João Gabriel Bracaioli Araújo Vitório

SELEÇÃO DINÂMICA DE CLASSIFICADORES PARA ROTULAÇÃO DE MÚSICAS EM EMOÇÃO E GÊNERO

Curitiba - PR, Brasil

2019

João Gabriel Bracaioli Araújo Vitório

SELEÇÃO DINÂMICA DE CLASSIFICADORES PARA ROTULAÇÃO DE MÚSICAS EM EMOÇÃO E GÊNERO

Projeto de Dissertação de Mestrado apresentado ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de mestre em Informática.

Pontifícia Universidade Católica do Paraná - PUCPR

Programa de Pós-Graduação em Informática - PPGIa

Orientador: Carlos Nascimento Silla Junior

Coorientador: Yandre Maldonado e Gomes da Costa

Curitiba - PR, Brasil

2019

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Edilene de Oliveira dos Santos CRB-9/1636

V845s
2019

Vitório, João Gabriel Bracaioli Araújo
Seleção dinâmica de classificadores para rotulação de músicas em emoção e gênero / João Gabriel Bracaioli Araújo Vitório ; orientador: Carlos Nascimento Silla Junior ; coorientador: Yandre Maldonado e Gomes da Costa. -- 2019
107 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2019
Bibliografia: f. 99-107

1. Informática. 2. Música – Classificação. 3. Estilo musical. 4. Emoções na música. 5. Banco de dados. I. Silla Junior, Carlos Nascimento. II. Costa, Yandre Maldonado e Gomes da. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. IV. Título

. CDD. 20.ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

15-2021

DECLARAÇÃO

Declaro para os devidos fins que o aluno **JOÃO GABRIEL BRACAIOLI ARAÚJO VITÓRIO**, defendeu sua dissertação de Mestrado intitulada **"SELEÇÃO DINÂMICA DE CLASSIFICADORES PARA ROTULAÇÃO DE MÚSICA EM EMOÇÃO E GÊNERO"**, na área de concentração Ciência da Computação, no dia 28 de agosto de 2019, no qual foi aprovado.

Declaro ainda que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade, firmo a presente declaração.

Curitiba, 15 de março de 2021.

Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná

Rua Imaculada Conceição,1155 Prado Velho CEP 80215-901 Curitiba Paraná Brasil
Fone: (41) 3271-1669 e-mail: secretaria@ppgia.pucpr.br

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Resumo

A música está presente entre as diferentes populações no mundo, sendo uma forma de manifestação cultural que também pode ser instrumentalizada para diversos fins. Assim, de acordo com a necessidade e objetivo a ser alcançado através da música, categorias específicas podem ser requeridas, ou utilizadas. Para sua classificação, é possível a utilização de diferentes formas ou métodos, como estilos, emoções que transmitem, ou mesmo gênero ao qual pertencem, podendo inclusive apresentar uma estrutura hierárquica, com categorias e subcategorias, se tornando um desafio a seleção e combinação de músicas em conjuntos a partir destas características semelhantes. Dada a relevância da problemática apresentada e visando a melhoria da classificação musical para uma posterior utilização de acordo com os diversos fins, a classificação de músicas se tornou o foco da Music Information Retrieval (MIR). Nesta perspectiva, o objetivo deste trabalho é a rotulação automática de músicas de acordo com o gênero e emoção. Para isso, foram utilizadas as representações de áudios (SSD, RH, RP, MFCC), de imagens (LBP) e de letras (Stemming, N-grams) extraídas das músicas das bases de dados BRMD (gênero e sentimento), LMMD e LMD. Primeiro foi realizada a análise do comportamento das representações sendo possível identificar um comportamento diferente para cada tipo de rótulo, como por exemplo, com a utilização da imagem com gênero e da imagem com sentimento. Em sequência, foram utilizadas técnicas de seleção dinâmica de conjuntos de classificadores, investigando a combinação de um conjunto de classificadores criados a partir das diferentes representações, tendo como objetivo aprimorar os resultados obtidos nos primeiros experimentos visando uma melhor eficiência na rotulação. Neste ponto, destacam-se os algoritmos KNORA's e DES-P, que apresentaram os melhores resultados quando combinadas todas as representações, no entanto, destaca-se que a seleção dinâmica, em geral, não alcançou o oráculo máximo possível. Esse trabalho traz importante contribuição para a Music Emotion Recognition (MER) ao analisar a base de dados BRMD com a rotulação de sentimento e os resultados obtidos demonstram que a seleção dinâmica melhora a assertividade na rotulação da música, em especial com o algoritmo k-Nearest Oracle Union (KNORA-U). A inovação deste estudo se deu pela utilização das bases de emoção para classificação automática de sentimentos através das bases Latin Music Database Mood Database e Brazilian Music Database, considerando que esta última ainda não havia sido utilizada em outros trabalhos.

