definição precisa de similaridade entre observações. O cálculo da distância euclidiana entre observações é um exemplo que fornece uma métrica simples, matematicamente conveniente, e que serve como base para o surgimento de algoritmos para aprendizado de métricas de distância com cenário supervisionado. Em geral, os algoritmos de aprendizado de métricas seguem o mesmo princípio: os bons vizinhos de uma observação devem estar mais próximos a ela que, por conseguinte, estará mais distante dos vizinhos ruins. A definição exata de perto e longe, bons e ruins vizinhos, varia de acordo com as configurações do problema,

podendo utilizar alguma combinação baseada na proximidade e no rótulo da observação.

Como motivação ao trabalho, viu-se a possibilidade de relacionar o campo de aprendizado de máquina em que a métrica representa o perfil individualizado de consumo e o aprendizado é mantido pelas correções, responsáveis pela definição a distância e vizinhança entre as observações (músicas), que o cliente realiza ao indicar a sua satisfação, ou não, com a classificação de uma determinada música. Junto ao aprendizado, almeja-se uma maior precisão devido à categorização com descritores semânticos obtidos através de informações acústicas que deverão impactar positivamente na acurácia de modelos de classificação em relação às informações subjetivas. Por o trabalho ser de grande valia ao mercado, estima-se que seu sucesso resulte em um modelo de aprendizado com potencial para ser replicado e adaptado a produtos comerciais existentes.

1.2 OBJETIVOS DO TRABALHO

O problema de tomar decisões online sobre o presente quando sua perda não é conhecida é uma área importante da pesquisa em aprendizado de máquina. Por exemplo, pode-se destacar a escolha de uma rota para ir ao trabalho, a compra de um produto via *e-commerce*, o investimento em riqueza no mercado financeiro e a seleção de uma lista de reprodução usando uma plataforma de streaming. Neste cenário online, o algoritmo de aprendizado deve ser capaz de minimizar o *regret* preocupado com a diferença entre sua perda acumulada e uma solução offline ótima. O desenvolvimento de modelos para aprender a similaridade musical a partir de arquivos de mídia de áudio é uma tarefa cada vez mais importante para a indústria de entretenimento. No entanto, a maioria desses algoritmos é desenvolvida em uma configuração offline, em que as informações passadas completas são usadas em modo *batch*.

O modelo proposto é constituído pela construção do vetor característico, a partir de extração de informações acústicas de um arquivo de áudio em formato MP3, redução de sua dimensionalidade, realização da classificação supervisionada pelo usuário, o que torna o modelo capaz de mensurar preferência musical do usuário baseada no estudo e avaliação da similaridade entre músicas utilizando uma abordagem de aprendizagem por métricas. Foi construído um modelo offline, com as observações processadas em batch, comparação com um classificador baseado em máquinas de vetores suporte (SVM), para validação do modelo, e uma versão online, com as observações processadas como um modelo streaming, semelhante a uma abordagem comercial.

A escolha da característica que representa a música é de extrema importância, vista a dificuldade existente ao utilizar metadados ou categorias que acarretem em informações subjetivas. Há inúmeras características que podem ser usadas para representar uma música, as características de baixo e alto níveis são mais utilizadas na literatura e são relacionadas como base para comparações. Características de baixo nível são valores

numéricos que representam o conteúdo de um sinal de acordo com tipos de domínio, como temporal, espectral e perceptivo. Já as características de alto nível estão relacionadas à capacidade de percepção e interpretação humana sobre uma determinada combinação sonora e como é feita sua definição, timbre e ritmo são dois exemplos. Ao manipular um arquivo MP3 com sinal sonoro, sabe-se que o vetor característico obtido possui alta dimensionalidade, com isso, torna-se importante realizar uma análise dos componentes principais para representação de uma música usando dimensionalidade inferior e avaliar seu impacto que esta redução reflete no processo de classificação. Com o processo de construção do vetor característico definido, pode-se realizar a implementação do modelo para classificação online integrando as etapas de extração de características e redução de dimensionalidade.

Considerou-se para cada amostra o uso de trinta segundos ao longo de um segmento de áudio. Extraiu-se o vetor de característica do áudio a partir do segmento musical usando *Mel-Frequency Cepstral Coefficient* (MFCC) (Loughran *et al.*, 2008) com objetivo de capturar as informações do sinal sonoro. Devido à grande dimensionalidade, fez-se um estudo para sua redução utilizando Análise de Componentes Principais (PCA) ao invés de uma abordagem de Seleção de Características. Esta representação de características utilizando somente MFCC é mais reduzida e homogênea em relação a características utilizadas em outros trabalhos. Por exemplo, em Wolff e Weyde (2014) é usado um conjunto características de baixo nível (*chroma* e timbre) e outro de alto nível (*loudness*, *beat* e *tatum*), McKinney e Breebaart (2003) fizeram uma estudo comparativo envolvendo quatro grupos de características sonoras (baixo nível, MFCC, psicoacústicas e modelo auditivo) e em Bergstra *et al.* (2006) utilizaram diferentes métodos de processamento de sinal para extrair e incluir um conjunto de características como MFCC, FFT, *Real Cepstral Coefficients* (RCEPS) e *Zero-crossing Rate* (ZCR) na composição de um vetor característico.

O problema de aprendizado de métrica é resolvido como um problema de otimização a considerando a minimização de um conjunto de distancias parametrizadas mensuradas sobre pares de amostras e sujeitas às restrições de desigualdade triangular. Além disso, os valores de distância devem ser simétricos e não negativos. Neste contexto, diferentes soluções são verificadas, como o aprendizado de uma matriz de parâmetros completa ou uma matriz diagonal, resultando em um vetor de parâmetros. Neste último caso, aprende-se uma métrica que pesa as diferentes dimensões do espaço do problema. Essa abordagem pode ser considerada como o uso de uma perda contrastante (Hadsell *et al.*, 2006) que tenta minimizar uma distância parametrizada entre amostras semelhantes e maximizar entre aquelas diferentes.

O método proposto para aprender a similaridade musical tem uma relação direta com a solução ao Problema de Predição Estruturada proposta por Coelho *et al.* (2017).

Baseia-se no cumprimento de um conjunto de restrições que atestam a pertinência de cada amostra musical em relação ao seu centroide de gênero em relação às alternativas. Essas restrições representam a condição de que a distância parametrizada entre uma amostra e seu respectivo centroide deve ser menor que a de qualquer outro centroide do conjunto de treinamento. O trabalho desenvolvido por Wolff e Weyde (2014) também utiliza uma abordagem análoga para aprender a semelhança musical, mas, neste caso, os autores consideram a aprendizagem de uma métrica a distância a partir de comparações relativas (Schultz e Joachims, 2004) envolvendo para cada restrição uma tripla de amostras de áudio.

Nos experimentos, usou-se dois conjuntos de dados, o primeiro é o conjunto público, GTZAN, que consiste em 1000 segmentos de áudio com 30 segundos cada, igualmente divididos em 10 gêneros; o segundo é um conjunto artificial denominado MUSIC que também contém 1000 segmentos de áudio com 30 segundos cada, mas com 5 somente gêneros. Ambos os conjuntos de dados construíram uma variedade de segmentos de áudio contendo 15 segundos. Além disso, o conjunto de dados MUSIC gerou mais duas variações com o número de segmentos de áudio igual a 250 e 500 segmentos contendo 30 segundos e 5 gêneros com o mesmo número de músicas.

Para versão offline do modelo, realiza-se uma avaliação abrangente do desempenho do modelo, realizando um conjunto de experimentos de treinamento e teste. Compara-se os resultados obtidos pelo modelo com um classificador multiclasse baseado em SVM treinado com uma estratégia de um contra todos. Estuda-se também a influência do aprendizado de métrica, com variações no comprimento do segmento musical, na dimensionalidade das observações e no tamanho do conjunto de treinamento e seus respectivos impactos no sucesso da generalização.

Já para avaliar versão online, inicialmente realizou-se a classificação da música em modo *batch*, produzindo uma solução ótima offline inalterada. Em seguida, realizou-se os experimentos utilizando os conjuntos de dados MUSIC e GTZAN em suas versões contendo 1000 músicas e 30 segundos, os resultados obtidos online foram comparados considerando os resultados obtidos a partir da versão offline. Uma análise analítica da convergência foi feita relatando o *regret* médio relacionado à precisão definida pelos erros médios. Além disso, analisou-se as convergências dos vetores da métrica e dos centroides para os respectivos vetores da solução offline.

1.3 ORGANIZAÇÃO

Além deste capítulo introdutório, apresenta-se no capítulo seguinte a fundação teórica para abordagem que contém os trabalhos relacionados ao contexto de aprendizagem de músicas por meio de vários tipos de características, junto aos seus problemas. Também introduz-se o conceito de aprendizagem de métricas e estratégias adotadas para solucionar

problemas de classificação de músicas. No Capítulo 3, é feita uma descrição sobre tipos de características acústicas de músicas, um detalhamento da característica que foi adotada neste trabalho junto ao processo de construção do vetor característico. O Capítulo 4 traz a formulação matemática para o problema de aprendizagem de métricas de similaridade e a descrição das duas soluções que foram propostas. O Capítulo 5 apresenta os experimentos e resultados obtidos para cada solução. Encerra-se no Capítulo 6 com as conclusões e possíveis desdobramentos para esse trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Muitas áreas de pesquisa na recuperação de informações musicais envolvem tarefas de classificação, como classificação de gêneros, geração de playlists, transcrição de áudio para textos e etc. A informação fundamental que suporta a classificação musical inclui conjunto de dados musicais (amostras), gravações de áudio, partituras e dados (playlists individuais, resenhas de álbuns, estatísticas de comerciais) que também podem incluir metadados sobre as músicas como identificação do artista, título, compositor, artista, gênero, data, entre outros. A variedade de dados para representação de músicas foi reportada por Gupta (2014), que também descreveu a dificuldade transformá-los em informação e a capacidade cognitiva de algoritmos de aprendizado de máquina em associar um conjunto de características na classificação de músicas.

A classificação automática da música é uma tarefa difícil devido à falta de dados confiáveis que podem ser utilizados com base, segundo Corrêa e Rodrigues (2016). Esta confiabilidade está relacionada à falta de padronização de rótulo a dados semelhantes, o que é crucial para executar adequadamente os métodos de reconhecimento de padrões. Bases de dados musicais aumentaram consideravelmente em número e tamanho nos últimos anos, o que incentivou ao desenvolvimento de ferramentas com maior precisão para extração de conhecimento musical. Os crescentes avanços nos algoritmos de aprendizado de máquina foram motivados justamente devido à abundância de músicas disponíveis em formatos de áudio digital, a crescente qualidade e acessibilidade dos dados musicais online e a disponibilidade de bibliotecas para a extração de características musicais.

Seguindo os ideias gerais de um problema de classificação, a questão mais importante que aborda a semelhança de música é a classificação dos segmentos de acordo com as categorias descritas por seu conteúdo musical (Corrêa e Rodrigues, 2016). Neste caso, espera-se que cada categoria seja composta por músicas com características em comum, enquanto músicas de categorias diferentes devem ser menos semelhantes. A classificação automática de gêneros musicais desempenha um papel fundamental no contexto de classificação, indexação e recuperação de músicas (McFee *et al.*, 2010), permitindo que sites e plataformas de música de dispositivos gerenciem e rotulem seu conteúdo musical.

A modelagem computacional da similaridade musical é cada vez mais importante para personalização e otimização na classificação, recuperação e pesquisa de informações musicais na percepção e cognição de músicas. Wolff e Weyde (2014) definem um método para avaliação da similaridade como o maior responsável pelo bom desempenho em problemas de recuperação e recomendação de música. Ainda em Wolff e Weyde (2014), alinhada ao aumento dos dados musicais que motivaram o avanço das pesquisas em aprendizado, estão a variedade de mídias de armazenamento, físicas ou não, de músicas que se tornaram mais baratas e populares encorajando ainda mais o progresso da área.

A dificuldade em medir a similaridade entre músicas é dificultada quando é feita a partir de avaliações por pessoas. Slaney *et al.* (2008) apresentam um cenário em que duas músicas que são semelhantes a dois amantes do jazz, podem ser muito diferentes de alguém que não ouve jazz e conclui que este é inerentemente um problema mal condicionado. Os usuários de plataformas que possuem recursos de recomendação de música têm sido frequentemente negligenciados devido aos critérios de percepção e recuperação de similaridade musical.

Além do desapontamento dos usuários que não se enquadram nesse pressuposto, as abordagens de recuperação fixa também podem impor uma influência cultural, especialmente quando os fatores envolvidos na comparação não são transparentes. Por outro lado, a recuperação adaptada pelo usuário tem o potencial de fornecer resultados de pesquisa personalizados mais adequados às necessidades do usuário do que uma sugestão padronizada (Wolff e Weyde, 2014).

Vlegels e Lievens (2017) relatam a tentativa de descobrir grupos que representam a relação de similaridade entre músicas de um conjunto de artistas e de informações de perfil de usuários obtidas de diferentes locais e distribuições de gênero e idade. Essas informações são extraídas de respostas a perguntas sobre aspectos sócio-demográficos e seu comportamento cultural em uma ampla gama de domínios como artes, cultura cotidiana, atividades de lazer, esporte e recreação. O objetivo principal era aprender a preferência musical do cliente usando a relação entre seu perfil e as informações sobre seus artistas favoritos. Para isso, os autores exploram o conhecimento musical fornecido por uma rede social para construir uma relação entre as pessoas e os artistas. Os resultados mostraram que pesquisas usando características baseadas somente em informações culturais e preferências de gêneros são inadequadas ou insuficientes para uma classificação eficiente.

Diversas abordagens, que analisam a similaridade musical em problemas de aprendizado de máquina, possuem uma deficiência em comum: ignoram o feedback do usuário. Este problema acaba tornando o sistema responsável por definir um consenso sobre a percepção da similaridade musical (Barrington *et al.*, 2009). Um exemplo destacado por McFee *et al.* (2010) mostra que uma banda sempre cantará músicas de apenas um gênero, ou músicas de uma região sempre serão inseridas no mesmo grupo, devido à influência cultural, e outros fatores são pressupostos e não são transparentes ao usuário. Este modelo tende a não ser eficiente em um cenário no qual o consumo musical de um usuário representa uma exceção existente dentro de um grupo rotulado por um consenso.

Para contornar estas limitações existem algumas alternativas para construção de um conjunto de dados para representação de músicas. Para realização desta tarefa, muitas das abordagens comerciais utilizam a filtragem colaborativa rotulando suas músicas a partir de avaliações feitas por seus usuários. Este modelo é uma abordagem muito utilizada também em propostas de recomendação online de músicas (McFee *et al.*, 2010).

2.1 FILTRAGEM COLABORATIVA

Uma técnica comumente usada para avaliar a similaridade entre músicas com informações de feedback do usuário é a filtragem colaborativa (Gupta, 2014). Esta técnica visa coletar conhecimento a partir de muitas avaliações realizadas por usuários sobre conteúdos multimídia disponibilizados por plataformas digitais. Sua adoção é justificada devido à possibilidade de obter muitas avaliações sobre um dado específico e buscar pela melhor forma de apresentar este dado que atenda a todas as avaliações.

Este conjunto de avaliações abre campo para aplicações de informações extraídas em problemas de classificação e, principalmente, em problemas de recomendação visto que é fácil compreender o perfil de consumo de um usuário a partir do conteúdo que ele avalia. A maior dificuldade em avaliar a similaridade musical baseada em informações providas por filtros colaborativos é garantir a existência de um grande número de músicas juntamente a um grande número de avaliações e que estas sejam coerentes.

Quando o conjunto de dados, obtido a partir dos filtros, é constituído por informações consistentes e confiáveis, o mesmo tende a ter bom desempenho em tarefas de recomendação (McFee *et al.*, 2010). Entretanto, o problema da cauda longa, quando existe um número ínfimo de músicas com muitas avaliações e a maioria delas com poucas avaliações (a representação gráfica possui uma imagem semelhante a um pequeno animal com cauda enorme), é muito recorrente e impacta negativamente a aplicabilidade em diversos cenários. McFee *et al.* (2010) propõem um método para solucionar este problema que combina informações do artista, com sua música, extraídas a partir do processo de filtragem, desta maneira, desde que o artista já possua um certo número de avaliações, é possui construir um novo conjunto de dados para representar as músicas ao associá-las a seu artista.

Esta proposta parece melhorar o desempenho em problemas de recomendação, porém só muda a perspectiva na inconsistência do conjunto dos dados. A desproporcionalidade entre o número de músicas e o número de artistas também irá interferir no processo de rotulação e, consequentemente, retornar ao problema da cauda longa.

É apresentado por P. W. Ellis *et al.* (2002) um estudo sobre medidas para avaliar a similaridades a partir de vários conjuntos de dados formados através de questionários respondidos por pessoas e concluíram que, no geral, mensurar a similaridade entre artistas a partir de conjuntos de dados com esse tipo de origem era algo subjetivo pois existem bastante variáveis entre os usuários e suas avaliações. Sua conclusão acaba lançando dúvidas sobre conceito de uma única "verdade fundamental" em que pode ser usada como dados de treinamento ou referência.

Barrington *et al.* (2009) demonstram que o processo de filtragem colaborativa é capaz de oferecer dados consistentes à avaliação de similaridades entre músicas. Os

experimentos desenvolvidos tiveram sucesso na avaliação da similaridade entre os dados e obtiveram êxito ao resolver um problema de recomendação musical a partir de dados semelhantes. Entretanto o modelo proposto falhou em não recomendar dados desconhecidos, o bem conhecido problema do partida fria e, como não são recomendadas, também não serão avaliadas pelo usuário juntando-se às músicas contidas na cauda longa do mercado da música.

Kim *et al.* (2009) propõem o uso de uma técnica para inserir avaliações, sintetizadas por um modelo, em novos dados ainda não avaliados por usuários. São exploradas algumas abordagens para avaliar a similaridade entre os dados e, posteriormente, gerar as avaliações que foram compostas por gêneros, descritores acústicos e tags sociais. Esta proposta poderia ser uma alternativa para solucionar os maiores problemas da filtragem colaborativa, porém a conclusão do trabalho foi que, mesmo sem ser uma técnica baseada em noções semânticas para definição da similaridade, o desempenho da filtragem colaborativa foi superior à solução proposta pelos autores.

Os experimentos propostos por Slaney *et al.* (2008) avaliam a eficácia das métricas de similaridade comparando-a com os vizinhos mais próximos das músicas em termos de nomes de artistas. Os autores discutem duas desvantagens associadas ao uso da identidade do artista como verdade da similaridade: em primeiro lugar, eles observam a grande variedade de estilos musicais que qualquer artista pode ter. Em segundo lugar, discute-se um desequilíbrio na distribuição das informações de filtragem colaborativa em seus dados em relação a artistas e álbuns. Os usuários podem ouvir e "curtir" todas as músicas de um artista porque sua playlist é baseada em artistas. O mesmo problema se aplica ao gênero musical ou a qualquer outra categoria tipicamente usada para organizar música. Esta conclusão vai ao encontro da análise vista em McFee *et al.* (2010).

Como visto nos trabalhos citados, o uso da técnica de filtragem colaborativa para composição de dados que representam músicas é uma prática constantemente adotada em tarefas de aprendizado de máquina. No entanto, foi visto nestes trabalhos que ainda há problemas na adoção desta técnica e que a maioria dos trabalhos convergia para eles, independentemente da abordagem utilizada. Herrada (2010) enumerou os principais problemas, junto às causas e consequências, como esparsidade devido a avaliações inexistentes na base; subjetividade já que usuários com comportamentos diferentes podem divergir em avaliações sobre os mesmos dados; e escalabilidade porque a complexidade para cálculo das avaliações de similaridades tendem a aumentar proporcionalmente em relação ao número de avaliações. Os problemas de partida fria e cauda longa também são analisados em Herrada (2010) destacando o impacto no desempenho ao tentar realizar em tarefas de recomendação.

A maior dificuldade em avaliar a similaridade musical baseada em informações provenientes de filtros colaborativos, de acordo com Slaney *et al.* (2008), é a existência

de um grande número de dados e ruídos incertos que levam a uma enorme incoerência no processo de avaliação e, consequentemente, no desempenho de qualquer aplicação. A confiabilidade dos dados e de seus rótulos é o fator crucial para o bom desempenho de modelos de aprendizado, as críticas neste requisito questionam a credibilidade da técnica e induz à busca por alternativas para extração de características dos dados.

Slaney e White (2007) discutem duas desvantagens associadas ao uso de atributos categóricos a partir de escolhas de usuários como base para avaliação da similaridade: em primeiro lugar, eles observaram a grande variedade de estilos musicais que qualquer artista pode ter. Em segundo lugar, discute-se um desequilíbrio na distribuição das informações de filtragem colaborativa em seus dados em relação a artistas e álbuns. Os usuários podem ouvir e "curtir" todas as músicas de um artista porque sua playlist é baseada em artistas e este mesmo problema se aplica ao gênero musical ou a qualquer outra categoria tipicamente usada para organizar música (Wolff e Weyde, 2014).

Com esse contexto subjetivo e inconsistente, uma alternativa para solucionar estes problemas é buscar por informações extraídas diretamente da música capazes de caracterizá-la de forma fidedigna. Preferencialmente sem interferência dos usuários e que o procedimento de extração seja aplicado uniformemente a todas as músicas existentes no conjunto de dados. Com estas informações, pode-se esperar uma caracterização com maior consistência e que seja permitido aprender a preferência do usuário de uma maneira objetiva ao analisar diretamente a composição estrutural da música. A uniformidade no processo de extração e construção do vetor de característica tende a melhorar o desempenho geral no processo de aprendizado, conforme Slaney *et al.* (2008).

2.2 CARACTERIZAÇÃO MUSICAL USANDO INFORMAÇÕES ACÚSTICAS

Os vetores característicos construídos a partir de informações acústicas possuem informações mais confiáveis devido ao processo de formação ser uniforme a todas as amostras. A comparação de música baseada em conteúdo do áudio pode incorporar a extração de características acústicas, psicoacústicas e de teoria musical derivadas de informações de áudio (Wolff e Weyde, 2014). A aplicabilidade de tais características e medidas de distância é altamente dependente do contexto da música, da plataforma de distribuição e do usuário. Entretanto, os modelos de aprendizado podem ajudar a garantir que o sistema seja adequado às necessidades dos usuários e às intenções do mercado.

Pensar em adotar características acústicas como forma de representação de músicas remete a questões relevantes que irão orientar como será realizado o processo de extração. Por exemplo, sabe-se que existem inúmeras características acústicas que podem ser extraídas, qual é a mais representativa? É possível fazer a combinação entre várias? O processo de extração é aplicado sobre toda música? Sobre parte dela? Qual parte? Estes questionamentos são extremamente relevantes para avaliar o custo computacional,

a melhoria no desempenho e real ganho em relações à outras técnicas de extração de características.

Supondo que deseja-se extrair características de uma música, com aproximadamente três minutos de duração, utilizando uma FFT para construir o vetor característico. A dimensionalidade deste vetor será de exatamente **3.969.530** componentes. Se utiliza-se a informação bruta como característica, com o som no domínio do tempo, tem a dimensionalidade igual a **180.024** componentes. Obviamente, devido ao tamanho, não são boas características para utilizar em problemas de aprendizado, porém destaca-se que trabalhar com a música em seu tamanho original também é extremamente complicado, mesmo adotando outras técnicas.

Extrair boas características está longe de ser uma tarefa fácil e envolve técnicas e teorias de processamento de sinais, psicoacústicas, percepção musical e análise de dados em geral. Existem estudos sobre modelos para efetuar uma busca sobre as partes mais representativas de uma música. Porém, Pampalket (2006) mostra que a informação extraída sobre trinta segundos de uma música seria suficiente para representá-la. Além disso, Xin *et al.* (2014) concluem que as principais características são expostas durante a primeira metade da música, e usualmente sua segunda metade é composta pela repetição de uma parte considerável da primeira.

O foco do trabalho apresentado por McKinney e Breebaart (2003) é sobre características para classificação de áudio e música. Os autores realizam uma comparação entre conjuntos de características mais comumente utilizados, propriedades de sinal de baixo nível e o MFCC, através de dois novos conjuntos de características construídos por eles. Avaliou-se o desempenho das características na classificação de um conjunto de arquivos de áudio genéricos e musicais existentes em gêneros populares. Mensurou-se também como a representação do comportamento temporal das características pode influenciar a classificação. Os dois conjuntos de características desenvolvidos pelos autores são baseados em modelos perceptivos à audição humana. Os resultados mostram que o comportamento temporal é importante para a classificação de música e áudio. Além disso, a classificação apresentou melhor desempenho, em média, quando baseou-se em características de modelos de percepção auditiva, e não em características padrão (sem a devida adaptação ao limiar auditivo).

McKinney e Breebaart (2003) fazem a caracterização das músicas através de uma análise acústica fornecida através de uma API web fornecida pelo site The Echo Nest[1]. A partir do envio de uma música para a API, é feita a análise que resulta em 18 características a serem usadas para representar as músicas. Esta análise tem a função de dividir a música em segmentos e organiza cada um com uma seção composta por segmentos com qualidades acústicas semelhantes. Esses segmentos são de 80ms até vários segundos de duração. Além

[1] http://the.echonest.com. Acessado em 29 de Junho de 2019.

de propriedades globais, para cada música é calculado o volume, o tempo de início e as outras medidas da variação nos segmentos.

Xin *et al.* (2014) propõem utilizar o MFCC como características de áudio. A escolha por esta característica deve-se à sua forma de modelagem do espectro sonoro que permite refletir o tom do áudio no grau desejado pelos autores. A partir da extração do MFCC é feito a classificação baseando-se na semelhança acústicas entre as músicas.

A abordagem de Velankar *et al.* (2015) é um modelo para avaliar a similaridade entre músicas baseando-se nas diferentes variações de melodia. Para os experimentos, não foi considerado o ritmo que está associado à música, pois é inerente à maioria das melodias e esta foi considerada como uma sequência de notas, inclusive com os parâmetros associados.

Samal *et al.* (2014) descrevem um estudo experimental e o desenvolvimento de um modelo para reconhecimento de voz. É apresentado um método eficiente, aplicado à área de segurança, para realizar o reconhecimento de voz utilizando MFCC como característica em sua representação. O método atende à identificação e autenticação das pessoas.

Influenciados pela ampla adoção do MFCC, Trang *et al.* (2015) realizam um estudo para medir o desempenho do método convencional de extração de características do MFCC em termos de precisão de reconhecimento e a velocidade para aprendizado durante o processo de treinamento do modelo para um sistema de reconhecimento de voz.

Loughran *et al.* (2008) visam examinar o uso de MFCC na classificação de instrumentos musicais. Amostras de piano, violino e flauta são analisadas para obter seus vetores característicos. O objetivo é tentar distinguir entre instrumentos musicais usando apenas MFCCs, além de fazer um estudo sobre a quantidade de coeficientes cepstrais são necessários e úteis para a classificação precisa do instrumento.

Estudos relacionados à música atingem diversas áreas de forma multidisciplinar. Vyas (2014) apresenta um método para identificação de sentimentos despertados em pessoas ao ouvir uma peça musical baseado em características acústicas. Três características foram combinadas para definir o rótulo (sentimento) da peça musical, são elas: MFCC, energia da estrutura e diferença de pico. As experiências foram realizadas em um banco de dados de segmentos de música de várias categorias. A precisão de acertos na identificação do sentimento encoraja a continuidade no desenvolvimento da abordagem.

Neste trabalho, considera-se para cada amostra o uso de um segmento de áudio com trinta segundos de duração. Em seguida, extrai-se um vetor de características do segmento musical usando MFCC (Loughran *et al.*, 2008) com o objetivo de capturar o sinal de áudio. Devido ao grande número de características extraídos, fez-se um estudo de redução de dimensionalidade usando a Análise de Componentes Principais (PCA) em vez de uma abordagem de seleção de características. A opção pela representação com

característica baseada apenas em recursos do MFCC é mais reduzida e homogênea em comparação com outros trabalhos. Por exemplo, Wolff e Weyde (2014) usam um conjunto de características de baixo nível (vetores de chroma e timbre) e de alto nível (loudness, beat e tatum médio e variância), McKinney e Breebaart (2003) que fizeram um estudo comparativo envolvendo quatro grupos de características de áudio (sinal de baixo nível, MFCC, modelo psicoacústico e auditivo) e Bergstra *et al.* (2006) extraem e incluem um conjunto de características de áudio de diferentes métodos de processamento de sinal de áudio, como MFCC, Fast Fourier Transform (FFT), *Real Cepstral Coefficients* (RCEPS) e de *Zero Crossing Rate* (ZCR).

McKinney e Breebaart (2003) realizou um estudo sobre o impacto que comportamentos temporais e estáticos de um conjunto de características podem ter sobre o desempenho de classificação de áudios gerais e gêneros musicais. A conclusão deste trabalho, uma constatação vista na maioria dos problemas de classificação, é que a variação da precisão no processo de classificação é fortemente condicionada pela composição do vetor de características, sem considerar o recurso computacional utilizado para construí-los. Dentre os conjuntos de características analisados, dois apresentaram bom desempenho ao classificador, Envelopes Temporais do Banco de Filtros Auditivo (AFTE) e Coeficientes mel-cepstrais (MFCC).

Motivado por pesquisas na área de recuperação de informação musical, Yen *et al.* (2014) realizam um estudo sobre a identificação das principais característica da voz de um artista, a partir de uma música, e recuperar outras relacionadas a ele. Para caracterizar a voz, os autores utilizam um espectrograma para encontram as regiões dominantes utilizando frequência junto às informações harmônicas. Bergstra *et al.* (2006) fazem um estudo sobre características acústicas para melhor representação musical em um problema de classificação do gênero da música e artistas. Seis características foram selecionadas, dentre elas destaca-se ZCR, MFCC e FFT.

O MFCC é uma técnica padrão de pré-processamento no processamento da fala. Originalmente, seu desenvolvimento visava o uso em problemas de reconhecimento automático de voz (Oppenheim, 1969), porém diversos estudos a utilizaram em experimentos de para recuperação de informação de música, classificação e muitas outras tarefas (Pampalket, 2006).

Tzanetakis e Cook (2002), relatam um estudo de análise de características em um processo de classificação musical usando o conjunto de dados GTZAN. Eles usaram como características a textura timbral, ritmo e frequência.

Li *et al.* (2003) apresentam um dos primeiros estudos para avaliar o desempenho de uma variedade de características acústicas no processo de classificação de gênero. Ele fornece um grande estudo comparativo de vários métodos de extração e classificação de características além de investigar o desempenho da classificação de vários classificadores

em conjuntos de características diferentes.

A alta complexidade para avaliar a similaridade de músicas, relatada por McFee *et al.* (2010), indica a necessidade de incorporar características acústicas, psicoacústicas ou teóricas derivadas do conteúdo contido no áudio para obter melhores resultados de classificação. Observe que, a partir da música de referência, pode-se usar uma consulta para retornar vários outros com características semelhantes, indicando novas preferências e também rotular amostras desconhecidas com base em métricas de similaridade (Pampalket, 2006; Slaney *et al.*, 2008; Wolff e Weyde, 2014). A conclusão apresentada por Wolff e Weyde (2014) evidencia a importância de cada característica de uma música conhecida e sua similaridade medida em relação a outras músicas é altamente dependente do contexto em que elas estão inseridas. Nesse sentido, percebe-se a importância que os modelos de aprendizagem têm quando ajudam a garantir que um sistema de recomendação seja adaptável à preferência de cada cliente.

2.3 APRENDIZADO DE MÉTRICAS

A busca por melhores meios para mensurar a semelhança, ou diferença, entre dados é onipresente principalmente em problemas de aprendizado de máquina, reconhecimento de padrões e mineração de dados. Entretanto, devido à variedade de métricas e formas de aplicação existentes, elaborar uma métrica para um problema específico é uma tarefa, geralmente, difícil. O desafio em obter métricas adaptáveis aos problemas, com aprendizado automático a partir de um conjunto dados qualquer, estimulou o avanço em diversos estudos sobre aprendizado de máquina e áreas correlacionadas nos últimos anos (Bellet *et al.*, 2013).

Toma-se como exemplo um problema em que há duas imagens disponíveis e tem-se como objetivo encontrar faces semelhantes com base na identidade, então deve-se escolher uma função de distância que enfatize as características apropriadas (cor do cabelo, cor da pele, proporções de distâncias entre pontos específicos, etc.). Entretanto, também pode-se ter uma aplicação em que almeja-se determinar a semelhança das imagens a partir da pose de alguns indivíduos e, portanto, exigindo que a função de distância capture a similaridade. Claramente, outras características são mais aplicáveis neste cenário, diferentes das utilizadas na busca por faces semelhantes.

Para lidar com múltiplas semelhanças ou métricas de distância, pode-se tentar determinar manualmente uma função de distância apropriada para cada aplicação, por meio da escolha de características mais representativas ao contexto explorado junto à combinação destas características. Kulis (2012) mostra que utilização de diferentes métricas em algoritmos de aprendizado de máquina pode mudar completamente os resultados de análises realizadas em bases de dados. Variar as maneiras de medir distâncias, ou similaridades, entre dados impacta diretamente na escala das informações obtidas sobre

estes dados e, como um efeito em cascata, influencia na tomada de decisões dependentes desta medida.

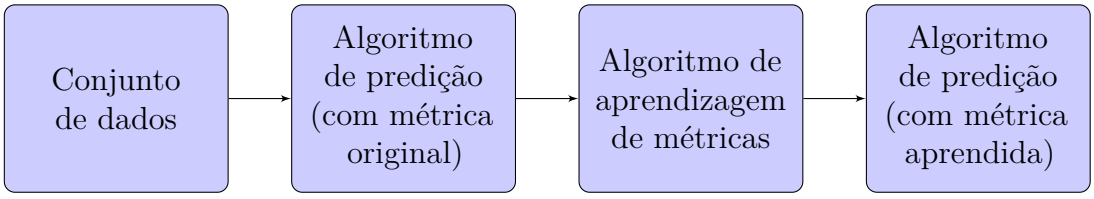
No contexto de aprendizado, o uso de métricas para avaliação de similaridade entre dados é adotado em várias abordagens clássicas, como na classificação tem-se o KNN de Cover e Hart (1967), um classificador que baseia-se uma métrica para identificar os vizinhos mais próximos de uma determinada instância; muitos algoritmos de agrupamento, como o proeminente k-means de Lloyd (1982), dependem de medições de distância entre o centroide do grupo e uma instância para definição de seu rótulo; na recuperação de informações, instâncias são geralmente classificadas de acordo com sua relevância para uma determinada consulta, sendo a relevância semelhante a um processo de ranqueamento e relacionada à similaridade entre as instâncias e a consulta.

Em geral, os algoritmos de aprendizagem de métrica seguem o mesmo princípio: os bons vizinhos de uma instância devem estar mais próximos do que seus vizinhos ruins. As definições exatas de bom e ruim variam entre as configurações do problema, mas normalmente derivam de alguma combinação de acordo com proximidade e rótulo (McFee e Lanckriet, 2010). Neste sentido, os algoritmos de aprendizagem de métricas são frequentemente avaliados através da precisão na predição dos rótulos em relação aos vizinhos mais próximos de uma determinada instância.

O desempenho destes métodos e, consequentemente, suas variações dependem da qualidade da métrica alinhada à capacidade em identificar como similar, ou dissimilar, os pares de instâncias que são, de fato, semanticamente próximas, ou distantes. Existem métricas de propósito geral (por exemplo, a distância euclidiana ou a similaridade de cosseno para vetores de característica), mas elas geralmente podem não capturar características de comportamento único dos dados de interesse. Melhores resultados são esperados quando a métrica é projetada especificamente para a tarefa em questão. Imagine o esforço na definição e ajuste manual no cálculo das métricas no exemplo dado anteriormente para avaliar a similaridade entres as imagens, seria enorme e não há garantia de consistência em caso de mudanças das instâncias. O objetivo do aprendizado de métricas é automatizar o aprendizado através de escolhas precisas das funções de distâncias ideais em busca de uma métrica ótima independentemente do conjunto de dados (Kulis, 2012).

Em um nível alto, considera-se uma boa métrica, a partir de uma amostra de teste q, quando ao ordenar um conjunto de treinamento, com a distância em ordem crescente em relação a q, resulta em bons vizinhos na frente da lista e maus vizinhos no final (McFee e Lanckriet, 2010). Sob esta perspectiva, pode-se considerar a predição de vizinhos mais próximos como um problema de classificação e a taxa de erro na predição dos rótulos como uma função de perda sobre este problema. O problema de aprendizagem de métricas também pode ser considerado um caso especial de recuperação de informações no paradigma de consulta por amostra.

Figura 1 – Processo de predição com aprendizado de métricas



Fonte: Elaborado pelo autor

Os algoritmos de aprendizagem de métricas têm o objetivo de encontrar parâmetros para medição de distâncias entre vetores multidimensionais que sejam capazes de melhorar o desempenho em uma predição de rótulo de uma amostra. A nova métrica é aprendida a partir de um conjunto de dados e faz com que o processo de predição obtenha melhores resultados ao comparar-se com os resultados obtidos com a métrica genérica para o mesmo conjunto. Esse processo está resumido no fluxograma contido na Figura 1, baseado a partir do trabalho de Bellet *et al.* (2013).

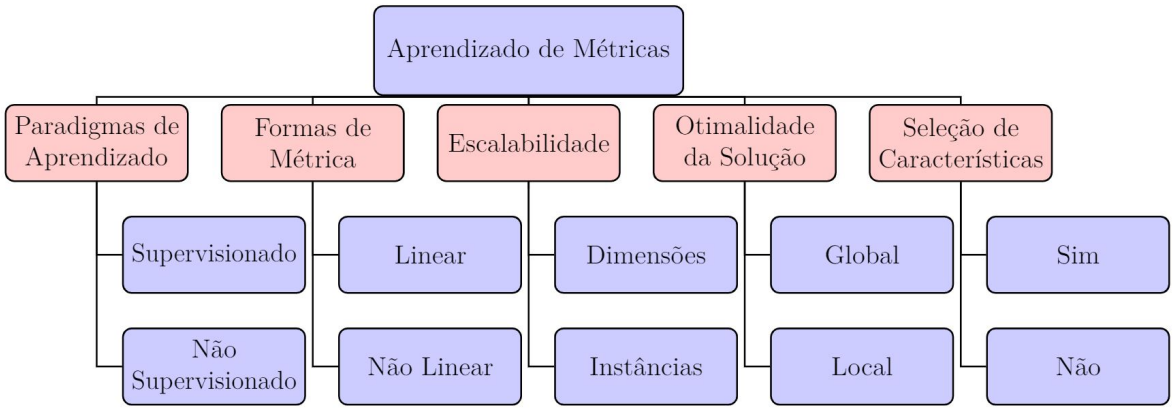
Uma noção métrica consistente é um desafio de várias áreas que estimulam ainda mais o avanço dos estudos para aprimorar sua forma de aprendizado. A motivação a este avanço é constatada na aplicação em problemas tão diversos quanto a previsão de links em redes (Shaw *et al.*, 2011), a representação do estado na aprendizagem por reforço (Taylor *et al.*, 2011), recomendação musical (McFee *et al.*, 2010), problemas de particionamento (Lajugie *et al.*, 2014), verificação de identidade (Ben *et al.*, 2012), arquivamento de páginas da Web (Law *et al.*, 2012), síntese de desenhos animados (Zhang e Yeung, 2010), entre outros (Bellet *et al.*, 2013).

Dependendo da disponibilidade dos exemplos de treinamento, os problemas de aprendizado de métricas podem estar contidos em duas categorias: aprendizado supervisionado e não supervisionado (Yang, 2006). Ao contrário da maioria dos algoritmos de aprendizado supervisionado, em que as instâncias de treinamento recebem rótulos das classes, aqui as instâncias de treinamento são lançadas em restrições de pares: as restrições de similaridade em que as instâncias pertencem às mesmas classes e restrições de dissimilaridade indicando que as instâncias pertencem à classes diferentes.

Ainda em Yang (2006), a otimalidade da solução de um problema deste tipo pode estar contida em escopo global ou local. O primeiro aprende a métrica de distância em um sentido global, ou seja, para satisfazer todas as restrições de pares simultaneamente com o objetivo de manter todos os pontos de dados dentro das mesmas classes próximos, enquanto separa todos os pontos de dados de classes diferentes (Bar-Hillel *et al.*, 2003). A segunda abordagem é aprender uma métrica de distância em uma configuração local, ou seja, apenas para satisfazer as restrições de pares locais. Isso é particularmente útil para recuperação de informações e para os classificadores baseados nos vizinhos mais próximos, pois os dois métodos são mais influenciados pelas instâncias de dados que estão próximas

dos exemplos de teste, ou consulta (Domeniconi e Gunopulos, 2002).

Figura 2 – Propriedades dos Algoritmos de Aprendizado de Métrica



Fonte: Elaborado pelo autor

A Figura 2, construída baseada nas definições de Yang (2006) e Kulis (2012), representa as possíveis características que um algoritmo aplicado a um problema de aprendizado de métricas pode possuir. A seguir, uma síntese de cada uma delas:

- Paradigma de aprendizado
  - Supervisionado: o aprendizado de métricas supervisionado pode utilizar a informação sobre a classe da instância, mas a informação principal aqui é a distância que os pares de instâncias de treinamento possuem, obtida através de uma função $d(x, y)$. Esta distância é a forma de avaliar e orientar o algoritmo de aprendizado que utiliza uma nova função $\hat{d}(x, y)$. A organização das distâncias entre pares de instâncias é feito por S para o conjunto de instâncias similares, D para dissimilares e, em alguns casos, P para triplas de instâncias que estabelecem uma relação de similaridade e dissimilaridade entre as instâncias;
  - Não supervisionado: por não possuir nenhuma informação de classe e grau de similaridade entre instâncias, o aprendizado supervisionado inicia-se com o processo de mensuração das distâncias entre parte das instâncias e a partir daí buscar pelo aprendizado por meio de relações entre distâncias conhecidas. Há uma forte conexão entre aprendizado de métricas não supervisionado e redução de dimensão. Em geral, as abordagens de redução de dimensionalidade têm como essencial o aprendizado de uma métrica de distância sem informações de rótulo para a construção do novo conjunto de dados com menor dimensionalidade;
- Transformações
  - Linear - sua eficiência está condicionada à disponibilidade das instâncias do problema, porém há uma facilidade para otimizar e, geralmente, levam a formulações convexas que possibilitam uma otimização global da solução. Sua adoção

é mais comum em problemas supervisionados. Em problemas não supervisionados destaca-se o uso em soluções que utilizam redução de dimensionalidade junto a algoritmos para análise de componentes principais.

  - Não linear - muitas vezes dão origem a formulações não convencionais, mas têm capacidade de capturar variações não lineares nos dados. Em geral, aprender uma transformação não linear é difícil e ao contrário das transformações lineares, que podem ser expressas como uma matriz de parâmetros, o conjunto de transformações não lineares não é facilmente parametrizado.

- Escalabilidade

  - Em dimensões: capacidade de o algoritmo ser executado em problemas em que as instâncias possuam um número elevado de dimensões. Todavia, como muitas vezes os algoritmos de aprendizado de métricas são formulados com o objetivo de aprendizagem de matrizes quadradas, projetar algoritmos que escalem bem em número de dimensões é um desafio considerável.
  - Em instâncias: capacidade de o algoritmo ser executado em problemas em que haja um numero elevado de instâncias. Muitas vezes os algoritmos de aprendizado de métricas são formulados para satisfazer as restrições entre pares ou triplas de instâncias, há uma dificuldade para obter escala em relação ao aumento do número de amostras.

- Otimalidade da Solução

  - Global - abordagens nesta categoria tentam aprender métricas que mantenha próximas todas as instâncias de uma mesma classe, enquanto afasta instâncias de classes diferentes. Há uma garantia que a métrica é ótima, porém é necessário que a formulação represente a um problema de otimização convexa.
  - Local - o objetivo é aprender uma métrica de distância em uma configuração local, ou seja, apenas para satisfazer as restrições de pares locais. E particularmente útil para recuperação de informações e para os classificadores KNN, pois as duas abordagens são influenciadas pelas instâncias de dados que estão próximas da instância de teste/consulta.

- Seleção de Características

  - Ocorre quando o algoritmo permite a anulação de parâmetros que definem o peso de uma ou mais dimensões

Uma formulação informal do problema de aprendizado de métricas poderia ser dada da seguinte maneira: dada uma função de distância inicial $d(x, y)$ entre objetos x e y (por

exemplo, a distância euclidiana), juntamente com informações supervisionadas a respeito de uma distância ideal, construa uma nova função de distância $\hat{d}(x, y)$ que é "melhor" que a função de distância original (também pode-se facilmente substituir "distância" por "semelhança" e d por s para alguma função de similaridade $s(x, y)$).

Também devemos considerar o aprendizado de métricas de forma não supervisionado em nossa formulação. Tomemos, por exemplo, a redução de dimensionalidade: métodos lineares, como uma análise de componentes principais, podem ser vistos como a construção de uma transformação linear P a ser aplicada globalmente aos dados, de maneira não supervisionada. A distância resultante entre objetos é, portanto, $d(Px, Py)$, e pode-se afirmar que esta é também uma forma de aprendizado métrico.