Palavras-chave: Música, Classificação automática, Gênero musical, Emoção, Seleção Dinâmica.

Abstract

Music is present in different cultures around the world, being an important component for cultural movements that can also be instrumented for a diverse range of use. Thus, according to necessity and goal that the song is willing to achieve, different music styles (or categories) may be utilized, or even required. In order to classify them, it is possible to use different approaches and methods, such as music styles, their transmitted emotions, their genre, or also present some hierarchical structure, with categories and subcategories, becoming a challenge to select and combine songs in collections considering their similar features. Given the relevance of the above mentioned problem, and aiming to enhance the current musical classification studies for a further reference, the classification of songs became the focus of Music Information Retrieval (MIR). In this perspective, the objective of this work is automatic labeling of songs according to their genre and mood. In order to achieve such goal, audio (SSD, RH, RP, MFCC), images (LBP), and letters (Stemming, N-grams) representations were utilized. They were extracted from songs contained in the BRMD (genre and mood), LMMD, and LMD databases. Firstly, a representation behavior analysis was performed, allowing the identification of a different behavior for each label type, for instance, with the utilization of the genre image representation and the mood image representation. Next, techniques of dynamic classifier selection were utilized in order to investigate the combination of classifiers in a pool created from different representations, aiming to enhance the first experiments' results, in order to improve labeling efficiency. At this stage, KNORA's and DES-P algorithms stand out. Such algorithms presented the best results when combined with all representations. However, in general, it has been noticed that the dynamic selection has not reached the maximum possible oracle. This paper brings up an important contribution to Music Emotion Recognition (MER) when analyzing the BRMD database with their mood labeling. Moreover, the results show that the dynamic selection improves music labeling assertiveness, especially when employing the k-Nearest Oracle Union algorithm (KNORA-U). The innovative aspect of the present paper has to do with the employment of emotion databases for automated classification of mood through the Latin Music Database Mood Database and the Brazilian Music Database, taking into account that the latter had not been adopted by other studies yet.

keyworks: Automatic Classification, Genre, Emotion, Mood, Dynamic Selection.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BRMD	Brazilian Music Database
CNN	Convolutional Neural Networks
DWCH	Daubechies Wavelet Coeficient Histogram
DWT	Discrete Wavelet Transform
GLCM	Gray level Co-occurrence Matrices
GMM	Gaussian Mixture Models
GMRF	Gaussian Markov Random Fields
IDF	Inverse Document Frequency
IOIHC	Inset-Onset Interval Histogram Coefficients
k-NN	k-Nearest Neighbor
KNORA	K-Nearest-ORAcle
LBP	Local Binary Pattern
LCA	Local Class Accuracy
LMD	Latin Music Database
LMMD	Latin Music Mood Database
MIR	Music Information Retrieval
MFCCs	Mel-frequency cepstrum coefficients
MP-F	F-measure
OLA	Overall Local Accuracy
OSC	Octave-based Spectral Contrast
PCM	Pulse-code modulation
RH	Rhythm Histogram
SMC	Sistemas de Múltiplos Classificadores
SoX	Sound eXchange
SSD	Statistical Spectrum Descriptor
SREM	Sistemas de Reconhecimento de Emoção Musical
STFT	Short Term Fourier Transform
SVM	Support Vector Machine
TF	Term Frequency

Sumário

1 Introdução

A música está presente em todas as civilizações, sendo um fenômeno complexo e fascinante, conforme Gjerdingen e Perrott (2008). Não apenas para entretenimento ou prazer, está relacionada diretamente a formações culturais, sociais, psicológicas e fisiológicas, ou seja, possui uma ampla gama de propósitos. De acordo com a necessidade, as músicas podem ser rotuladas, sendo possível a utilização de diferentes formas ou métodos para sua classificação, como estilos, emoções e gêneros, podendo também apresentar uma estrutura hierárquica, com categorias e subcategorias.