Em geral, a maioria dos algoritmos de aprendizado tenta encontrar uma distância que mantenha todos os pares de amostras próximos, desde que sejam equivalentes, ou afastados, contanto que sejam diferentes. A formulação inicial do problema definiu-se por: considere $X = \{x_1, x_2, ..., x_n\}$ como um conjunto de amostras, onde n representa o número de amostras no contidas no conjunto. Cada $x_i \in R^m$ representa uma instância onde m é seu número de dimensões. A seguir apresenta-se o conjunto de restrições de similaridade pela notação

$$S = (x_i, x_j)|\ x_i \text{ e } x_j \text{ pertençam a mesma classe}$$

e o conjunto de restrições de dissimilaridade denota-se por

$$D = (x_i, x_j)|\ x_i \text{ e } x_j \text{ pertençam a diferentes classes.}$$

A métrica de distância denotada pela matriz $A \in R^{m \times m}$, e a distância entre duas amostras quaisquer x e y é expresso por:

$$d_A^2(x, y) = \| x - y \|_A^2 = (x - y)^T A(x - y) \tag{2.1}$$

Entre os estudos iniciais sobre aprendizado de métricas está o trabalho de Wagstaff *et al.* (2001) que propôs o uso de informações adicionais no algoritmo k-means de Macqueen (1967). A proposta consistia em inserir restrições de similaridade, ou dissimilaridade, entre pares de vetores de um subconjunto dos dados. Essas restrições foram denominadas *"must-link"*, indicando que dois vetores devem estar contidos no mesmo grupo, e *"cannot-link"*, representando dois vetores que devem estar contidos em grupos diferentes. Porém o método não garantia a convergência para uma solução que atendesse a todas as restrições de similaridade e dissimilaridade estabelecidas.

Pouco tempo depois Xing *et al.* (2002) apresentam o trabalho mais representativo da área, no qual ele realiza a formulação do aprendizado de métricas como um problema de otimização convexa com restrições, o que possibilitou aprender uma métrica de distância global que minimiza a distância entre os pares de instâncias similares, sujeitas às restrições das instâncias dissimilares que estão bem separadas.

Também em Xing *et al.* (2002) é definido os bons vizinhos como todos os pontos com o mesmo rótulo e resolve a métrica por programação semidefinida. As distâncias para pares de instâncias semelhantes são limitadas por uma constante e as distâncias de pares diferentes são maximizadas (Yang, 2006). Dadas as restrições de equivalência em S e as restrições de desigualdade em D, Xing *et al.* (2002) formularam o problema da aprendizagem métrica no seguinte problema de programação convexa (Vandenberghe e Boyd, 1996):

$$\min_{A \in R^{mxm}} \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A^2 \tag{2.2}$$
$$\text{sujeito a: } A \geq 0, \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A^2 \geq 1.$$

A modelagem proposta por Xing *et al.* (2002), então, ocorre da seguinte forma: utiliza-se como entrada um conjunto de relações de similaridade S, envolvendo pares de vetores (x_i, x_j) que pertencem a um mesmo cluster, e um conjunto de relações de dissimilaridade D, envolvendo pares (x_i, x_j) que pertencem a clusters distintos. Então é computado o somatório das distâncias ao quadrado $d_A^2(x_i, x_j)$ de todos os pares de vetores do conjunto S, considerando que a matriz A é definida inicialmente como a matriz inversa da matriz de covariância (M^{-1}), ou seja, a distância de Mahalanobis ao quadrado.

Após isso, a matriz A tem seus parâmetros (autovalores) corrigidos iterativamente, de forma que seja minimizado esse somatório. Contudo, tornam-se necessários mecanismos para que a matriz A se mantenha semidefinida positiva, bem como, não zere todos os seus parâmetros. Para isso, foi formulada uma restrição de dissimilaridade que mantivesse a característica de convexidade.

Embora o problema formulado por Xing *et al.* (2002) esteja contido na categoria de programação convexa, Yang (2006) destaca duas razões para que a eficiência da solução não seja a melhor possível: ele não cai em nenhuma classe especial de programação convexa, como programação quadrática (Gill *et al.*, 1981) e programação semidefinida (Vandenberghe e Boyd, 1996), e como resultado, ele acaba sendo resolvido como uma abordagem genérica, em que pode ser incapaz de tirar proveito da estrutura especial do problema; como apontado por Zhang *et al.* (2003), o número de parâmetros de A é quase quadrático em relação ao número de características, o que dificulta a escalabilidade do problema em cenários com um grande número de características.

O trabalho de Xing *et al.* (2002) deu origem a diversas extensões através de outros trabalho (Kulis, 2012; Yang, 2006), porém, entre eles, destaca-se o trabalho de Schultz e Joachims (2004) devido a forma com que ele interpreta a relação de similaridade e dissimilaridade entre instâncias. Baseando-se em Joachims (2002), viu-se que a representação do resultado de uma consulta na forma A está mais próximo de B do que A está para

C, é um exemplo de aplicação que mostra a possibilidade de aprendizado de métricas de similaridade. No caso do estudo de base, avaliou-se a semelhança entre documentos a partir de uma consulta. Considere uma lista ranqueada contendo os resultados desta consulta em que os documentos selecionados a partir de um clique podem ser considerados semanticamente mais próximos do que os documentos observados pelo usuário, mas decidiram não clicar (ou seja, $A_{clicado}$ está mais próximo de $B_{clicado}$ que $A_{clicado}$ está em $C_{nao-clicado}$).

Por outro lado, concluir que $A_{clicado}$ e $C_{nao-clicado}$ não são similares pode ser uma afirmação incorreta, visto que $C_{nao-clicado}$ possui uma posição alta no ranking de resultados e, provavelmente, está mais próximo de $A_{clicado}$ do que a maioria dos outros documentos do conjunto com posições abaixo da sua no ranking. Seguindo Schultz e Joachims (2004), também concorda-se que o feedback com informações contendo as distâncias relativas possuem maior disponibilidade em aplicativos do que exemplos quantitativos (por exemplo, a distância entre A e B é 7,35) ou um feedback qualitativo absoluto (por exemplo, A e B são semelhantes, A e C não são semelhantes) como considerado em (2.2).

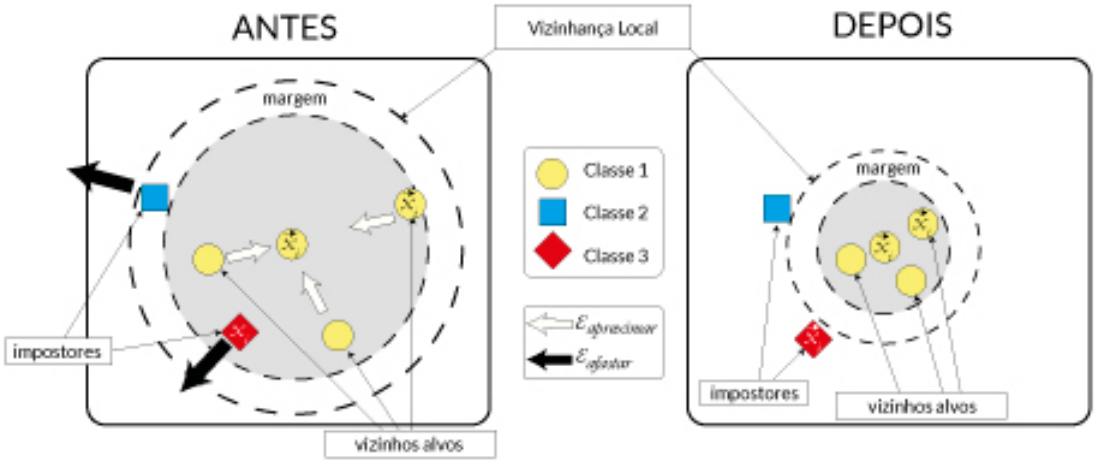
Weinberger e Saul (2009); Weinberger *et al.* (2006) propõem o aprendizado de métricas com foco em classificação pelo algoritmo KNN. Uma abordagem é apresentada com o objetivo de maximização da margem por meio de uma função objetivo convexa. O método, denominado LMNN (Large Margin Nearest Neighbor), tem sua formulação semelhante ao método de Schultz e Joachims (2004), utilizando as restrições de distâncias relativas, R, mas também fazendo uso das comparações par a par, S, com distâncias absolutas. Este modelo é um dos métodos mais populares de aprendizagem de métricas Bellet *et al.* (2013), e sua formulação é disposta na Equação (2.3).

$$\min_{A \geq 0} \sum_{(i,j) \in S} d_A(x_i, x_j) + \lambda \sum_{(i,j,k) \in R} [1 + d_A(x_i, x_j) - d_A(x_i, x_k)]_+ \tag{2.3}$$

A intuição por trás do LMNN é bastante direta: o objetivo é que uma determinada instância compartilhe o mesmo rótulo de seus vizinhos mais próximos, enquanto outras instâncias que possuem rótulos diferentes devem se afastadas. A distância relativa com margens impõe tal intuição. A Figura 3 ilustra a ideia por trás do LMNN: o termo vizinho alvo refere-se a uma instância que deve ser semelhante, enquanto um impostor é uma instância que é o vizinho próximo, entretanto tem o rótulo diferente. O objetivo do LMNN é minimizar o número de impostores por meio de restrições de distância relativa. O conjunto S é definido como todos os pares de vizinhos semelhantes, o que é frequentemente dado pelo conjunto de pares (i, j), onde x_j é um dos k vizinhos mais próximos na mesma classe que x_i; o conjunto R é definido como todos as triplas (i, j, k) de tal forma que x_i e x_j são vizinhos semelhantes e x_k é uma instância com rótulo diferente.

No contexto de aprendizado online, em Blum (1998) discute-se alguns resultados,

Figura 3 – Comportamento do LMNN



Fonte: Figura adaptada a partir de Weinberger e Saul (2009)

modelos e problemas abertos da Teoria da Aprendizagem Computacional relacionada ao uso de algoritmos online. Particularmente, o trabalho descreve um conjunto de algoritmos online relacionados à solução do problema de "predição a partir de informações de especialistas" e um conjunto de algoritmos online relacionados a um cenário de aprendizado mais geral que é o problema de "aprender uma classe conceitual a partir de exemplos". Zinkevich (2003) introduz o conceito de programação convexa online para resolver o problema em jogos de repetição e apresenta um algoritmo para otimizar as funções convexas gerais com base no gradiente descendente que usa uma projeção gulosa. A principal contribuição deste trabalho é uma prova de que o *regret* médio do algoritmo de projeção gulosa converge para zero contra um algoritmo offline que possui todas as informações antes de tomar qualquer decisão.

Seguindo essa abordagem online, Bansal *et al.* (2003) formularam um problema de roteamento online como um problema de programação convexa online, aplicando o algoritmo de projeção guloso para obter uma solução online. Em Anava *et al.* (2013), um algoritmo online de projeção gulosa é proposto para resolver o problema da predição de séries temporais. Os autores também mostram que seu algoritmo aborda assintoticamente o desempenho do melhor modelo ARMA (média móvel autorregressiva). Além disso, Anava *et al.* (2013) desenvolveram um algoritmo online de projeção gulosa baseado no método de Newton para resolver o problema universal de gerenciamento de portfólio.

Os algoritmos para aprendizado de métricas podem fornecer funções de distância úteis para uma variedade de domínios e, em grande parte dos trabalhos referenciados, mostraram boa precisão para problemas em que pode-se acessar todas as restrições de distância de uma só vez. No entanto, em muitos aplicativos reais, as restrições estão disponíveis apenas de forma incremental, necessitando, portanto, de métodos que possam realizar atualizações online para a métrica aprendida.

2.3.1 **Soluções para classificação musical**

O interesse em utilizar o aprendizado de métricas para mensurar a similaridades entre músicas também é visto em diversos trabalhos como solução de problemas de classificação, agrupamento e, principalmente, recuperação de músicas com intuito de efetuar a recomendação de músicas semelhantes.

A proposta de Xin *et al.* (2014) é composta pela construção de um vetor característico utilizando apenas informações acústicas, avaliação da similaridade utilizando a métrica de Wasserstein (Vershik, 2013) para realização da classificação e em seguida realiza-se a recuperação de uma música mais semelhante definida a partir de outra medida que, neste caso, irá mensurar o ganho de informação entre duas músicas.

Wolff e Weyde (2014) apresentou um estudo da eficácia de várias características acústicas, de forma unitária ou composta por uma combinação, aplicadas à uma avaliação abrangente do desempenho de classificadores de diferentes abordagens em relação a classificação baseada no aprendizado de métricas. A avaliação voltou-se à aprendizagem métrica com base em máquinas de vetores de suporte (SVM) e aprendizado de métricas para ranqueamento (MLR) e aprendizado relativo à distância com redes neurais (RDNN). As abordagens que utilizaram o processo de aprendizado de métricas para classificação obtiveram os melhores desempenhos.

A aplicação de vários algoritmos de aprendizagem, feita por Slaney *et al.* (2008), para encontrar métricas de similaridade sobre características acústicas que são otimizadas para agrupar músicas do mesmo artista ou álbum. Já Slaney e White (2007) apresentam um método para aprender uma métrica de distância de Mahalanobis. Eles avaliam a similaridade a preferência dos usuários sobre as variações dos tipos de jazz em relação aos artistas. Seus experimentos avaliam a eficácia das métricas de similaridade comparando-a com os vizinhos mais próximos de uma certa música, em termos de nomes de artistas. Ainda é feita uma comparação entre métricas obtidas a partir de informações baseadas em filtragem colaborativa e baseadas em conteúdo extraído diretamente do arquivo de áudio.

McFee e Lanckriet (2010) realizam a parametrização de uma métrica de similaridade musical usando dados obtidos a partir de uma filtragem colaborativa combinados a características acústicas. Se usuário escutou e "curtiu" uma música prova ser uma avaliação altamente efetiva para fornecer recomendações musicais relevantes. A conclusão do trabalho apresenta uma técnica relevante para construção de uma modelo para recomendação, porém destaca que sua aplicabilidade é altamente dependente da disponibilidade dos dados de usuários no filtro colaborativo e da popularidade da música.

Diferentemente das propostas de McFee e Lanckriet (2010) e Slaney e White (2007), que utilizam o gênero e o artista da música, respectivamente, como alvo da predição, baseando-se na ideia de centroide mais próximo para classificar uma música, nossa proposta

também utiliza a mesma ideia da classificação pelo centroide mais próximo, porém solicita um feedback ao usuário para determinar a inserção do rótulo de gênero à música, ou aproximá-la de um outro grupo em que o usuário julga ser mais semelhante. Assim, torna-se possível adaptar o conjunto de músicas à preferência do usuário de forma individualizada.

O problema de aprendizado de métricas foi resolvido como um problema de otimização convexa, considerando a minimização online de um conjunto de distâncias parametrizadas medidas sobre pares de amostras e sujeito à desigualdade triangular, restrições simétricas e não negativas (Xing *et al.*, 2002). Seguindo Schultz e Joachims (2004) optou-se por aprender uma matriz diagonal, resultando em um vetor de parâmetros. Nesse caso, aprende-se uma métrica que pondera as diferentes dimensões do espaço do problema. Essa abordagem pode ser considerada como o uso de uma perda contrastiva (Hadsell *et al.*, 2006) que tenta minimizar uma margem parametrizada entre amostras similares e maximizar entre àquelas dissimilares, semelhante à demonstração de Weinberger e Saul (2009); Weinberger *et al.* (2006) em (2.3). A métrica aprendida define a preferência de cada usuário individualmente, pois cada um é responsável por seu próprio feedback. Portanto, de acordo com o este feedback, o modelo sofre uma perda quando o usuário discorda da classificação e o algoritmo faz a correção e ajustando adequadamente o conjunto de parâmetros.

O método proposto para aprender a similaridade musical tem uma relação direta com a solução ao problema de Predição Estruturada proposta por Coelho *et al.* (2017). O método baseia-se em satisfazer um conjunto de restrições de comparação entre pares. Essas restrições são escalonadas na ordem $O(n)$ com o número de amostras e representam a condição de que a distância parametrizada entre uma música e seu centroide de gênero deve ser menor que em relação a qualquer outra alternativa. Além disso, acrescenta-se uma função de perda contrastiva baseada em margem, garantindo que exemplos musicalmente semelhantes sejam incorporados junto com esses respectivos grupos de gênero. Como mencionado anteriormente, este trabalho tem uma grande similaridade com o modelo de comparações relativas proposto em Wolff e Weyde (2014) que tem uma abordagem SVM estruturado. Nesse modelo, cada restrição representa a relação de similaridade entre uma tripla de amostras refletindo o fato de que uma amostra x_i é mais semelhante à amostra x_j do que a amostra x_k (Schultz e Joachims, 2004). A principal desvantagem dessa formulação é o número de restrições que escala em ordem $O(n^3)$ com o número de amostras.

2.4 CARACTERIZAÇÃO MUSICAL

A classificação de gêneros musicais é abordada por diversos estudos na literatura, com o uso de diferentes descritores musicais e esquemas de classificação. Este tipo de problema de classificação é, geralmente, disposto sob cinco perspectivas diferentes,

dependendo do tipo de informações utilizadas para caracterização musical, segundo Silla e Freitas (2009) são eles: baseados em conteúdo de áudio, conteúdo simbólico, letras, metadados (como filtros colaborativos) e abordagens híbridas.

O problema de classificação baseado em conteúdo de áudio explora características acústicas das músicas em sinais de áudio digital, enquanto que em conteúdo simbólico as características são extraídas de formatos de dados simbólicos como MIDI, no qual os eventos musicais são apresentados em um nível mais alto de abstração. Já os baseados sobre as letras das músicas, utilizam de técnicas de mineração de texto para extrair informações e analisar semanticamente para realizar a classificação. A utilização de metadados compõe soluções análogas às relatadas na seção de filtros colaborativos. Representando à perspectiva mais recente, os sistemas baseados em modelos híbridos usam recursos de música originados de duas ou mais abordagens anteriores (Corrêa e Rodrigues, 2016).

Neste trabalho, utiliza-se apenas características acústicas extraídas diretamente do arquivo de áudio. Existem diversas maneiras para representar as músicas que variam entre escolha de característica junto às técnicas para extração e construção do vetor característico. A seguir, serão apresentadas algumas das principais características (Fu *et al.*, 2011) junto a cenários em que elas são recomendadas.

2.4.1 CARACTERÍSTICAS DE BAIXO NÍVEL

A importância das características de baixo nível é devido à facilidade de extração e sua capacidade de representação do áudio, constatada através do bom desempenho nos problemas de classificação de música. Corrêa e Rodrigues (2016) apresenta uma síntese das principais características e cenários em que são recomendadas.

Como elemento básico da música, timbre é um termo que descreve a qualidade do som, podendo ser assimilado à cor nas imagens. Timbres diferentes são produzidos por diferentes tipos de fontes sonoras, como diferentes vozes e instrumentos musicais. Fu *et al.* (2011) apresentam uma grande quantidade de características de timbre que são utilizadas para classificação, porém, apesar desta quantidade, elas apresentam muitas semelhanças visto que o processo de extração está diretamente relacionado à análise espectral do sinal de áudio e segue alguns passos padrões.

Uma música é, geralmente, longa e pode durar alguns minutos. Assim, o sinal de áudio de entrada para cada música pode conter milhões de amostras com uma frequência de amostragem acima de 10 kHz. Em vez de analisar todo o sinal em uma única vez, realiza-se sua divisão em muitos quadros locais que facilitam a análise e a extração das características de timbre. Destaca-se neste processo duas vantagens: processo de extração de características mais eficiente, uma vez que só precisa-se aplicar análise espectral a sinais de curto prazo e quadro a quadro; é mais eficaz modelar as características de timbre em quadros com duração de 10 ms a 100 ms.

Após o processo de enquadramento, técnicas de análise espectral, como a transformada de Fourier, são aplicadas sobre cada quadro do sinal. A partir dos espectros de magnitude resultantes, podem-se obter algumas características, como Spectral Centroid (SC), Spectral Rolloff (SR), Spectral Flow (SF), Zero Crossing Rate (ZCR) e Spectral Bandwidth (SB), ao capturar estatísticas dos espectros. Para extrair características mais poderosas, como descritores de áudio Mel-frequency Cepstrum Coefficients (MFCC), Octave based Spectral Contrast (OSC) e Daubechies Wavelet Coefficient Histogram (DWCH) e Amplitude Spectrum Envelop (ASE), é realizada uma análise de sub-bandas decompondo o espectro de magnitude em sub-bandas e aplicando a extração de características em cada sub-banda. Os detalhes da configuração de sub-bandas e o processo para extração variam de acordo com os tipos de características. Normalmente, as sub-bandas estão dispostas com faixas espectrais logaritmicamente espaçadas (OSC, ASE, DWCH) por uma diferença de frequência constante (Brown, 1991).

Para MFCC, as sub-bandas são linearmente espaçadas em frequências mais baixas, até 1 kHz, e logaritmicamente espaçadas em frequências mais altas. A justificativa para usar uma escala logarítmica na faixa de frequência mais alta é simular o sistema auditivo humano, pelo qual uma resolução espectral mais fina é alcançada em frequências mais baixas. Verificou-se que o arranjo era um *trade off* razoável, devido ao sucesso para classificação musical utilizando as características obtidas pelo MFCC (Fu *et al.*, 2011), mais especificamente em problemas de classificação de gêneros (Lidy e Rauber, 2005) e sentimentos (Chua, 2007).

As características temporais formam outra classe importante de características de baixo nível que capturam a evolução temporal do sinal. Destaca-se que as características temporais são, geralmente, construídas em cima das características de timbre, ou em um espectrograma resultante de sua extração, obtidas sobre os vários quadros do sinal e integradas para criar a característica temporal. Desta forma, pode-se criar um conjunto mais rico de características para classificação através de diferentes combinações entre características temporais e de timbre.

O tipo mais simples de características temporais são momentos estatísticos como média, variância, covariância de características de timbre coletadas em um intervalo, ou janela. E uma combinação entre a média e variância dos vetores compostos por características de timbre apresentou bom desempenho para solucionar um problema de classificação de música (Tzanetakis e Cook, 2002). A janela para o cálculo de SM sobre características de timbre geralmente possui alguns segundos de duração e de quadros.

O conjunto de vetores de características de timbre, dentro das janelas, também pode ser tratado como dados de séries temporais e a aplicação de uma transformação espectral, como a FFT, à série temporal e obter novas características, incluindo padrão de flutuação (FP) (Pampalk *et al.*, 2002), padrão rítmico (Lidy *et al.*, 2007), coeficiente

rítmico (West, 2008) e características mais sofisticadas derivadas da análise do espectro de modulação (Lee *et al.*, 2009). Muitas destas características são geradas a partir de processos de extração semelhantes, analisando a modulação da amplitude do espectro (Fu *et al.*, 2011).

De maneira geral, qualquer uma das técnicas de processamento de sinais usadas na extração de características de timbre também pode ser adotada para obter características temporais. A diferença é que a extração de características de timbre é executada no sinal de áudio de entrada nas janelas, enquanto a extração de características temporal é executada na série de características de timbre extraídas dentro de janelas com um intervalo maior.

2.4.2 CARACTERÍSTICAS DE MÉDIO E ALTO NÍVEL

Apesar do bom desempenho das características de baixo nível para classificação musical, elas não capturam as propriedades intrínsecas da música perceptíveis às pessoas. Mesmo as características de baixo nível sendo usadas de forma recorrente em várias tarefas de classificação de música, as de médio e alto nível possuem aplicação sobre certos problemas que requerem recursos capazes de destacar diferenças semânticas de nível mais alto, como consulta, por exemplo, (Jang e Lee, 2008; Unal *et al.*, 2008) e detecção de músicas interpretadas em forma de covers (Tsai *et al.*, 2005; Gómez e Herrera, 2006). Dentre esta categoria, destacam-se:

- Harmonia: combinação de notas simultaneamente para produzir acordes;
- Ritmo: intervalos musicais com padrões recorrentes;
- Tom: percepção da frequência fundamental do som.

Embora estas características sejam facilmente identificadas pelos ouvintes de música, não é fácil defini-las explicitamente e extraí-las de maneira confiável a partir de sinais de áudio para fins de análise musical.

Entre elas, o ritmo é a mais utilizada na classificação musical baseada em áudio. Com ela é possível determinar padrões musicais que ocorrem, se repetem e estão relacionados à "dançabilidade" da música. Batida e tempo (batida por minuto, BPM) são dois sinais importantes que descrevem o conteúdo rítmico de uma música que foi utilizada na classificação musical.

Tzanetakis e Cook (2002) utilizam histograma das batidas (BH) modelar as distribuições das regularidades exibidas no sinal onde as características rítmicas podem ser obtidas como grandezas e localizações dos picos dominantes e BPM. As características do ritmo demonstraram bom desempenho para classificação de sentimentos (Yazhong Feng

et al., 2003; Yang e Lee, 2004; Korhonen *et al.*, 2006). O objetivo destas aplicações são explicados, pois o clima de uma música é altamente correlacionado ao ritmo. Por exemplo, músicas tristes geralmente têm ritmo lento, enquanto músicas agitadas costumam ter um ritmo rápido.

O tom é outro componente importante da música e é determinado pelo que o ouvido julga ser a frequência mais fundamental do som. Esta associação do tom à frequência fundamental levanta alguns questionamentos, pois como a percepção do tom de uma música, que ocorre de forma subjetiva, pode representar a medição de uma frequência, que é uma medida objetiva. E também, visto que uma nota musical tocada na maioria dos instrumentos consiste em uma série de frequências relacionadas a harmônicos, incluindo a frequência fundamental, e é normalmente percebida como um som com um único tom.

Por mais que o conceito de tom esteja atrelado à frequência fundamental, a enciclopédia de música, Grove Music Online[2], acrescenta alguns outros fatores em sua definição, como diferenças de timbre, intensidade sonora e contexto em que a música está inserida, também afetam a extração de informações de tom em sinais de áudio reais. Sua representação também é por meio de níveis e pode ser feita por com o histograma do tom, e com aplicações nos mesmos cenários do ritmo (Tzanetakis *et al.*, 2002). Há aplicações em classificação de gêneros musicais utilizando o tom junto às características de timbre de baixo nível, como MFCC (Tzanetakis e Cook, 2002), e classificação de sentimentos (Li e Ogihara, 2003).

A harmonia é alcançada pela progressão de acordes, uma série de acordes tocados sucessivamente. Um acorde é o componente fundamental da harmonia, que envolve a combinação simultânea de duas ou mais notas. Informações de acordes, como sequências de acordes (CS), podem ser extraídas dos dados de áudio de música, usando vários métodos de detecção e reconhecimento de acordes (Fujishima, 1999; Lee, 2006).

Todos estes métodos começam com a detecção de tom, usando uma técnica padrão ou aprimorada (Gómez e Herrera, 2004), para identificar a frequência fundamental e seus parciais. Em seguida, cada característica de histograma de tom é comparada ao modelo de acordes para identificar a existência de possíveis acordes. As características de acorde complementam as características de tom na correspondência de similaridade baseada na melodia e na detecção de covers de músicas (Ellis e Poliner, 2007; Bello, 2007). Em problemas de classificação de gênero, a harmonia, por si só, não possui tanta utilização, embora, combinada às características de timbre e ritmo teve sua aplicação sobre classificação de sentimentos (Heng-Tze Cheng *et al.*, 2008) e busca pelo início e fim de um respectivo som dentro de uma música (T. Lidy e Inesta, 2007).

Resumindo, escolher características para representar um áudio depende muito dos

[2] https://www.oxfordmusiconline.com/grovemusic, Acessado em 10 de agosto de 2019

problemas abordados. As características de timbre são adequadas para classificação de gêneros e identificação de instrumentos (Loughran *et al.*, 2008), mas não são apropriadas para comparar a similaridade de melodia de duas canções (Fu *et al.*, 2011). Para a classificação de sentimentos, vários trabalhos adotam características extraídas a partir do ritmo (Yang e Lee, 2004; Korhonen *et al.*, 2006). Embora as características de pitch e harmonia não sejam muito populares em sistemas de classificação de gênero, artista, humor, etc., elas são muito importantes para recuperação de música similares e detecção de covers de músicas em nível melódico (W. H. Tsai e Wang, 2005; Gomez, 2006; Bello, 2007; J. Serra e Serra, 2008), onde as características de timbre não conseguem bons resultados. Em geral, não há um conjunto único de características, independentemente do problema, que possam superar de maneira consistente às demais.

Ainda em Fu *et al.* (2011), é reportado o desempenho de vários classificadores ao resolver o problema de classificação de gêneros, com as músicas dispostas no conjunto GTZAN, usando diversas combinações de características diferentes tipos; e em McKinney e Breebaart (2003) é feito uma extensa avaliação sobre o desempenho de quatro conjuntos de características de baixo nível, MFCC e dois novos conjuntos de características baseadas sobre o modelo de percepção auditiva, também para resolver um problema de classificação de gêneros. Estes estudos foram muito importantes para prosseguimento desta pesquisa, pois a proposta deste trabalho visa a classificação de gêneros através de características acústicas, porém não tem-se como objetivo principal um estudo paralelo de seleção de características e apresenta-se desenvolvimento de um processo de aprendizado como uma nova abordagem para solução do problema de classificação de gêneros musicais. Em ambos os trabalhos, a conclusão obtida foi que as características perceptíveis à nossa audição oferece um melhor desempenho aos classificadores, sendo a característica usada como base na maioria dos experimentos foi MFCC. Devido a estes estudos, também adota-se esta característica para representação das músicas em nosso modelo.

2.5 MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Entre as principais componentes de uma música estão informações importantes da voz do artista, que são altamente perceptíveis, tais como, dialeto, contexto, estilo de falar, estado emocional, etc; e também dos instrumentos que mesmo com pouco tempo de exposição a uma música e tocados no mesmo volume, a maioria das pessoas pode facilmente distinguir entre instrumentos musicais familiares. Há outras características que dificilmente são perceptíveis, tais como tons, frequências dos formantes, intensidade (Siqueira, 2012).

A capacidade humana de distinguir entre instrumentos musicais tem sido objeto de investigação há vários anos. Sabe-se que a qualidade da sensação auditiva pela qual um ouvinte pode distinguir entre dois sons de intensidade, duração e altura iguais estão

relacionadas ao timbre. Por isso, pode-se dizer que o reconhecimento de instrumentos musicais é amplamente dependente do timbre (Loughran *et al.*, 2008).

Segundo Giannakopoulos e Pikrakis (2014), a técnica para extração dos MFCC faz uma análise de características obtidas a partir de pequenos quadros da música, baseando-se no uso do espectro do sinal convertido para uma escala de frequências denominada MEL. Estes coeficientes são uma representação definida como o espectro de um sinal janelado no tempo, derivado da aplicação da transformada de Fourier e dispostos sobre escalas de frequência não lineares.

Com a aplicação da transformada, obtém-se os coeficientes com o sinal no domínio da frequência e a diferença de larguras entre as janelas está relacionada à escala da frequência utilizada. Sabendo que os valores em Hertz não refletem tão bem a percepção auditiva humana, (Stevens *et al.*, 1937) desenvolveu uma escala que visa imitar as características únicas perceptíveis pelo ouvido humano. Através de experimentos com diversos ouvintes criou-se a escala mel, cujo nome veio da palavra melodia, que mostrou-se eficaz no processo de distinção entre som diferentes, fazendo com que 2 mel soe duas vezes mais aguda que 1 mel. Por exemplo, um sinal senoidal de 880 Hz não soa duas vezes mais agudo que um de 440 Hz e nem quatro vezes mais agudo que um de 220 Hz.

O desempenho superior da técnica MFCC, em diversos trabalhos, é atribuído ao fato de ela representar melhor os aspectos perceptíveis do espectro do som de curta duração com auxílio dos filtros espaçados linearmente para baixas frequências e logaritmicamente para altas frequências que são usados para capturar as características mais importantes do som (Okida e Queiroz, 2006).

Diferentes cenários foram explorados com intuito de avaliar o desempenho de diferentes características na representação musical e solução de alguns problemas. A maioria destes cenários e seus problemas já foram citados em seções anteriores junto às outras características musicais adotadas. Entretanto, a adesão do MFCC como característica principal para representação musical tem sido tão frequente na literatura que iremos apresentar mais alguns trabalhos em que ela foi aplicada de forma exclusiva, combinada com outra característica ou junta à alguma derivação.

Há dois componentes importantes a serem considerados para uma melhor classificação de gêneros musicais, extração de características e classificador. Lee *et al.* (2009) incorpora dois tipos diferentes de tipos de características para classificação de gênero, timbre, usando MFCC, e ritmo, com o histograma das batidas. Para decidir qual característica de timbre seria utilizada, foi feito uma avaliação entre sete características e o MFCC apresentou o melhor desempenho para classificar gêneros. Após a escolha das características, os autores variaram o número dos coeficientes do MFCC e a dimensionalidade do vetor característico e mensuraram o desempenho do classificador.

Li *et al.* (2003) realizaram a comparação das características MFCCs, FFT, ritmo e tom de forma isolada e com combinações entre si buscando a solução mais efetiva para o problema de classificação de gênero. Além destas, há uma proposta de uma nova característica construída a partir de trechos de outras características de timbre, sendo todas elas avaliadas sobre 5 classificadores diferentes. Seus resultados apresentam a acurácia obtida pelos classificadores em cada conjunto de características e, por serem relevantes à esta pesquisa, serão discutidos posteriormente.

Pesquisas para determinar e distinguir entre diferentes classes de instrumentos se tornaram mais populares à medida que o campo da análise de áudio se expandiu mais para a análise musical e uma série de estudos analisou os MFCCs na identificação de sons. Herrera *et al.* (2000) apresenta uma revisão exaustiva dos métodos usados na identificação automática de instrumentos musicais. A partir desta revisão, é evidente que as qualidades temporais e espectrais são necessárias para a identificação precisa do instrumento. O estudo analisa o uso de MFCCs ao longo da duração temporal de uma nota.

Brown (1999) distinguiu entre oboés e sons de saxofone calculando MFCCs e aplicou o algoritmo kmeans para formar grupos de cada instrumento. Eronen e Klapuri (2000) incluiu MFCCs como uma de suas características em examinar uma ampla gama de instrumentos orquestrais. Já Logan (2000), examinou alguns dos pontos mais sutis do MFCC na análise musical, além da voz, mostrando o quão útil é sua aplicação neste cenário.

Outras abordagens vão ao encontro do uso dos MFCCs, como na classificação binária para diferenciação entre fala e música em um sinal sonoro apresentada por Giannakopoulos e Pikrakis (2014) em que obteve-se resultados significativos à capacidade discriminativa desta característica na solução do problema. Já Vyas (2014) desenvolveu um método para classificar trechos de músicas de acordo com rótulos de sentimentos. Para identificar o sentimento, utilizaram a combinação do MFCC com duas características, a energia do quadro e o pico da amplitude sonora e compararam com um enorme conjunto de dados previamente rotulado.

Os problemas de recuperação de música despertam certo interesse do mercado ao possibilitar o desenvolvimento de vários recursos para um produto. Um sistema de recuperação de MP3 usando MFCC para caracterizar o sinal de áudio foi proposto por Xin *et al.* (2014) e avaliado a partir de experimentos para recuperar músicas em chinês e em outros idiomas. A solução de recuperação dos autores é complementada por uma classificação e agrupamento feito anteriormente.

Seguindo esta abordagem de aprendizado de máquina, Bergstra *et al.* (2006) propuseram um algoritmo de aprendizagem baseado em uma versão multiclasse do ADABOOST (Bellet *et al.*, 2013). Os autores fizeram um estudo comparativo do desempenho de seu algoritmo com outras técnicas de aprendizado de máquina, como SVM e Redes Neurais

Artificiais sobre um problema de classificação de segmentos de músicas. É importante destacar que o desempenho do SVM é melhor quando apenas as características do MFCC são usadas e o comprimento dos segmentos é de cerca de trinta segundos de duração. McKinney e Breebaart (2003) realizou a classificação de arquivos de áudio usando análise discriminante quadrática (Kulis, 2012). O modelo usa uma mistura de gaussianas multidimensionais e, consequentemente, cada gênero tem seus próprios parâmetros de média e variância. Os autores também fizeram um estudo comparativo da representação de características e MFCC produziram melhores resultados para classificação.

O bom desempenho para identificar características únicas na voz, capacidade de distinguir instrumentos e tonalidades são atributos que credenciam o MFCC como características mais relevante a ser utilizada para representação de músicas (Xin *et al.*, 2014; P. W. Ellis, 2007). Além dos resultados de outros trabalhos já citados, esta conclusão também pode ser obtida através do estudo comparativo envolvendo quatro grupos de características de áudio (baseadas em sinal de baixo nível, MFCC, psicoacústicas e modelo auditivo), feito por McKinney e Breebaart (2003), e baseando-se na análise realizada sobre conjunto de características de baixo nível (*chroma* e timbre) e alto nível (sonoridade, batida e *tatum* média e variâncias), apresentada por Wolff e Weyde (2014).

O número de coeficientes que deve ser utilizado no MFCC é outra questão no processo de extração devido à informação específica do sinal que cada um representa. Vale a pena notar que, dependendo da tarefa em questão, diferentes subconjuntos dos MFCCs são adotados. Por exemplo, tornou-se habitual em muitas aplicações de processamento de música selecionar os primeiros 13 MFCCs porque eles são considerados portadores de informação discriminativa suficiente no contexto de várias tarefas de classificação e recuperação (Giannakopoulos e Pikrakis, 2014). Paralelamente ao objetivo principal do trabalho, McKinney e Breebaart (2003); Pampalket (2006); Lee *et al.* (2009); Xin *et al.* (2014); Wolff e Weyde (2014) fizeram a busca pelo número de coeficientes ideais e também propuseram 13 como o melhor número.

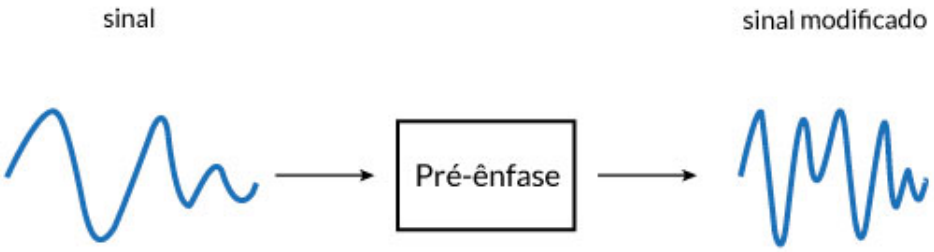
2.5.1 **Pré-ênfase**

A pré-ênfase objetiva identificar a distorção espectral que não traz informação adicional e eliminá-la através da aplicação de um filtro, (H), ocasionando um nivelamento no espectro. O nivelamento é representado pela ênfase que é aplicada sobre cada quadro do sinal, (k), destacando as informações mais relevantes para audição. O filtro é representado na Equação (2.4) e o nivelamento é ilustrado na Figura 4.

$$H(k) = 1 - \alpha.k^{-1}, \tag{2.4}$$

onde, recomendado por Picone (1993), $-1.0 \leq \alpha \leq -0.4$.

Figura 4 – Nivelamento do Sinal



Fonte: Elaborado pelo autor

Para as frequências em que o ouvido apresenta maior sensibilidade, acima de 1KHz, a pré-ênfase amplifica a área do espectro visando fornecer os aspectos perceptíveis mais importantes do espectro do som ao algoritmo de análise espectral (Markel e Gray, 1982).

2.5.2 **Janelamento**

O sinal sonoro é uma grandeza aleatória. Porém, durante curtos intervalos de tempo, assume-se que as características do sinal apresentam uma pequena variação. Baseando-se nisso, a segmentação por meio dos quadros com mesmo "tamanho" com duração entre 10 e 30 milisegundos (Picone, 1993). No entanto, esta segmentação acarreta em um problema: a quebra de ondulações em quadros vizinhos, gerando perda de informação.

A solução para este problema é obtida pela realização da divisão via sobreposições, fazendo com que um quadro inicie um pouco antes do anterior terminar. Desse modo, mesmo que um trecho importante seja cortado no fim de um quadro, ele estará inteiro no quadro seguinte.

A divisão do sinal traz também um segundo problema: a descontinuidade. Devido à partição, cada trecho começa e termina bruscamente, o que prejudica a extração de características. Então, torna-se necessário suavizar o quadro, multiplicando-o por uma função janela, mostrado na Figura 5. Existem diversos tipos de função janela. Uma das mais utilizadas no campo de processamento de sinais sonoros é a janela de Hamming (Wickramarachi, 2003), dada pela expressão:

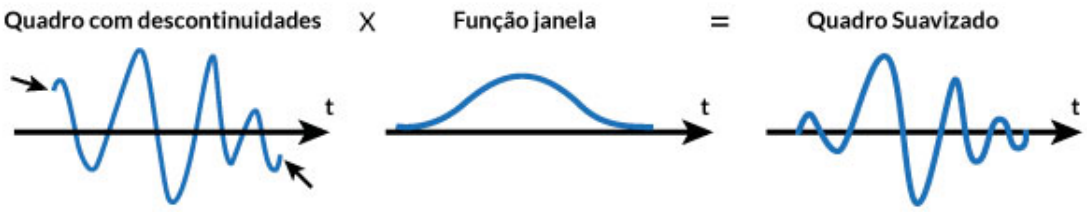
$$w(k) = 0.54 - 0.46 \cos\left(\frac{2\pi k}{K-1}\right), \tag{2.5}$$

onde K é o número total de quadros.

2.5.3 **Transformada de Fourier**

Não se utiliza o sinal no domínio do tempo devido à dificuldade em detectar diferenças entre seus quadros. Existem várias técnicas para a análise espectral do sinal do

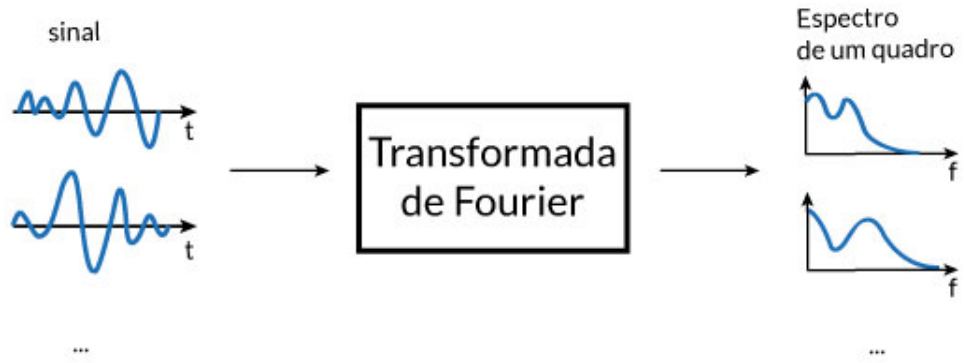
Figura 5 – Aplicação do janelamento



Fonte: Elaborado pelo autor

som, sendo que Transformada Discreta de Fourier (DFT) a mais usada e que consiste em determinar as componentes de frequência predominantes em um dado segmento de som. O alto custo computacional da DFT, $O(n^2)$, é um problema para esta abordagem, porém, na prática, para esta tarefa são usados algoritmos mais eficientes com custo computacional na ordem $O(nlogn)$, genericamente chamados FFT (Fast Fourier Transform) (Rabiner e Schafer, 1978).

Figura 6 – Transformada de Fourier

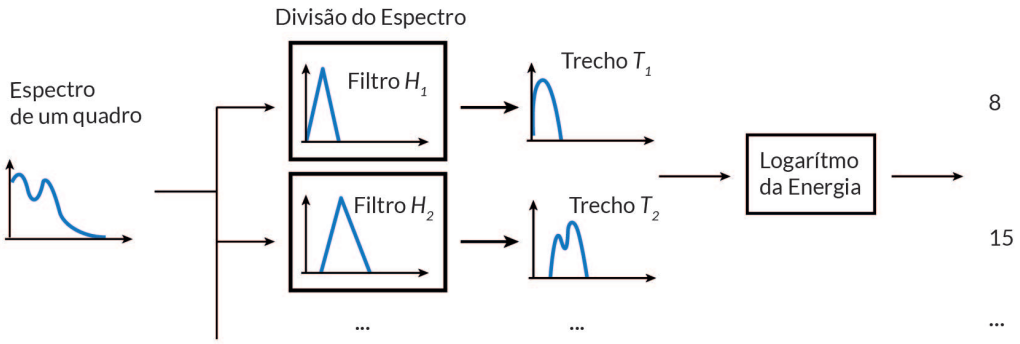


Fonte: Elaborado pelo autor

2.5.4 **MFCC**

O método para extração dos MFCCs (Mel-Frequency Cepstral Coefficients), ou coeficientes mel-ceptrais, faz a divisão do espectro por filtros, com os mesmos princípios da divisão do sinal em quadros. Ou seja, o começo de um filtro está sempre um pouco antes do final do filtro anterior, para reduzir a perda de informação. E, similares às janelas de Hamming, são aplicados filtros triangulares, para evitar o problema da descontinuidade. Em seguida, a extração dos valores é feita com o logaritmo da energia de cada trecho. A Figura 7 ilustra o processo.

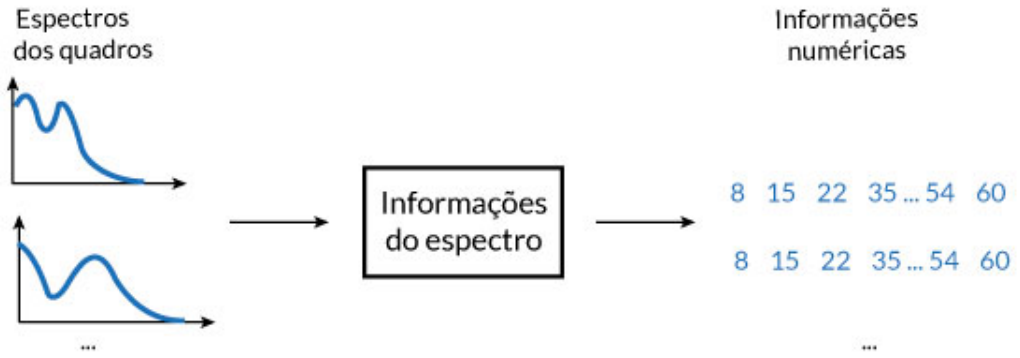
Figura 7 – Processo de extração dos coeficientes



Fonte: Elaborado pelo autor

Os coeficientes representam informações numéricas extraídas do espectro de cada quadro do sinal separadamente, ilustrados na Figura 8. Esta etapa faz a distinção entre as técnicas de extração, na qual aplica-se um banco de B filtros à potência espectral.

Figura 8 – Extração de valores do espectro



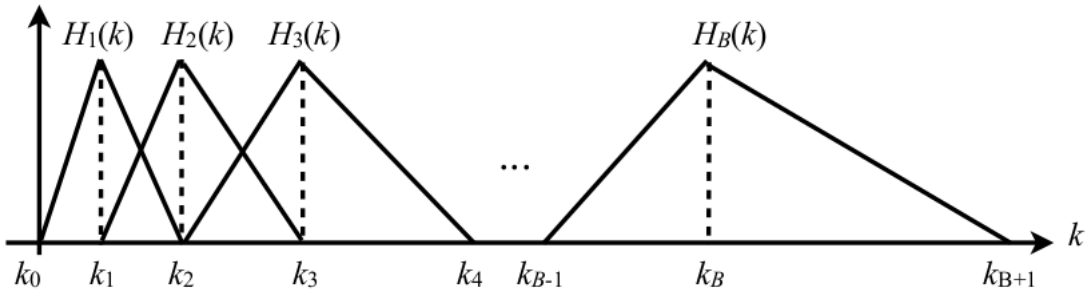
Fonte: Elaborado pelo autor

O banco de filtros é formado por filtros triangulares e sobrepostos, espaçados com larguras crescentes de acordo com a escala de frequência MEL, representada pela Equação (2.6), que, como foi mencionado, imita a resposta em frequência do sistema auditivo humano. A figura 9 ilustra o processo, sendo $H(k)$ o filtro aplicado sobre cada quadro k.

$$m(f) = 1125 \ln\left(1 + \frac{f}{700}\right). \tag{2.6}$$

Matematicamente, os filtros MEL são definidos por Gordillo (2013) seguindo a

Figura 9 – Filtros triangulares



Fonte: Elaborado pelo autor

resposta em frequência:

$$H_m[k] = \begin{cases} 0 & k < k[m-1] \\ \frac{2(k-k[m-1])}{(k[m+1]-k[m-1])(k[m]-k[m-1])}, & k[m-1] \leq k \leq k[m] \\ \frac{2(k[m+1]-k)}{(k[m+1]-k[m-1])(k[m+1]-k[m])}, & k[m] \leq k \leq k[m+1] \\ 0 & k > k[m+1] \end{cases} \tag{2.7}$$

Cada filtro calcula a média do espectro em torno da frequência central, e têm diferentes larguras de banda. Para determinar matematicamente os segmentos, parte-se das frequências extremas de cada quadro que são as frequências de corte do banco de filtros em Hz. Esses valores são usados para dividir o intervalo em $B + 1$ partes. Quanto maior é a frequência maior é a largura de banda, como já visto na Figura 9.

Em seguida, obtém-se o logaritmo de energia da saída de cada um dos filtros MEL.

$$\hat{S} = \ln\left(\sum_{m=0}^{\frac{N}{2}-1} S[k]H_m[k]\right), \quad 1 < m < M \tag{2.8}$$

onde $S[k]$ como o resultado da FFT aplicada sobre cada quadro e $H_m[k]$ são os filtros na escala mel, e N o número de amostras da FFT.

Finalmente, os coeficientes MFCC são obtidos aplicando a transformada do cosseno (DCT) ao logaritmo dos coeficientes de energia obtidos no item anterior:

$$c[n] = \sum_{m=0}^{M-1} \hat{S}[m] \cos\left(\frac{\pi n(m+0,5)}{M}\right), \quad 0 < n < M-1. \tag{2.9}$$

Nesta abordagem, adotou-se $M = 13$ e tem-se o vetor característico com a seguinte composição:

$$C_{mel} = c_0, c_1, c_2, ..., c_{12}.$$

2.6 QUANTIZAÇÃO VETORIAL

Tendo-se um universo de vetores multidimensionais, são estabelecidos vetores (com a mesma dimensionalidade) representantes desse universo. Esses vetores representantes são chamados centroides. O conjunto de todos os centroides é chamado de *codebook*. Os centroides são, em número, muito menor que o número de vetores que compõem o universo. Assim, pode-se discretizar qualquer vetor multidimensional em um, entre os centroides previamente conhecidos. Quantização é a conversão de um vetor de entrada em um código relacionado a vetores de mesma dimensão. A quantização vetorial é um dos métodos ideais para mapear uma enorme quantidade de vetores de um espaço para um número predefinido de grupos (Mahesha e Vinod, 2012).

O processo de quantização vetorial de um sinal consiste em codificação de blocos de amostras (ex. vetores), em vez de amostras individuais. Em geral, sobre um vetor multidimensional aleatório $x = x_0, x_1, ..., x_{k-1}$ atua um operador de codificação $Q(.)$, que associa a x um vetor de reconstrução $y = y_0, y_1, ..., y_{k-1}$ que está contido em um conjunto finito de N vetores discretos, chamado dicionário (codebook), onde N representa o número de níveis de dicionário. A quantização vetorial pode ser representada como: $y = Q(x)$. O vetor y é escolhido no dicionário e associado a x de tal forma que a distorção $d(x, y)$ introduzida pela quantização é mínima. Um exemplo de uma medida de distorção entre vetores é a distância euclidiana, e esta foi a medida usada nos experimentos abordados por este trabalho.

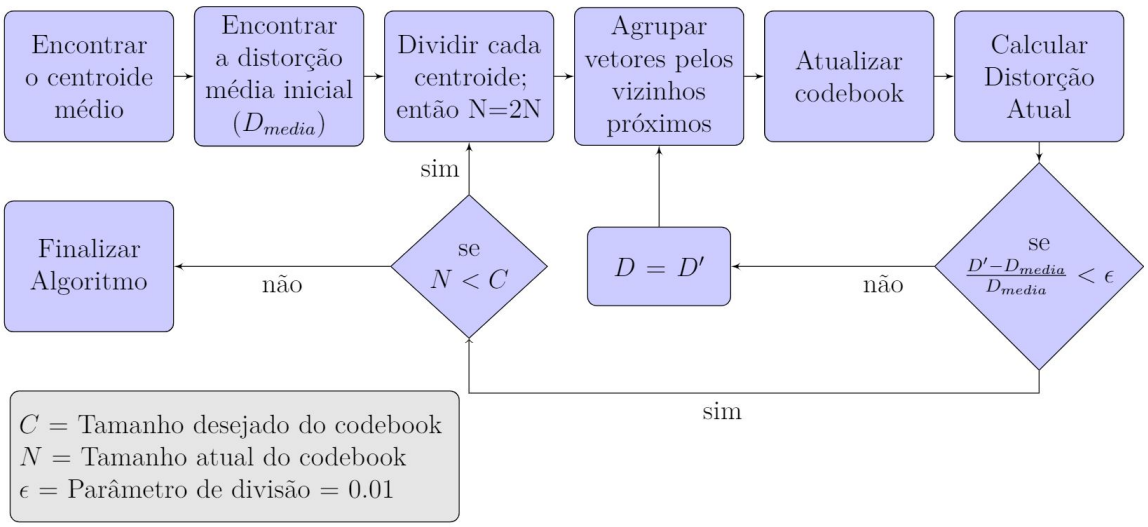
Um dos métodos para obter o codebook é o algoritmo Linde-Buzo-Gray (LBG), também conhecido como Algoritmo Generalizado de Lloyd (GLA), uma vez que é a generalização vetorial do algoritmo escalar proposto por Lloyd (1982). O algoritmo LBG ajusta iterativamente um codebook inicial de modo a reduzir a medida de distorção até que um mínimo local, por algum critério de convergência, seja atingido (Linde *et al.*, 1980). O fluxograma da Figura 10 demonstra o funcionamento do algoritmo.

A quantização vetorial é um método geralmente aplicado na compressão de sinais de fala (Kanawade e Gundal, 2017), imagens (Carafini, 2015) e vídeos (Esakkirajan *et al.*, 2009). No entanto, também encontra aplicações no campo do processamento de sinais, classificação e extração de dados. Neste trabalho, o objetivo é a compressão dos 13 MFCCs de uma forma que represente a música com a menor distorção.

2.7 PCA

No processo de extração dos MFCCs, utiliza-se a DCT. E este uso tem a capacidade de reduzir o número de coeficientes gerados após utilizar as técnicas de parametrização especificadas (Okida e Queiroz, 2006; Siqueira, 2012). A redução é feita através de uma propriedade da DCT conhecida como compactação da energia, concentrando os valores

Figura 10 – Quantização Vetorial usando LBG



Fonte: Elaborado pelo autor

mais significativos nas primeiras posições do vetor, abrindo uma grande possibilidade de reduzir a dimensionalidade do vetor característico e aumentar a eficiência computacional das tarefas.

Além da concentração da informação mais importante nos primeiros coeficientes, muitos dos dados existentes no codebook são redundantes (Loughran *et al.*, 2008) e, por isso, torna-se necessário aplicar um método para extrair as informações mais significativas de cada amostra. Isto pode ser alcançado através da Análise de Componentes Principais (PCA). Essencialmente, é feito uma transformação linear em componentes ortogonais de modo que sua variância permaneça constante, mas esteja concentrada nas dimensões inferiores. Após a transformação, a matriz de dados é composta por um vetor de coeficientes para cada amostra.

O PCA, um recurso muito utilizado para a extração de características em problemas de aprendizado de máquina, como compactação de dados, classificação, agrupamento, entre outros, vai ao encontro desta abordagem. Dado um conjunto de dados $X \in R^{N \times d}$ com N amostras e d dimensões, o PCA pode ser facilmente executado através da decomposições da matriz de covariância das amostras $1/NX^TX$ ou da decomposição em vetores singulares da matriz de dados X.

O primeiro trabalho avaliando o impacto da redução da dimensionalidade dos MFCCs usando PCA é apresentado por Trang *et al.* (2015). Eles usaram duas versões do método de extração de características do MFCC aplicadas ao problema de reconhecimento da fala, investigando a eficiência do modelo. Já Loughran *et al.* (2008), foram um pouco além e avaliou também a variação do número de coeficientes que seriam utilizados no MFCC, junto à redução do número de componentes usando PCA, aplicados ao problema de identificação de instrumentos existentes em uma música.

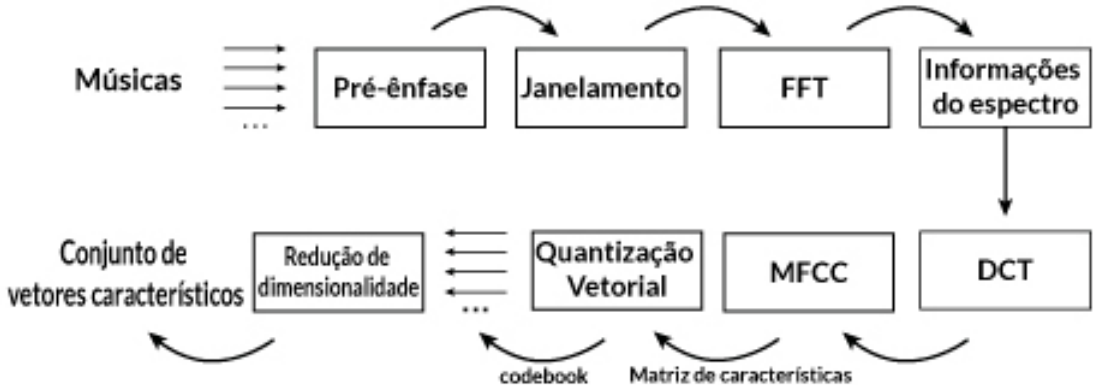
A extração de valores de características do MFCC de segmentos de áudio torna o volume de dados extremamente grande e instâncias diferentes de duração podem causar uma variação no espaço de dimensionalidade. Para reduzir o custo de extrair e manter todos os vetores com a mesma dimensionalidade, Pampalket (2006) mostrou que as informações extraídas ao longo de trinta segundos de uma música em particular são suficientes para representá-la. Além disso, as principais características das músicas são expostas durante a primeira metade e, geralmente, a segunda metade é composta pela repetição de uma parte considerável da primeira metade (Xin *et al.*, 2014). Neste sentido, para fins de extração de características, construímos um segmento de áudio MP3 com quinze ou trinta segundos de cada música. Além disso, para melhorar a qualidade do áudio de sonoro capturado, fiz-se um salto de 15 segundos após o início. Estatisticamente, muitos desses dados são redundantes e, por isso, é necessário empregar um método para extrair as informações mais significativas (Loughran *et al.*, 2008), e adotaremos a aplicação do PCA.

Essencialmente, a análise de componentes principais transforma os dados ortonormalmente de modo que a variação dos dados permaneça constante, mas esteja concentrada nas dimensões inferiores. A matriz de dados sendo transformada consiste em um vetor de coeficientes para cada amostra. Assim, existe agora uma matriz de dados para cada vetor do MFCC. A matriz de covariância da matriz de dados é então calculada. Os componentes principais para o conjunto de dados podem ser recuperados dos autovetores dessa matriz de covariância. Foi realizado uma análise dos componentes e concluiu-se que os 5 primeiros componentes concentram a maior parte da variância de todo o conjunto e, consequentemente, a informação mais significativa das amostras. Além destes, trabalhou-se com outros componentes de menor variância com intuito de mensurar o impacto da redução da dimensionalidade no modelo de classificação proposto. A dimensionalidade máxima de cada conjunto de dados criado após o PCA é limitada ao número de amostras de música. Para música com 30 segundos, o vetor de característica obtido do MFCC tem uma dimensionalidade igual a 1293 e por 15 segundos o tamanho é 647. O processo de criação de cada conjunto de dados foi feito com a extração de características seguida pela análise dos componentes principais.

Desenvolveu-se duas abordagens de classificação que necessitaram de duas aplicações diferentes do algoritmo para realizar o PCA, uma que analisa todas as músicas de uma única vez, como se fosse em modo batch, e uma versão que simula um ambiente de streaming, com os dados sendo reconhecidos de forma iterativa. Para esta versão, seguiu-se a proposta de Yang *et al.* (2018), no qual os autores apresentam um algoritmo capaz de computar os top-k componentes principais na configuração de streaming com $O(Bd)$ requisito de memória, em que B é o tamanho do conjunto de dados e d a dimensionalidade das amostras. A velocidade de convergência do novo algoritmo é muito mais rápida do que as abordagens existentes, principalmente devido ao uso eficaz de informações anteriores.

Todo processo de construção do vetor característico é realizado em sete etapas e estão ilustrados no fluxograma disposto na Figura 11.

Figura 11 – Processo de construção do vetor característico



Fonte: Elaborado pelo autor

3 MÉTODO PROPOSTO

Nesse capítulo será formulado a proposta de solução para o problema de aprendizado de métrica como uma solução offline, em que todas as amostras do conjunto de dados são conhecidas previamente, e online, de modo que as amostras são apresentadas de forma iterativa.

3.1 DISTÂNCIAS PARAMETRIZADAS E RELAÇÕES DE SIMILIRIDADES

Seja um conjunto de n pontos em um espaço d-dimensional definido como $\{x_i, i = 1, ..., n\} \subset R^d$. Considere também um conjunto de restrições propostas por um especialista que aponta a existência de um conjunto de similaridade entre pares S que pode ser particionado em k subconjuntos disjuntos: $S_1, S_2, ..., S_k$ cada um associado a um grupo. Assim sendo:

$$S : (x_i, x_j) \in S_l \rightarrow x_i \wedge x_j \text{ são similares}$$
$$S = S_1 \cup S_2 \cup ... \cup S_k$$

Por outro lado, se os pontos forem diferentes, tem-se:

$$D : (x_i, x_j) \in D_l \rightarrow x_i \wedge x_j \text{ são dissimilares}$$

Geralmente, o problema de aprendizado de métricas com relações de similaridade entre pares é formulado como um problema de otimização cujo objetivo é diminuir as distâncias de pares semelhantes enquanto aumenta a distância em relação aos diferentes. Essa abordagem envolve um número quadrático de termos na função objetivo e um problema de otimização quadrática. No entanto, esse problema pode ser reformulado como um problema de análise de grupo mais simples ao considerar a existência de duas propriedades de grafo:

$$\textit{transitividade}: \text{ se } (x_i, x_j) \text{ e } (x_j, x_k) \text{ são similares, então } (x_i, x_k) \text{ são similares.}$$
$$\textit{simetria}: \text{ se } (x_i, x_j) \text{ são similares, então } (x_j, x_i) \text{ são similares.}$$

Considere, também, uma medida de distância parametrizada entre dois pontos definidos como uma função de uma matriz A_{dxd} semidefinida positiva e simétrica (PSD):

$$d_A(x_i, x_j) = \| x_i - x_j \|_A^2 = (x_i - x_j)^T A (x_i - x_j) \tag{3.1}$$

com as seguintes propriedades:

$$d_A(x_i, x_j) \geq 0,$$
$$d_A(x_i, x_i) = 0,$$
$$d_A(x_i, x_j) = d_A(x_j, x_i),$$
$$d_A(x_i, x_j) \leq d_A(x_i, x_k) + d_A(x_k, x_j)$$

Nesse sentido, pode-se formular o problema de aprendizado métricas como um problema de análise de grupo considerando a relação de cada subconjunto S_l com um grupo. Então, deve-se resolver:

$$Min \sum_l \sum_{(x_i,x_j)\in S_l} \| x_i - x_j \|_A^2 \tag{3.2}$$
$$\text{sujeito a: } A \geq 0 \text{ (PSD)}$$

Ao considerar o caso em que a matriz de parâmetros é uma matriz diagonal, tem-se que aprender um vetor de parâmetros $w = [w_1, w_2, ..., w_d]$ ou um matriz diagonal equivalente W, cuja solução é equivalente a reescalonar o respectivo conjunto de pontos. Pode-se observar que, ao considerar o uso de uma matriz de identidade na equação (3.1), tem-se um conjunto de distâncias euclidianas. Caso contrário, se a matriz de covariância for adotada, tem-se um conjunto de distâncias de Mahalanobis. Para o caso mais geral, temos um conjunto de distâncias parametrizadas como uma função de uma matriz completa A. Para o caso diagonal, a restrição PSD é atendida se todos os componentes do vetor w não forem negativos, isto é: $w_i \geq 0$. Assim, a formulação da equação (3.2) para o caso diagonal pode ser reformulada como:

$$\sum_l \sum_{(x_i,x_j)\in S_l} \| x_i - x_j \|_A^2 = \sum_l \eta_l \sum_{x_i \in S_l} \| x_i - c_l \|_A^2 =$$
$$= \sum_l \eta_l \sum_{x_i \in S_l} (x_i - c_l)^T A (x_i - c_l) =$$
$$= \sum_l \eta_l \sum_{x_i \in S_l} w_1(x_{i1} - c_{l1})^2 + w_2(x_{i2} - c_{l2})^2 + ... + w_d(x_{id} - c_{ld})^2 \tag{3.3}$$
$$\text{sujeito a: } w_i \geq 0.$$

Essa equivalência é provada por Edwards e Cavalli-Sforza (1965) levando em conta o fato de que cada centroide dos clusters é calculado como a média dos vetores do seu respectivo cluster, ou seja:

$$c_l = \left(\frac{1}{\eta_l}\right) \sum_i x_i, \forall x_i \in S_l \tag{3.4}$$

Para a solução do problema apresentado na equação (3.2), usando uma matriz diagonal, Xing *et al.* (2002) propõem uma formulação relaxada que envolve a minimização

de uma função objetivo irrestrita com um termo de penalidade adicional:

$$Min \sum_{(x_i,x_j)\in S} \| x_i - x_j \|_A^2 - \log \sum_{(x_i,x_j)\in D} \| x_i - x_j \|_A, \tag{3.5}$$

onde S é o conjunto de pontos similares e D é o conjunto de pontos diferentes.

A solução proposta ao problema de aprendizado métricas é mais próxima à proposta por Schultz e Joachims (2004), que apresentaram uma extensão ao SVM (Cortes e Vapnik, 1995) e é baseada no cumprimento de um conjunto de restrições de relação comparativa. Essas relações comparativas têm a seguinte expressão envolvendo uma tripla de pontos:

$$x_i \text{ está mais próximo de } x_j \text{ do que } x_k.$$

Assim, pode-se deduzir que x_i é semelhante a x_j, mas não pode-se deduzir com certeza que x_i são x_k semelhantes, ou diferentes. Neste sentido, torna-se necessário modelar um número de $O(n^3)$ restrições, onde n é o número total de pontos, considerando a representação de cada subconjunto de triplas. Seja w o vetor de parâmetros associados a cada distância parametrizada. Então, pode-se modelar cada restrição como:

$$\forall i, j, k : d_w(x_i, x_k) - d_w(x_i, x_j) > 0. \tag{3.6}$$

Obviamente, podem existir nenhuma ou um conjunto de inúmeras soluções que satisfaçam o sistema de inequações apresentado. Neste sentido, pode-se nos basear na proposta de Coelho *et al.* (2017) em que propuseram uma solução semelhante ao SVM de margem flexível considerando a minimização da norma euclidiana do vetor de parâmetros w:

$$\begin{gathered} Min \frac{1}{2} \| w \|^2 + C. \sum_{i,j,k} \xi_{i,j,k} \\ \text{sujeito a: } \forall i, j, k : d_w(x_i, x_k) - d_w(x_i, x_j) \geq 1 - \xi_{i,j,k} \\ \xi, w \geq 0 \end{gathered} \tag{3.7}$$

onde ξ representa o vetor de variáveis de folga e C a constante de penalidade.

Para superar a desvantagem relacionada ao alto número de restrições, propõe-se, em nossa formulação, um conjunto de comparações entre cada ponto e o respectivo centroide de seu grupo reduzindo o número de restrições para $O(n)$. Além disso, na próxima seção mostra-se o uso de um método de relaxação baseado em uma versão estruturada do modelo Perceptron como técnica de solução. Com isso, evita-se ter de solucionar um problema mais complexo de programação quadrática.

3.2 ALGORITMOS PARAMETRIZADOS

Nesta subseção, descreve-se os algoritmos parametrizados aplicados nas fases de treinamento e teste da tarefa de classificação musical. Em vez de usar o algoritmo K-means

com distâncias euclidianas em um ambiente não supervisionado, empregamos o algoritmo Perceptron Estruturado que visa aprender uma métrica orientada a especialistas. Esse algoritmo é uma versão de amostra por amostra da margem máxima do K-means com informações secundárias que ajustam a métrica em uma configuração supervisionada de acordo com um conjunto de centroides predefinidos. Nesse sentido, chama-se esse algoritmo de *Maximal Margin Parameterized K-means* (MMP K-means). Para a fase de teste, empregamos o classificador baseado no centroide mais próximo com distâncias parametrizadas usando a métrica aprendida na fase de treinamento. Chamamos esse algoritmo de *Maximum Margin Parameterized Nearest Centroid* (MMP NCC).

O algoritmo k-means minimiza a função de distância, conhecida como intracluster, relacionada a um conjunto de pontos distribuídos no espaço euclidiano considerando um número de grupos previamente definidos. Mais especificamente, o algoritmo minimiza a soma dos quadrados das distâncias euclidianas de cada ponto com seu respectivo centroide, calculado como a média dos pontos de seu respectivo grupo.

A função de distância parametrizada do algoritmo k-means é construída com base na equivalência da soma das distâncias entre os vetores do mesmo grupo que compartilham uma relação de similaridade e a soma das distâncias intracluster. Assim, a única mudança necessária no algoritmo *Euclidean* K-means é na determinação do centro, em que, agora, as distâncias euclidianas parametrizadas para os respectivos centroides devem ser usadas.

O algoritmo *Nearest Centroid Classifier* (NCC) realiza a comparação das distâncias euclidianas de um novo ponto com o centroide de cada classe, classificando-o de acordo com o centroide vencedor. Por outro lado, o *Maximum Margin Parameterized Nearest Centroid* (MMPNCC) usa uma função de distância parametrizada para essa finalidade. Ao considerar um problema de classificação de duas classes com matrizes parametrizadas iguais, tem-se como hipótese de classificação uma função de decisão linear. Note que, ao escolher aprender duas diferentes matrizes de parâmetros, tem-se, como no caso geral de um discriminante de Fisher, duas matrizes de covariância diferentes, uma função de decisão quadrática.

De fato, Fisher (1936) propõe o primeiro algoritmo paramétrico para resolver o problema relacionado a classificação em reconhecimento de padrões. Para problemas de classificação binária com distribuição gaussiana multivariada, com os centros m_1 e m_2 e matrizes de covariância Σ_1 e Σ_2, a função de decisão pode ser expressa de acordo com a solução ótima de Bayes como a saída da função de sinal:

$$f(x) = \varphi\left((x - m_1)^T \Sigma_1^{-1}(x - m_1) - (x - m_2)^T \Sigma_2^{-1}(x - m_2) + \ln \frac{\mid \Sigma_2 \mid}{\mid \Sigma_1 \mid}\right) \tag{3.8}$$

De acordo com Cortes e Vapnik (1995), a estimativa desta função requer a determinação de um número quadrático de parâmetros, isto é, de ordem $O(d^2)$, onde d é a dimensão do problema. No entanto, quando o número de observações é reduzido em

comparação com o número de parâmetros, menor que $10.d^2$, essa estimativa não é mais viável. Neste sentido, Fisher (1936) recomenda o uso de uma função discriminante linear obtida a partir da Equação (3.8) quando as matrizes de covariância são iguais.

Seja w^* o vetor ótimo obtido a partir do processo de aprendizado métrico e W a matriz diagonal que representa os componentes de $w*$. Então, ao considerar um problema de classificação de duas classes com uma função de distância parametrizada com centroides m_1 e m_2, tem-se a seguinte função de decisão linear que representa o classificador MMPNCC:

$$f(x) = \varphi((x - m_1)^T W(x - m_1) - (x - m_2)^T W(x - m_2)) \tag{3.9}$$

3.3 APRENDIZADO DE MÉTRICAS COM PREDIÇÃO ESTRUTURADA: SOLUÇÃO OFFLINE

O problema de Predição Estruturada é caracterizado pela existência de um conjunto de treinamento $S = \{x_i, y_i, i = 1, ..., m\}$ formado por uma coleção de pares de entrada e saída, onde cada par é representado por um objeto estruturado $x(i)$ (entrada) e por um exemplo desejado $y(i)$ (saída). O modelo visa atender às restrições e correlações do conjunto estruturado de saída Y relativo ao conjunto de entrada X.

Pode-se formular o problema de aprendizado de métrica como um caso especial do modelo de Predição Estruturada no qual um conjunto de entrada X é formado por grafos completos e o conjunto de saída Y é formado por subgrafos de acordo com um conjunto de relações de similaridade fornecidas por um especialista.

O problema de inferência pode ser resolvido como um problema de minimização relacionado a uma função $S_x : Y(x) \rightarrow R$, que avalia cada saída em particular. Portanto, deve-se determinar: $y* = arg\{min_{y \in Y(x)} S_x(y)\}$. Esta classe de modelos pode ser parametrizada por um vetor w. Então, considerando $w.f(x, y) = S_x(y)$, tem-se a seguinte família linear de hipóteses:

$$H_w(x) = argmin_{y \in Y(x)}\{w.f(x, y)\}, \tag{3.10}$$

onde $(x, y) \in S = \{x(i), y(i), i = 1, ..., m\}$, e a saída, y, sendo sujeita a alguma função de restrição $g(x, y)$. O objetivo é estimar o vetor w tal que $H_w(x)$ mapeia qualquer saída desejada, y. Assim:

$$y(i) \approx argmin_{y \in Y(x(i))}\{w.f(x, y)\}, \tag{3.11}$$

Desta forma, considerando todas as possibilidades de saída, tem-se:

$$\forall i, \forall y \in y(i) : w.f(x(i), y(i)) \leq w.f(x(i), y) \tag{3.12}$$

A solução do problema de predição estruturada pode ser obtida por uma formulação de máxima margem de acordo com Taskar *et al.* (2005):

$$Min\frac{1}{2} \parallel w \parallel^2 \tag{3.13}$$
$$\text{sujeito a: } w.f_i(y_i) \leq min_{y \in Y(i)}\{w.f_i(y) + l_i(y)\}, \forall i,$$

onde $f_i(y) = f(x(i), y)$ e a função $l_i(y)$ é definida como uma função de perda que redimensiona o valor da margem geométrica. Considerando apenas o cumprimento das restrições, este problema pode ser resolvido com o uso de uma variação do algoritmo Perceptron Estruturado (Coelho *et al.*, 2012).

Agora, a regra de atualização, sem a função de perda, pode ser descrita como:

$$\begin{gathered}\text{para cada par}(x(i), y(i)), i = 1, ..., m, \text{faça} \\ \text{se}(w.f_i(y_i) > w.f_i(y^*))\text{então} \\ w \leftarrow w - \eta(f_i(y_i) - f_i(y^*)),\end{gathered} \tag{3.14}$$

onde $0 < \eta \leq 1$, é uma taxa de aprendizado constante e y^* é o melhor candidato computado para cada índice i por um algoritmo de otimização.

Usando uma analogia com a regra de atualização do vetor de parâmetro associado ao problema de aprendizado da métrica, pode-se dizer que $w.f_i(y_i)$ representa o valor da distância parametrizada fornecida pelo especialista, e $w.f_i(y^*)$ o valor da distância parametrizada calculada pelo algoritmo k-means. Esta função de distância pode ser calculada separadamente para cada cluster considerando a existência de m classes.

Nesse sentido, o problema de aprendizado de métricas pode ser resolvido calculando-se o vetor de parâmetros w. Considerando o fato que muitas soluções podem atender todas as restrições, também é possível adaptar o problema de predição estruturada, impondo uma margem para encontrar uma solução vetorial única. Isso equivale a minimizar a norma Frobenius da matriz diagonal W (Schultz e Joachims, 2004). Seguindo Coelho *et al.* (2012), propõe-se a seguinte formulação:

$$\max \gamma \tag{3.15}$$
$$\text{sujeito a: } w.(f_i(y^*) - f_i(y_i)) \geq \gamma.\|w\|, i = 1, ..., m$$

onde γ é o valor da margem.

A nova regra de atualização pode ser descrita como:

$$\begin{gathered}\text{para cada par } (x(i), y(i)), i = 1, ..., m, \text{faça} \\ \text{se } (w.f_i(y_i) > w.f_i(y*) - \gamma \parallel w \parallel) \text{ então} \\ w \leftarrow w(1 - \frac{\eta\gamma}{\parallel w \parallel}) - \eta(f_i(y_i) - f_i(y*))\end{gathered} \tag{3.16}$$

A abordagem apresentada até agora pode ser descrita como um processo de correção em batch que considera o erro total intracluster para cada classe em que o vetor w é atualizado usando o método gradiente. No entanto, considerando o erro total, o processamento em batch é responsável por grandes correções no vetor w, tornando o método gradiente instável e exigindo maior controle da taxa de aprendizado. Para superar esse problema, é possível considerar a regra de atualização para cada erro individual de acordo com o esquema de rótulos fornecido pelo especialista.

Em outras palavras, se a distância parametrizada entre uma amostra x_i e seu respectivo centroide c_l for maior que a distância do centroide de melhor candidato c_k, onde $k = arg\{min_{j \neq i} \|x_i - c_j\|_w\}$ então é feita a correção do vetor de parâmetros w para forçar o preenchimento dessa restrição. Então, ao usar a distância parametrizada entre dois vetores, $dw(x_i, c_l) = (x_i - c_l)^T W (x_i - c_l)$, tem-se que resolver o seguinte problema de maximização de margem:

$$\begin{gathered} \min \frac{1}{2}\|w\|^2 + C.\sum{}_i \xi_i \\ \text{sujeito a: } d_w(x_i, c_k) - d_w(x_i, c_l) \geq 1 - \xi_i, i = 1, ..., n, \\ \xi, w \geq 0 \end{gathered} \tag{3.17}$$

onde ξ representa o vetor de variáveis de folga e C é a constante de penalização.

Para evitar à solução de um problema de otimização quadrática, o problema de maximização da margem da equação (3.17) pode ser resolvido como:

$$\begin{gathered} \max \gamma \\ \text{sujeito a: } d_w(x_i, c_k) - d_w(x_i, c_l) + \lambda\alpha_i \geq \gamma. \parallel w \parallel, i = 1, ..., n, \\ \alpha, w \geq 0 \end{gathered} \tag{3.18}$$

onde $\lambda = \frac{1}{C}$ representa o inverso da constante de penalização.

Essa formulação permite o processo de relaxação da margem flexível semelhante à penalização quadrática do vetor ξ, proposto por Villela *et al.* (2016). Assim, a nova regra de atualização segue:

$$\begin{gathered} \text{para cada par } (x_i, c_l), i = 1, ..., n, \text{ faça} \\ \text{se } d_w(x_i, c_k) - d_w(x_i, c_l) + \lambda\alpha_i < \gamma.\|w\| \text{então} \\ w \leftarrow w(1 - \frac{\eta\gamma}{\|w\|}) - \eta(d_w(x_i, c_k) - d_w(x_i, c_l)), \\ \alpha \leftarrow \alpha(1 - \frac{\eta\gamma}{\|w\|}), \alpha_i \leftarrow \alpha_i + \eta \end{gathered} \tag{3.19}$$

A solução da equação (3.18) começa com um valor de margem igual a zero. Após a primeira execução do Perceptron Estruturado com margem, existe uma possibilidade

maior de que a margem de parada não seja a máxima. Essa margem é considerada como a margem com menor valor entre as classes, assim:

$$\gamma^t = min_{i=1,\ldots,m}\{\gamma_i\} \tag{3.20}$$

A nova margem para uma próxima iteração do algoritmo será o dobro da margem de parada da iteração anterior, ou seja:

$$\gamma^{t+1} \leftarrow 2.\gamma^t \tag{3.21}$$

Para o novo problema, pode-se usar o vetor final w da iteração anterior como solução inicial. A margem de parada é aumentada até que a solução não seja viável. Nesse caso, um processo de aproximação baseado em uma pesquisa binária pode ser usado para encontrar a margem máxima de parada permitida.

Para um esquema de rótulos predefinidos por um especialista, a equação (3.18) representa o problema inverso relacionado à análise de grupos. Ou seja, "qual deve ser uma métrica adequada que atenda às restrições intragrupo?" Por outro lado, se a métrica for predefinida, a posição dos centroides e, consequentemente, o esquema de rótulos será computada usando o mesmo conjunto de restrições baseadas nas comparações de distâncias.

Para avaliar esta abordagem, na seção de experimentos, objetiva-se comparar o uso do algoritmo para aprendizado de métricas MMPNCC, demonstrado na equação (3.9) com um algoritmo do estado da arte com *kernels* linear, polinomial e gaussiano, para solucionar problemas de classificação musical.

3.4 APRENDIZADO DE MÉTRICAS COM PREDIÇÃO ESTRUTURADA: SOLUÇÃO ONLINE

No algoritmo para a solução proposta ao problema de classificação com aprendizado de métricas online, os vetores w e α devem ser atualizados da mesma forma que a versão offline, de acordo com a equação (3.19), pois são implementados em modo iterativo, ou amostra por amostra. No entanto, os centroides possuem outra regra para serem atualizados. Inicialmente calcula-se os valores do centroide como a média móvel considerando o conjunto de rótulos corretos dos vetores de entrada. Isso é:

$$c_l = (\frac{1}{\eta_l})\sum_i x_i, \forall x_i \in S_l \tag{3.22}$$

onde η representa a taxa de aprendizado.

Após algumas iterações, os centroides são atualizados de acordo com o algoritmo *Learning Vector Quantization* (LVQ) (Vishnupriya e Meenakshi, 2018) adaptado a um conjunto de distâncias parametrizadas. Considere $J(w) = \sum_l \sum_{i|x_i \in S_l}(x_i - c_l).w.(x_i - c_l)^T$,

como a função de erro parametrizada relacionada ao problema de análise de grupo. Então, ao tomar o gradiente estocástico de J em relação a c_l considerando a amostra x_i, tem-se a seguinte regra de atualização para o centroide vencedor:

$$\begin{aligned} c_l &\leftarrow c_l + \eta w(x_i - c_l) \text{ se um erro não ocorreu,} \\ c_l &\leftarrow c_l - \eta w(x_i - c_l) \text{ se um erro ocorreu.} \end{aligned} \tag{3.23}$$

Essa regra de atualização aproxima o centroide vencedor em relação ao vetor de amostra quando não há erro e afasta-o caso contrário. Finalmente, o parâmetro margem é atualizado seguindo a regra:

$$\begin{aligned} \gamma &\leftarrow \gamma + \eta(1 - \nu) \text{ se um erro não ocorreu,} \\ \gamma &\leftarrow \gamma - \eta(1 - \nu) \text{ se um erro ocorreu.} \end{aligned} \tag{3.24}$$

Para avaliar o algoritmo online, propõe-se experimentos com dois conjuntos de dados, um artificial e construído pelos autores e outro, denominado GTZAN, já reconhecido na literatura. Para avaliar e validar a precisão do modelo online, a classificação da música foi realizada inicialmente em modo batch, produzindo uma solução ótima offline e inalterada. Em seguida, os resultados obtidos online foram comparados considerando os resultados da versão offline. Uma análise analítica da convergência foi feita reportando o *regret* médio relacionado à precisão definida pelos erros acumulativos médios. Também relata-se as convergências dos vetores da métrica e dos centroides em relação aos respectivos vetores da solução offline.

4 EXPERIMENTOS E RESULTADOS

Neste trabalho, utiliza-se dois conjuntos de dados diferentes. Um já conhecido devido a várias pesquisas na área de aprendizado de máquina e o outro foi construído artificialmente. O primeiro, chamado GTZAN[1], permite comparar, mesmo que indiretamente, a acurácia da nossa abordagem em relação a trabalhos significativos na área de aprendizado musical. O último, chamado MUSIC, tem como objetivo mostrar que o processo de similaridade musical usando aprendizado de métricas pode ser invariável com o conjunto de treinamento ou, em outras palavras, com a preferência do cliente.

O conjunto de dados GTZAN aparece em pelo menos 100 trabalhos publicados e é o conjunto de dados público mais usado para avaliação em pesquisa de aprendizado de máquina para classificação de gêneros musicais (Sturm, 2013). O conjunto de dados original consiste em 1.000 segmentos de áudio, cada um com 30 segundos de duração, divididos sobre 10 gêneros (Blues, Clássica, Country, Disco, Hip Hop, Jazz, Metal, Popular, Reggae e Rock), sendo cada um representado por 100 segmentos. Este conjunto original foi utilizando para criar uma variação com músicas contendo 15 segundos de duração.

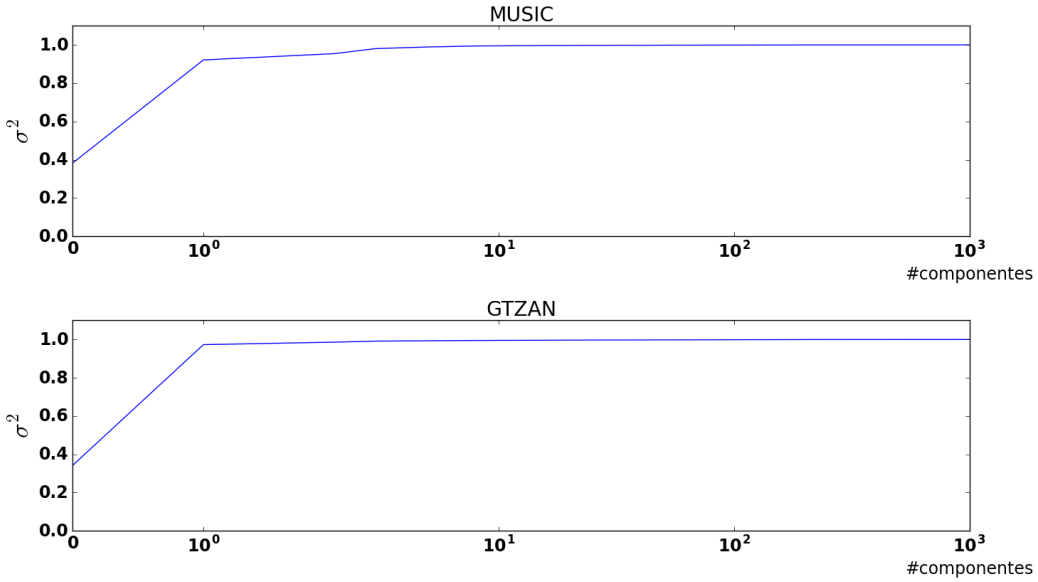
O conjunto MUSIC foi dividido em três subconjuntos contendo 250, 500 e 1000 segmentos de áudio. Todos os subconjuntos têm as músicas igualmente distribuídas em 5 gêneros (Rock, Clássica, Jazz, Eletrônica e Samba), com cada segmento de áudio tendo 30 segundos que foram extraídos após os primeiros 15 segundos de cada música, como sugerido por Pampalket (2006). Para conjuntos de dados com 1000 instâncias, também construiu-se subconjuntos com 15 segundos de duração.

Todo o processo de extração de características e construção do vetor característico foi aplicado sobre todos os conjuntos, junto às suas variações. De cada segmento, os primeiros 13 MFCCs foram extraídos porque esta informação é suficiente para a análise discriminativa no contexto das tarefas de classificação musical conforme a conclusão dos trabalhos referenciados na subseção sobre MFCC.

Para estudar o impacto da redução de dimensionalidade, analisou-se a variância explicitada pelos componentes. Selecionou-se os mais relevantes e definimos, empiricamente, alguns tamanhos de dimensão que seriam usados em cada um dos subconjuntos, com suas variações. A dimensionalidade máxima de cada subconjunto criado após o PCA é limitada ao número de instâncias. Originalmente, o segmento de áudio com 30 segundos, o vetor de características obtido tem uma dimensão com o tamanho igual à 1293 e com 15 segundos seu tamanho é 647. A Figura 12 apresenta a variância dos componentes após a etapa da quantização dos vetores característicos dos dois conjuntos de músicas que foram utilizados nos experimentos.

[1] http://marsyasweb.appspot.com/downloads/data_sets

Figura 12 – Variância acumulada dos componentes dos vetores característicos



Fonte: Elaborado pelo autor

As variações do conjunto de dados GTZAN estão dispostas sobre cinco subconjuntos que contêm 1000 músicas com 30 segundos e 5, 50, 100, 250, 500 e 1000 dimensões e também quatro subconjuntos com 1000 músicas com 15 segundos e 5, 50, 100, 250 e 500 dimensões. Em um processo semelhante, o conjunto de dados MUSIC originou subconjuntos com 250 músicas com 30 segundos e 50, 100 e 250 dimensões; 500 músicas com 30 segundos e 50, 100, 250, 500 dimensões; e 1000 músicas com 15 e 30 segundos com dimensões iguais ao GTZAN.

A variação do número de músicas, tempo de duração e dimensionalidade permite mensurar o comportamento da abordagem proposta em diversos cenários e avaliar seu desempenho. Um resumo dos dois conjuntos de dados é mostrado nas Tabelas 1 e 2 com o número de amostras, a duração, em segundos, de cada amostra, as dimensões de cada subconjunto e o número de classes.

Tabela 1 – MUSIC

amostras	**duração (s)**	**dimensões**	**classes**
250	30	50 - 100 - 250	5
500	30	50 - 100 - 250 - 500	5
1000	15	5 - 50 - 100 - 250 - 500	5
1000	30	5 - 50 - 100 - 250 - 500- 1000	5

Tabela 2 – GTZAN

amostras	**duração (s)**	**dimensões**	**classes**
1000	15	5- 50 - 100 - 250 - 500	10
1000	30	5- 50 - 100 - 250 - 500- 1000	10

Fonte: Elaborado pelo autor

4.1 AVALIAÇÃO DA SOLUÇÃO OFFLINE

No primeiro cenário, representado nas Tabelas 3 e 4, utilizou-se as variações do MUSIC com 250 e 500 músicas para avaliar a etapa de treinamento. Nesta etapa, utilizo-se o conjunto de dados com intuito de ratificar a importância do aprendizado com informações adicionais em relação à métrica Euclidiana. Neste sentido, comparou-se o resultado considerando o algoritmo K-means com métrica Euclidiana (Euclidean K-means), K-means parametrizado (MMP K-means). Os resultados apresentados representam a média e a variância de 20 repetições.

Tabela 3 – Desempenho de Treinamento - MUSIC com 250 músicas

	Euclidean K-means		**MMP K-means**	
dimensões	μ	σ^2	μ	σ^2
50	32,52	5.461	55,00	6.514
100	37,14	5.002	66,14	5.437
250	34,30	4.907	60,10	3.264

Tabela 4 – Desempenho de Treinamento - MUSIC com 500 músicas

	Euclidean K-means		**MMP K-means**	
dimensões	μ	σ^2	μ	σ^2
50	37,60	5.676	41,20	6.902
100	35,60	4.283	51,60	5.311
250	34,00	4.564	62,00	2.880
500	34,20	5.112	63,60	4.935

Fonte: Elaborado pelo autor

Aqui, o erro é considerado quando a diferença das respectivas distâncias em relação ao centroide correto do gênero possuir um valor negativo. O algoritmo de linha de base k-means com distâncias euclidianas tem o efeito de *underfitting* e consequentemente não pode aprender uma função de decisão correta. Além disso, pode-se observar que o algoritmo de aprendizado métrico não apresenta o efeito de *overfitting*. Nos experimentos ficou evidenciado que o aprendizado de métrica contribuiu efetivamente para melhora dos resultados na identificação dos grupos.

As tabelas comparativas apresentam os resultados representando a capacidade de generalização desta abordagem comparada ao algoritmo SVM clássico com margem flexível usando *kernel* linear, polinomial e gaussiano. Para análise estatística, todos os experimentos de teste foram realizados com 20 execuções para cada conjunto de dados, em todas as suas variações de dimensões. Para cada iteração, selecionou-se os dados aleatoriamente de forma balanceada, 50% dos dados para o conjunto de treinamento e os outros 50% dos dados para o conjunto de testes. Adotou-se uma estratégia um contra todas para construir a função de decisão de multi-classificação. Para os algoritmos de SVM, a constante de penalização, C, variou por 0.1, 0.2, .5, 1.0, 1.2 e 1.5 e os valores reportados são calculados como uma média. Os resultados apresentados representam os

valores médios para a porcentagem de associações corretas dos modelos e a variância obtida em cada experimento.

As Tabelas 5 e 6 apresentam os resultados da classificação obtidos, respectivamente, pelos conjuntos de dados MUSIC com 250 e 500 músicas em comparação com o algoritmo SVM com os três tipos de *kernel*. Considerando a variação na dimensionalidade do problema, o algoritmo parametrizado MMP NCC produz resultados superiores, principalmente quando os subconjuntos apresentam dimensões menores.

Tabela 5 – Resultados de Teste - MUSIC com 250 músicas

	Linear SVM		**Polynomial SVM**		**Gaussian SVM**		**MMP NCC**	
dimensões	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
50	64,79	10.154	60,15	9,11735	48,48	12,8856	68,51	4.159
100	66,26	10.408	59,14	8,50950	58,20	11,9119	67,54	3.489
250	66,20	10.744	58,23	8,59202	62,52	12,1056	70,01	3.157

Tabela 6 – Resultados de Teste - MUSIC com 500 músicas

	Linear SVM		**Polynomial SVM**		**Gaussian SVM**		**MMP NCC**	
dimensões	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
50	50,81	9.558	62,62	5,525	38,36	12,145	69.26	2.454
100	64,00	7.869	62,52	5,868	46,48	13,201	67.82	2.720
250	64,46	8.699	61,63	6,471	57,46	10,781	67.21	3.122
500	64,63	8.855	61,55	6,460	58,02	10,095	68.94	2.357

Fonte: Elaborado pelo autor

Além dessa maior precisão, o MMP NCC também possui um valor menor para a variância, o que demonstra a maior estabilidade do classificador. Neste cenário, também podemos considerar que nosso algoritmo é invariante à redução de dimensionalidade, obtendo o valor de precisão máxima usando apenas os 50 primeiros componentes principais do vetor de características, como pode ser visto na tabela 6.

Após avaliar o desempenho da nossa abordagem para as etapas de treinamento e teste para 250 e 500 músicas, apresentaremos nas próximas tabelas os resultados obtidos para o conjunto de dados MUSIC e GTZAN contendo 1000 músicas, nos cenários em que suas músicas possuem 15 e 30 segundos. Além disso, a menor dimensão representa os componentes estatisticamente mais relevantes, resultante da análise de componentes principais.

As Tabelas 7 e 8 apresentam os resultados obtidos para o conjunto de dados MUSIC com 1000 músicas, com cada uma com 15 e 30 segundos respectivamente. Com o aumento do tamanho do conjunto de treinamento, não observamos uma grande melhora na acurácia do teste em relação 6. No entanto, os resultados para os segmentos com 30 segundos, mostrados na tabela 8, mostram uma ligeira melhoria nos resultados de precisão e um pouco mais estáveis. Além disso, precisamos destacar nosso desempenho

para o subconjunto com 5 dimensões. Os componentes principais apresentam desempenho semelhante a todos os outros, evidenciando que a redução de dimensão é um recurso muito pertinente a ser aplicado ao vetor característico composto por MFFCs.

Tabela 7 – Resultados de teste - MUSIC com 1000 músicas e 15s

	Linear SVM		**Polynomial SVM**		**Gaussian SVM**		**MMP NCC**	
dimensões	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
5	38,87	2,118	36,86	2,525	33,73	1,592	64.86	1.770
50	34,69	13,786	62,19	5,164	31,43	15,420	66,65	1,467
100	49,77	12,102	61,31	5,119	37,38	14,568	66,06	1,540
250	62,88	7,463	60,65	5,670	44,39	12,533	66,58	1,524
500	62,54	7,438	60,60	5,718	45,94	12,320	66.49	1.404

Tabela 8 – Resultados de teste - MUSIC com 1000 músicas e 30s

	Linear SVM		**Polynomial SVM**		**Gaussian SVM**		**MMP NCC**	
dimensões	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
5	40,47	1,575	38,61	2,745	36,06	1,394	68,16	2,217
50	37,26	14,839	62,96	5,026	33,62	16,198	68,74	1,383
100	52,82	11,247	62,64	5,725	40,98	15,411	67,93	1,860
250	65,09	8,424	62,61	6,242	50,08	12,544	67,75	1,506
500	65,32	8,273	62,43	6,426	55,57	10,721	68.01	2.140
1000	65,09	8,383	62,18	6,372	55,40	11,393	68,30	1,370

Fonte: Elaborado pelo autor

As Tabelas 9 e 10 apresentam os resultados obtidos para o conjunto de dados GTZAN com 1000 músicas com segmentos com 15 e 30 segundos respectivamente. Esses resultados apresentam o desempenho dos algoritmos para um problema com mais classes, o que dificulta a predição e, consequentemente, impacta o desempenho do classificador. O SVM permaneceu abaixo do desempenho do algoritmo MMPNCC, como nos outros problemas. É importante ressaltar que o SVM Linear supera os kernels polinomial e gaussiano, destacando que o problema de aprender a similaridade musical tem uma solução melhor com um classificador baseado em uma hipótese linear. O SVM apresentou um desempenho pior para as dimensões menores, seguindo o comportamento dos resultados anteriores, mas o algoritmo parametrizado permanece com pouca variação e, novamente, obteve melhores resultados com as dimensões menores. O subconjunto com os componentes principais não obteve o melhor desempenho médio, porém, nota-se pela variância, tendeu ao melhor resultado.

Finalizando, relatou-se o desempenho de classificação do algoritmo parametrizado, exibindo as matrizes de confusão dos dois conjuntos. Na Figura 14 tem-se o melhor desempenho do algoritmo MMPNCC quando aplicado sobre o conjunto de dados GTZAN com 1000 músicas e 30s. Na Figura 13 tem-se o mesmo melhor resultado quando aplicado ao conjunto de dados MUSIC com 1000 músicas e 30s.

Tabela 9 – Resultados de teste - GTZAN com 1000 músicas e 15s

	Linear SVM		**Polynomial SVM**		**Gaussian SVM**		**MMP NCC**	
dimensões	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
5	16,49	1,835	17,53	1,884	21,74	1,366	62.29	3.477
50	34,89	19,291	56,64	9,205	17,67	22,439	62,23	3,981
100	51,48	11,231	57,03	9,690	29,23	19,164	61,98	3,027
250	57,01	11,272	55,68	9,917	39,93	17,552	62,37	3,257
500	56,92	11,378	56,01	10,277	41,64	16,154	61,77	2,730

Tabela 10 – Resultados de teste - GTZAN com 1000 músicas e 30s

	Linear SVM		**Polynomial SVM**		**Gaussian SVM**		**MMP NCC**	
dimensões	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
5	17,68	1,693	17,66	2,321	18,57	1,085	61,04	4,140
50	40,25	17,448	56,86	8,238	24,34	22,208	63,11	2,452
100	58,43	10,876	57,76	10,067	36,62	17,905	63,46	3,149
250	58,10	11,638	56,52	9,783	52,15	13,130	62,23	2,519
500	58,29	11,987	56,44	10,744	56,67	12,109	61,58	3,308
1000	57,70	11,962	56,27	11,259	56,34	11,710	60,85	2,800

Fonte: Elaborado pelo autor

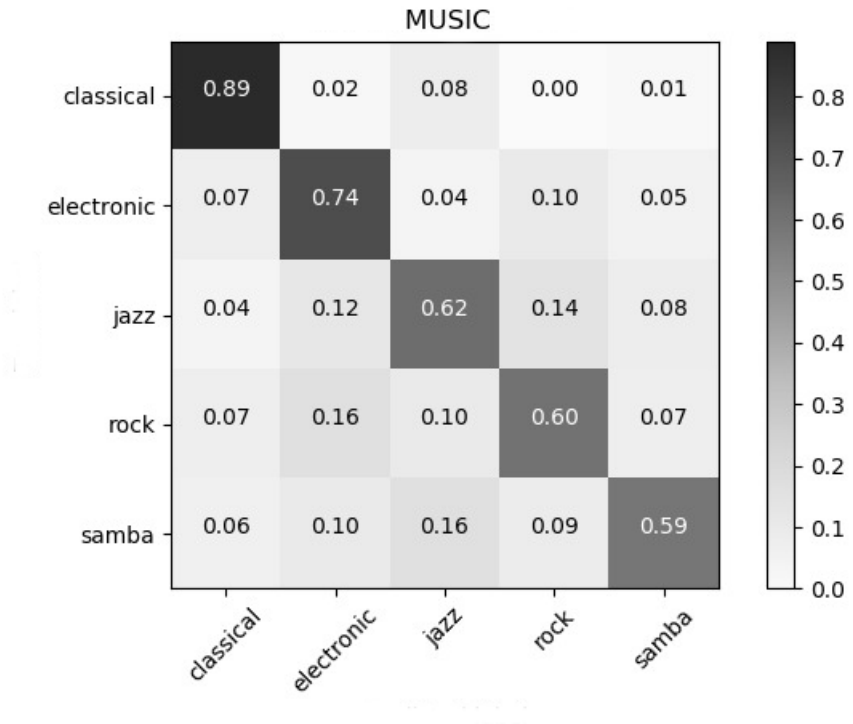
Pode-se observar que os valores de acurácia e a distribuição de erros ao longo dos diferentes gêneros estão muito próximos, exceto pelo gênero “Clássica” que apresenta uma estrutura musical bem formada, garantindo a estabilidade ao algoritmo quando aplicado a diferentes conjuntos de dados. No entanto, a medida média de desempenho, cerca de 63,5% para 10 gêneros e 69% para 5 gêneros, apontou que o problema de classificação musical é um problema difuso onde os grupos de gênero não possuem limites distintos dificultando uma melhor classificação.

Isso pode ser demonstrado observando os resultados de ambas as matrizes de confusão, nas quais os maiores valores de erros de classificação ocorrem quando os gêneros possuem uma estrutura musical mais semelhante. Por exemplo, no conjunto de dados GTZAN, pode-se destacar o erro de classificação entre gêneros rock e metal. Isso é compatível com o fato de que o metal pode ser considerado um subgênero do rock. Por outro lado, no conjunto de dados MUSIC, a classificação errônea entre os gêneros samba e jazz pode ser considerado pelo fato de que esses gêneros contêm vários elementos musicais comuns.

4.1.1 **Análise Comparativa**

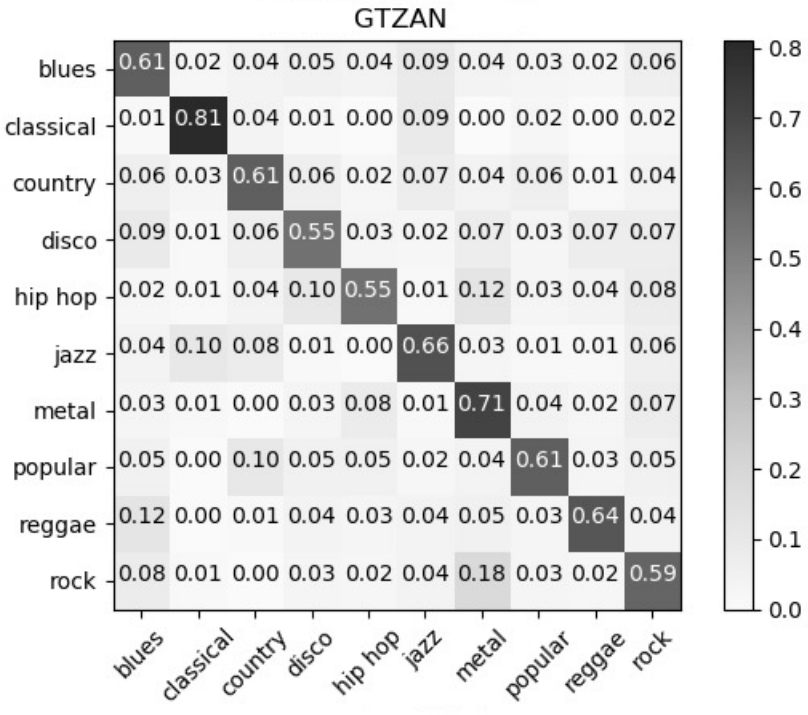
Os resultados obtidos com o conjunto de dados da GTZAN são comparavéis com os resultados encontrados na literatura, sendo superiores na maioria dos trabalhos referenciados e com uma proposta que utiliza um conjunto reduzido de treinamentos. O desempenho do modelo de aprendizado de métrica usando 50% dos dados para treinamento

Figura 13 – Matriz de Confusão MUSIC



Fonte: Elaborado pelo autor

Figura 14 – Matriz de Confusão GTZAN



Fonte: Elaborado pelo autor

e 50% para testes indica que, com o aumento do número de restrições de treinamento, o modelo tende a evoluir melhor seu poder de generalização quando comparado ao SVM e outros classificadores referenciados.

A fim de ter resultados com possibilidade de comparação a outras abordagens da área de classificação de músicas, também submeteu-se os dois conjuntos de dados apenas com os componentes principais a um processo de classificação com validação cruzada, dividindo por 10 partes iguais, sendo 9 delas para treinamento e 1 para teste, com 20 repetições. Logicamente, sabe-se que não é uma comparação justa ao olhar apenas os percentuais de acertos e ignoro-se o processo de construção de características e métodos para treinamento e teste, porém apresenta-se os resultados e, em sequência, alguns trabalhos que utilizaram GTZAN e MFCC para construir o vetor de características. Para GTZAN obteve-se $(76,67\% \pm 6,965506)$.

Embora o desenvolvimento do aprendizado de similaridade musical seja realizado em diferentes contextos com abordagens distintas, apresentou-se aqui uma análise comparativa dos resultados obtidos com alguns trabalhos que também utilizam o conjunto de dados GTZAN. Tzanetakis e Cook (2002) relataram um estudo de análise de características em um processo de classificação musical usando o conjunto de dados GTZAN. Eles usaram como características a textura timbral, conteúdo rítmico e conteúdo de tom. Os resultados da classificação são calculados usando uma técnica de validação cruzada de dez vezes, em que o conjunto de dados a ser avaliado é particionado aleatoriamente para que 10% seja usado para testes e 90% seja usado para treinamento. Usando os conjuntos de características propostos, os autores obtêm 61% de precisão para dez gêneros musicais.

Li *et al.* (2003) realizam uma comparação do desempenho de vários classificadores sobre um cenário de classificação musical usando o conjunto de dados GTZAN com vários subconjuntos de recursos. Os valores de precisão também são calculados por meio da técnica de validação cruzada de dez vezes. Os resultados obtidos usando características do MFCC foram: SVM um-contra-um: 58.40%, SVM um-contra-todos: 58.10%, Modelos de Mistura Gaussiana (GMM): 46.40%, Análise Linear Discriminante (LDA): 55.50% e k-vizinhos mais próximo (KNN): 53.70%.

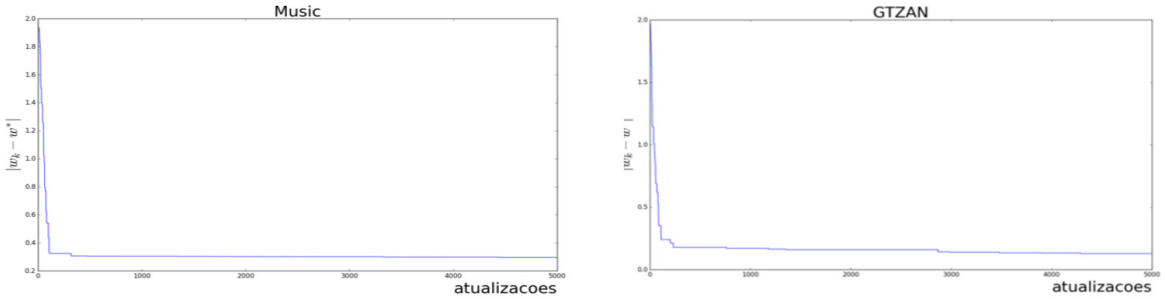
Atualmente, muitos autores consideram o uso de abordagens baseadas em *Deep Learning* como o estado da arte em diversas áreas de aprendizado de máquina, como por exemplo, em reconhecimento de imagem e fala. No entanto, pode-se observar no trabalho de Vishnupriya e Meenakshi (2018) que os melhores resultados de teste alcançados para o conjunto de dados GTZAN usando MFCC como características é de cerca de 47%, usando 80% dos dados para treinamento e 20% para teste.

4.2 AVALIAÇÃO DA SOLUÇÃO ONLINE

Durante os experimentos, os dois conjuntos de dados com os componentes principais da abordagem offline foram utilizados para avaliar a taxa de convergência do algoritmo online em relação à solução do algoritmo offline. Os experimentos foram realizados para simular o comportamento do usuário de um aplicativo de reprodução de música via streaming. Para esta simulação, o conjunto de dados da música deve ser fornecido em ordem aleatória e individualmente. Além disso, permitiu-se um máximo de cinco repetições para cada música.

A Figura 15 exibe o processo de convergência do vetor de métricas computando a distância euclidiana $\|w_k - w^*\|$, onde w_k é o vetor calculado em k-ésima atualização e w^* é a solução ótima, calculada no modo offline. Destaca-se o pequeno número de iterações necessárias para aproximar ao w^*, provando a capacidade do algoritmo para aprender uma métrica individual ao usuário.

Figura 15 – Convergência do Vetor de Métricas

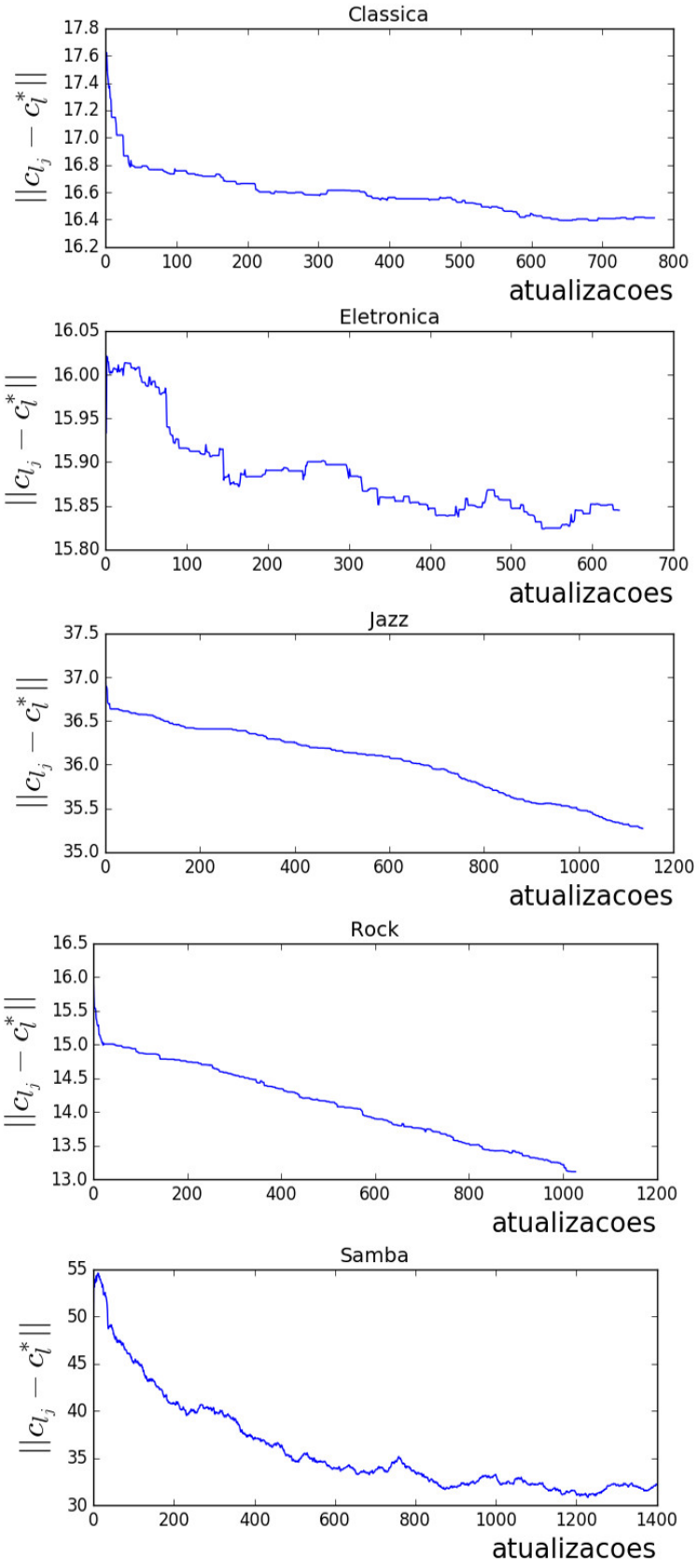


Fonte: Elaborado pelo autor

Para a avaliação da convergência dos centroides, em cada erro do algoritmo online calculou-se a distância euclidiana entre a solução online e a offline, inalterada, que foi calculada anteriormente. A Figura 16 mostra a convergência para o conjunto de dados MUSIC e a Figura 17 mostra a convergência dos centroides para o conjunto de dados GTZAN.

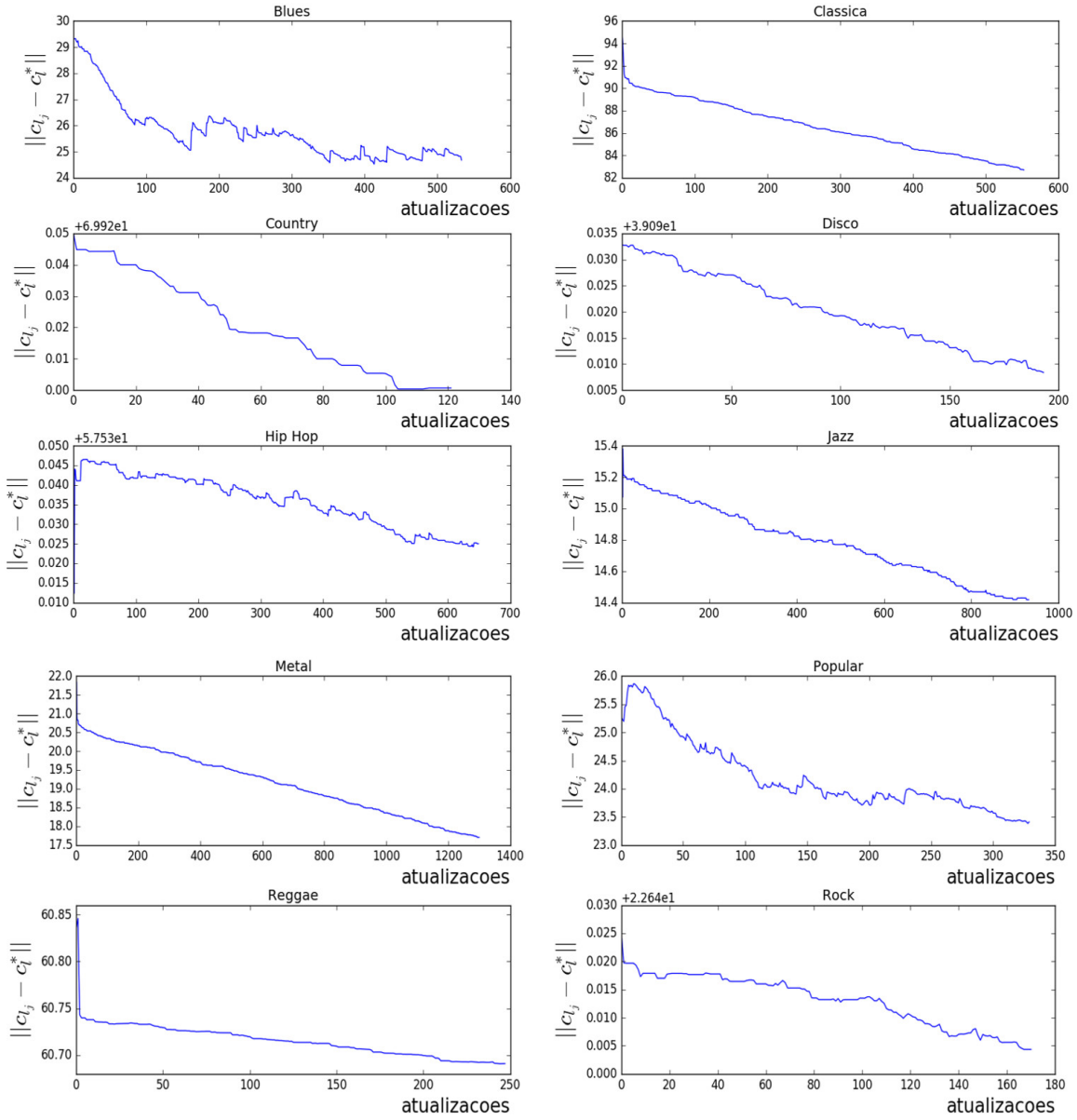
Durante os experimentos, a mesma taxa de aprendizado foi adotada para os dois algoritmos. A taxa de erro offline tem um valor igual a $32,57\%$ para o conjunto de dados Music e $38,89\%$ para o conjunto de dados GTZAN. A avaliação da taxa de erro online apresentada na Figura 18, que representa o *regret* médio, foi calculada com uma média móvel nas últimas 50 atualizações. Apesar de, no início, ocorrer uma alta taxa de erro, após poucas iterações, pode-se observar que o *regret* médio converge para zero.

Figura 16 – Convergência dos centroides - MUSIC



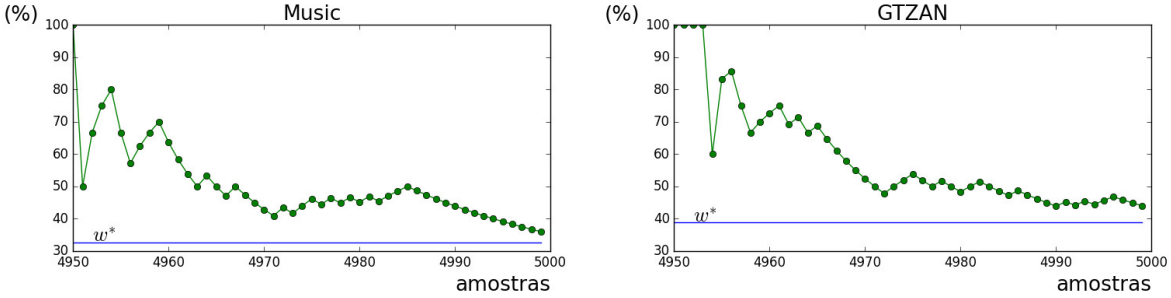
Fonte: Elaborado pelo autor

Figura 17 – Convergência dos centroides - GTZAN



Fonte: Elaborado pelo autor

Figura 18 – Variação da Taxa de Erro Online



Fonte: Elaborado pelo autor

5 CONCLUSÕES E TRABALHOS FUTUROS

Esse trabalho apresentou uma nova abordagem como solução do problema de classificação de gêneros de músicas. A proposta apresentada resume-se em utilizar informações acústicas estruturais extraídas diretamente dos arquivos de áudio em formato MP3 para composição do vetor característico, aplicar uma redução de dimensionalidade sobre o conjunto de vetores e a submetê-los a um processo de classificação usando uma técnica de aprendizado de métricas de similaridades.

Inicio-se a construção da solução final a partir de um modelo de aprendizado em modo batch, com todas as músicas apresentadas ao modelo de uma única vez, e os experimentos mostraram que o modelo de classificação baseado na aprendizagem métrica tende a melhorar seu desempenho geral de treinamento e testes, alcançando valores de predição consistentes ao estado da arte, superando o SVM linear de margem flexível. Além do desempenho inferior ao modelo proposto, a maior variância apresentada pelo SVM indica uma grande variação na predição dos dados futuros comprometendo diretamente a confiabilidade do modelo relacionado para este tipo de problema.

O modelo apresentado foi submetido a diversos cenários de testes que visavam mensurar seu desempenho com variações no número de músicas, tempo de duração e quantidade de dimensões. Em todos estes cenários, o desempenho obtido pelo modelo proposto foi consistente e invariante, o que validou a estratégia de redução de dimensionalidade e viabilizou uma evolução do modelo em batch para um modelo com configuração online, com as músicas sendo apresentadas de forma iterativa.

A solução online, também baseada apenas em informações acústicas, representa uma abordagem capaz de aprender o perfil de consumo musical de um cliente de forma individual. Evoluiu-se a solução em batch, para uma solução com regras de correções iterativas que é comparada à solução de origem para avaliar sua convergência. Os experimentos indicaram que os resultados da aprendizagem online convergem para os mesmos resultados obtidos pela aprendizagem offline. Isto permite ressaltar que, apesar da ausência de treinamento, os resultados apresentados demonstraram que a abordagem de aprendizado de métricas é muito eficiente e estável para um cenário online, estimulando a aplicação desta abordagem à plataformas comerciais de streaming de música, como Spotify, Deezer e outros.

Com a validação do modelo de classificação, pode-se estender o trabalho a outras abordagens. Mesmo com o bom desempenho, o percentual de acurácia ainda pode evoluir. Para isso, um estudo de seleção de características pode ser pensado para incrementar novas características acústicas que permitam uma melhor diferenciação na estrutura das músicas.

Pensando em uma aplicação comercial, pode-se utilizar a informação do vetor de métricas, que representa o cliente, e que também pondera o centroide, para desenvolver

um modelo de ranqueamento de músicas mais atrativas ao cliente de forma individualizada que permite abordar problemas de recuperação e recomendação de músicas em forma de busca por músicas similares e indicação de músicas de forma unitária em organizadas em playlists. Além disso, este perfil de consumo musical combinado à informações de metadados, como idade e região, permite um largo campo de estudos para comercialização de músicas.

REFERÊNCIAS

Anava, O.; Hazan, E.; Mannor, S.; Shamir, O. (2013). Online learning for time series prediction. In *COLT*.

Bansal, N.; Blum, A.; Chawla, S.; Meyerson, A. (2003). Online oblivious routing. In *Proceedings of the Fifteenth Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA '03, pages 44–49, New York, NY, USA. ACM.

Bar-Hillel, A.; Hertz, T.; Shental, N.; Weinshall, D. (2003). Learning distance functions using equivalence relations. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 11–18. AAAI Press.

Barrington, L.; Oda, R.; Lanckriet, G. (2009). Smarter than Genius? Human Evaluation of Music Recommender Systems. In *International Society for Music Information Retrieval Conference*, volume 9, pages 357–362, Kobe, Japan.

Bellet, A.; Habrard, A.; Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *CoRR*, **abs/1306.6709**.

Bello, J. (2007). Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats. In *International Conference on Music Information Retrieval*, pages 239–244, Vienna, Austria.

Ben, X.; Meng, W.; Yan, R.; Wang, K. (2012). An improved biometrics technique based on metric learning approach. *Neurocomput.*, **97**, 44–51.

Bergstra, J.; Casagrande, N.; Erhan, D.; Eck, D.; Kégl, B. (2006). Aggregate Features and ADABOOST for Music Classification. *Machine Learning*, **65**(2-3), 473–484.

Blum, A. (1998). On-line algorithms in machine learning. In *Developments from a June 1996 Seminar on Online Algorithms: The State of the Art*, pages 306–325, Berlin, Heidelberg. Springer-Verlag.

Brown, J. C. (1991). Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, **89**(1), 425–434.

Brown, J. C. (1999). Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *The Journal of the Acoustical Society of America*, **105**(3), 1933–1941.

Carafini, A. (2015). *Quantização vetorial de imagens coloridas através do algortimo LBG*. Ph.D. thesis, Federal University Rio Grande de Sul, Rio Grande do Sul, Brazil.

Chua, B.-Y. (2007). *Automatic extraction of perceptual features and categorization of music emotional expression from polyphonic music audio signals*. PhD dissertation, Monash University, Churchill, Australia.

Coelho, M. A.; Neto, R. F.; Borges, C. C. (2012). Perceptron Models for Online Structured Prediction. In *Proceedings of the 13th international conference on Intelligent Data Engineering and Automated Learning*, volume 7435, pages 320–327, Berlin, Heidelberg. Springer-Verlag.

Coelho, M. A.; Borges, C. C.; Neto, R. F. (2017). Uso de predição estruturada para o aprendizado de métrica. In *Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods*.

Corrêa, D. C.; Rodrigues, F. A. (2016). A survey on symbolic data-based music genre classification. *Expert Systems with Applications*, **60**, 190 – 210.

Cortes, C.; Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Cover, T.; Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**(1), 21–27.

Domeniconi, C.; Gunopulos, D. (2002). Adaptive nearest neighbor classification using support vector machines. *Proc NIPS*.

Edwards, A. W. F.; Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics*, **21**, 362–375.

Ellis, D. P. W.; Poliner, G. E. (2007). Identifying cover songs with chroma features and dynamic programming beat tracking. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 1429–1432.

Eronen, A.; Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 2, pages II753–II756 vol.2.

Esakkirajan, S.; Veerakumar, T.; Navaneethan, P. (2009). Adaptive vector quantization based video compression scheme. In *2009 International Multimedia, Signal Processing and Communication Technologies*, pages 40–43.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

Fu, Z.; Lu, G.; Ting, K. M.; Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, **13**(2), 303–319.

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using common lisp music. In *International Conference on Music Information Retrieval*, pages 464–467.

Giannakopoulos, T.; Pikrakis, A. (2014). Audio Features. In *Introduction to Audio Analysis*, pages 59–103. Academic Press, Oxford.

Gill, P. E.; Murray, W.; Wright, M. H. (1981). *Practical Optimization*. Academic, New York.

Gomez, E. (2006). *Tonal description of music audio signals*. Ph.D. thesis, Dept. Technol., Universitat Pompeu Fabra, Barcelona, Spain.

Gordillo, C. D. A. (2013). *Reconhecimento de Voz Contínua Combinando os Atributos MFCC e PNCC com Métodos de Robustez SS, WD, MAP e FRN*. Ph.D. thesis, PUC RIO DE JANEIRO, Rio de Janeiro, Brazil.

Gupta, S. (2014). Music Data Analysis: A State-of-the-art Survey. *CoRR*, **abs/1411.5**.

Gómez, E.; Herrera, P. (2004). Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04*.

Gómez, E.; Herrera, P. (2006). The song remains the same: identifying versions of the same piece using tonal descriptors. In *International Conference on Music Information Retrieval*, pages 180–185.

Hadsell, R.; Chopra, S.; LeCun, Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Heng-Tze Cheng; Yi-Hsuan Yang; Yu-Ching Lin; I-Bin Liao; Chen, H. H. (2008). Automatic chord recognition for music classification and retrieval. In *2008 IEEE International Conference on Multimedia and Expo*, pages 1505–1508.

Herrada, O. C. (2010). The Long Tail in Recommender Systems. In *Music Recommendation and Discovery*, pages 87–107. Springer Berlin Heidelberg.

Herrera, P.; Amatriain, X.; Batlle, E.; Serra, X. (2000). Towards instrument segmentation for music content description a critical review of instrument classification techniques. In *International Conference on Music Information Retrieval*, Plymouth, Massachusetts, USA. Content Based Retrieval.

J. Serra, E. Gomez, P. H.; Serra, X. (2008). Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. Audio, Speech, Lang. Process*, **16**(6), 1138–1151.

Jang, J. R.; Lee, H. (2008). A general framework of progressive filtering and its application to query by singing/humming. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(2), 350–358.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Kanawade, P. R.; Gundal, S. S. (2017). Tree structured vector quantization based technique for speech compression. In *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, pages 274–279.

Kim, J.; Tomasik, B.; Turnbull, D. (2009). Using artist similarity to propagate semantic information. In *Proc. ISMIR Conf.*, pages 375–380.

Korhonen, M. D.; Clausi, D. A.; Jernigan, M. E. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **36**(3), 588–599.

Kulis, B. (2012). Metric learning: A survey. *Foundations and Trends in Machine Learning*, **5**, 288–364.

Lajugie, R.; Bach, F.; Arlot, S. (2014). Large-margin metric learning for constrained partitioning problems. *Proceedings of the 31St International Conference on Machine Learning*, pages 297–305.

Law, M. T.; Gutierrez, C. S.; Thome, N.; Gancarski, S.; Cord, M. (2012). Structural and visual similarity learning for web page archiving. In *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6.

Lee, C.; Shih, J.; Yu, K.; Lin, H. (2009). Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Transactions on Multimedia*, **11**(4), 670–682.

Lee, K. (2006). Automatic chord recognition from audio using enhanced pitch class profile. In *International Conference on Music Information Retrieval*.

Li, T.; Ogihara, M. (2003). Detecting emotion in music. In *International Conference on Music Information Retrieval*, volume 2003.

Li, T.; Ogihara, M.; Li, Q. (2003). A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 282–289, New York, NY, USA. ACM.

Lidy, T.; Rauber, A. (2005). Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *International Conference on Music Information Retrieval*, pages 34–41.

Lidy, T.; Rauber, A.; Pertusa, A.; Quereda, J. M. I. (2007). Improving genre classification by combination of audio and symbolic descriptors using a transcription systems. In *International Conference on Music Information Retrieval*.

Linde, Y.; Buzo, A.; Gray, R. (1980). An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, **28**(1), 84–95.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, **28**(2), 129–137.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. *Proc. 1st Int. Symposium Music Information Retrieval*.

Loughran, R.; Walker, J.; O'Neill, M.; O'Farrell, M. (2008). The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification. *In proceedings of the international computer music conference*, pages 24–29.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Mahesha, P.; Vinod, D. (2012). Vector quantization and mfcc based classification of dysfluencies in stuttered speech. *Bonfring International Journal of Man Machine Interface*, **2**, 01–06.

Markel, J. E.; Gray, A. H. (1982). *Linear Prediction of Speech*. Springer-Verlag, New York.

McFee, B.; Lanckriet, G. (2010). Metric learning to rank. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 775–782, USA. Omnipress.

McFee, B.; Barrington, L.; Lanckriet, G. (2010). Learning Similarity from Collaborative Filters. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR*, pages 345–350.

McKinney, M. F.; Breebaart, J. (2003). Features for audio and music classification. *International Society for Music Information Retrieval Conference, ISMIR*, pages 151–158.

Okida, C. M.; Queiroz, R. A. B. (2006). Investigação de classificação de sons através de redes neurais artificiais: um estudo de caso - classificação de música.

Oppenheim, A. V. (1969). Speech Analysis-Synthesis System Based on Homomorphic Filtering. *The Journal of the Acoustical Society of America*, **45**, 458–465.

P. W. Ellis, D. (2007). Classifying music audio with timbral and chroma features. In *International Society for Music Information Retrieval Conference*, pages 339–340.

P. W. Ellis, D.; Whitman, B.; Berenzweig, A.; Lawrence, S. (2002). The quest for ground truth in musical artist similarity. In *ISMIR*, page 170–177.

Pampalk, E.; Rauber, A.; Merkl, D. (2002). Content-based organization and visualization of music archives. In *Proceedings of the ACM International Multimedia Conference and Exhibition*, pages 570–579.

Pampalket, E. (2006). *Computational models of music similarity and their application in music information retrieval*. Ph.D. thesis, Vienna University of Technology, Vienna, Austria.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, **81**(9), 1215–1247.

Rabiner, L. R.; Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Pearson Education.

Samal, A.; Parida, D.; Ranjan Satapathy, M.; Mohanty, M. (2014). On the use of mfcc feature vector clustering for efficient text dependent speaker recognition. *International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013*, pages 305–312.

Schultz, M.; Joachims, T. (2004). Learning a Distance Metric from Relative Comparisons. In S. Thrun; L. K. Saul; B. Schölkopf; B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 41–48. MIT Press.

Shaw, B.; Huang, B.; Jebara, T. (2011). Learning a distance metric from a network. In J. Shawe-Taylor; R. S. Zemel; P. L. Bartlett; F. Pereira; K. Q. Weinberger; K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1899–1907. Curran Associates, Inc.

Silla, C. N.; Freitas, A. A. (2009). Novel top-down approaches for hierarchical classification and their application to automatic music genre classification. *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 3499–3504.

Siqueira, J. K. (2012). *Reconhecimento de voz contínua com atributos mfcc, ssch e pncc, wavelet denoising e redes neurais*. Ph.D. thesis, PUC RIO DE JANEIRO, Rio de Janeiro, Brazil.

Slaney, M.; White, W. (2007). Similarity based on rating data. In *ISMIR*.

Slaney, M.; Weinberger, K.; White, W. (2008). Learning a metric for music similarity. In *International Conference on Music Information Retrieval*, pages 313–318.

Stevens, S. S.; Volkmann, J.; Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, **8**, 185–190.

Sturm, B. L. (2013). The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *CoRR*, **abs/1306.1**.

T. Lidy, A. Rauber, A. P.; Inesta, J. (2007). Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In *International Conference on Music Information Retrieval*.

Taskar, B.; Chatalbashev, V.; Koller, D.; Guestrin, C. (2005). Learning Structured Prediction Models: A Large Margin Approach. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 896–903, New York, NY, USA. ACM.

Taylor, M.; Kulis, B.; Sha, F. (2011). Metric learning for reinforcement learning agents. In *10th International Conference on Autonomous Agents and Multiagent Systems 2011, AAMAS 2011*, volume 2, pages 777–784.

Trang, H.; Tran, L.; Bui Hoang Nam, H. (2015). Proposed combination of pca and mfcc feature extraction in speech recognition system. *The 2014 International Conference on Advanced Technologies for Communications (ATC'14)*, **2015**, 697–702.

Tsai, W.-H.; Yu, H.-M.; Wang, H.-M. (2005). Query-by-example technique for retrieving cover versions of popular songs with similar melodies. In *International Conference on Music Information Retrieval*.

Tzanetakis, G.; Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, **10**(5), 293–302.

Tzanetakis, G.; Ermolinskiy, A.; Cook, P. R. (2002). Pitch histograms in audio and symbolic music information retrieval. In *International Conference on Music Information Retrieval*.

Unal, E.; Chew, E.; Georgiou, P. G.; Narayanan, S. S. (2008). Challenging uncertainty in query by humming systems: A fingerprinting approach. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(2), 359–371.

Vandenberghe, L.; Boyd, S. (1996). Semidefinite programming. *SIAM Rev.*, **38**(1), 49–95.

Velankar, M.; Sahasrabuddhe, H.; Kulkarni, P. (2015). Modeling melody similarity using music synthesis and perception. *Procedia Computer Science*, **45**, 728 – 735. International Conference on Advanced Computing Technologies and Applications (ICACTA).

Vershik, A. (2013). Long history of the monge-kantorovich transportation problem. *The Mathematical Intelligencer*, **35**.

Villela, S. M.; de Castro Leite, S.; Neto, R. F. (2016). Incremental p-margin algorithm for classification with arbitrary norm. *Pattern Recognition*, **55**, 261–272.

Vishnupriya, S.; Meenakshi, K. (2018). Automatic Music Genre Classification using Convolution Neural Network. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4.

Vlegels, J.; Lievens, J. (2017). Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences. *Poetics*, **60**, 76–89.

Vyas, G. (2014). Automatic mood detection of indian music using mfccs and k-means algorithm. *2014 7th International Conference on Contemporary Computing, IC3 2014*.

W. H. Tsai, H. M. Y.; Wang, H. M. (2005). A query-by-example technique for retrieving cover versions of popular songs with similar melodies. In *Int. Conf. Music Information Retrieval*.

Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 577–584, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Weinberger, K.; Blitzer, J.; K. Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, **10**.

Weinberger, K. Q.; Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, **10**, 207–244.

West, K. (2008). *Novel Techniques for Audio Music Classification and Search*. PhD dissertation, University East Anglia, Norwich, U.K.

Wickramarachi, P. (2003). Effects of windowing on the spectral content of a signal.

Wolff, D.; Weyde, T. (2014). Learning music similarity from relative user ratings. *Information Retrieval*, **17**(2), 109–136.

Xin, L.; Xuezheng, L.; Ran, T.; Youqun, S. (2014). Content-based retrieval of music using mel frequency cepstral coefficient (MFCC). *Computer Modelling & New Technologies*, **18**(11), 1356–1361.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; Russell, S. (2002). Distance Metric Learning, with Application to Clustering with Side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, volume 15 of *NIPS'02*, pages 521–528, Cambridge, MA, USA. MIT Press.

Yang, D.; Lee, W. (2004). Disambiguating music emotion using software agents. In *International Conference on Music Information Retrieval*, pages 10–14.

Yang, L. (2006). *Distance Metric Learning: A Comprehensive Survey*, volume 2. Michigan State Universiy.

Yang, P.; Hsieh, C.-J.; Wang, J.-L. (2018). History PCA: A New Algorithm for Streaming PCA. *arXiv*.

Yazhong Feng; Yueting Zhuang; Yunhe Pan (2003). Music information retrieval by detecting mood via computational media aesthetics. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 235–241.

Yen, F. Z.; Luo, Y.-J.; Chi, T.-S. (2014). Singing Voice Separation Using Spectro-Temporal Modulation Features. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, pages 617–622.

Zhang, Y.; Yeung, D.-Y. (2010). Transfer metric learning by learning task relationships. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1199–1208, New York, NY, USA. ACM.

Zhang, Z.; Kwok, J.; Yeung, D.-Y. (2003). Parametric distance metric learning with label information. *Proc International Joint Conference on Artificial Intelligence*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 928–935. AAAI Press.