Com a expansão da internet no início do século XXI e o surgimento de ferramentas de *streaming* de música digital, esta passou a ser mais acessível às pessoas, que por sua vez, passaram a buscar ferramentas digitais capazes de realizar seleções musicais que atendam às suas necessidades individuais em diferentes aspectos e demandas, o que se tornou o foco da *Music Information Retrieval* (MIR), que é uma ciência interdisciplinar de crescente interesse, relacionada à recuperação de informações e conhecimentos relacionados à música (ORIO, 2006).

Muitas plataformas de *streaming* de música como *Spotify*, *Pandora* e *Saavn* usam serviços automatizados para recomendações de músicas aos seus usuários e para que esta tarefa seja bem sucedida, é necessária uma correta rotulação da música, pois é a partir desta rotulação que será possível atender aos anseios dos ouvintes, que objetivam categorias específicas de músicas, de acordo com seus desejos e particularidades. Ou seja, a correta rotulação é essencial para recomendação e promoção adequada das músicas desejadas, como por exemplo, pelos inúmeros usuários das plataformas de *streaming* (SRIDHARAN; MOH; MOH, 2018).

A rotulação automática de uma música é definida como uma tarefa de classificação. A classificação, por sua vez, é uma das tarefas na área de conhecimento de padrões e tem como função atribuir classes, dentro de várias possibilidades, a uma amostra de teste (representada por um vetor de características não vistas anteriormente). As classes podem ser os vários rótulos que uma música pode ter e seu objetivo é descrever o conteúdo semântico de determinadas faixas.

Visando facilitar a experiência do usuário, os rótulos podem ter atributos relacionados de forma automática, através da utilização de classificadores, ou outra possibilidade é de que esta marcação, ou rotulação, possa ser realizada de forma colaborativa, utilizando sistemas de marcação social, onde um indivíduo aplica curtas anotações de texto aos itens, geralmente para organizar seu conteúdo pessoal, conforme explicam Lin, Chung e Chen (2018).

Partindo de características comuns e visando que essa classificação seja realizada de maneira automática, é possível a criação de um classificador dotado de inteligência artificial para a rotulação de músicas. Na criação e modulação deste classificador são envolvidas decisões tão complexas quanto sua execução, incluindo a forma de treinamento: se será supervisionado ou não supervisionado, com amostras rotuladas ou com amostras cujas classes são desconhecidas ou ainda se as amostras serão consideradas no processo de treinamento. Este classificador deve ser configurado em um modelo com parâmetros para uma seleção mais objetiva, alcançando assim um melhor desempenho e resultado global.

A seleção do classificador mais adequado passa por um processo de minimização de erros, pois, um único classificador pode obter um desempenho satisfatório para reconhecer determinado conjunto de amostras de testes, no entanto falhar em outros conjuntos. Essa possível falha ocorre em razão da maneira de aprendizado do algoritmo do classificador para execução da classificação e, objetivando melhorar a assertividade do resultado, foi criado o conceito de "conjuntos de classificadores" (VRIESMANN et al., 2012).

Conjuntos de classificadores (classifier ensemble), também conhecidos como sistemas de multi-classificadores (*Multi-Classifier Systems* - MCSs) e que serão apresentados de maneira detalhada na "Seção 3.3" têm sido utilizados para alcançar maior robustez nos resultados, melhorando as taxas Individuais no reconhecimento de classes em processos de rotulação automática, explorando a ideia de que conjuntos formados por diferentes classificadores, ou seja, com diferentes formas de aprendizado, podem oferecer informações complementares, aumentando a efetividade do processo de reconhecimento.

Podendo ocorrer tanto de maneira estática quanto dinâmica, a partir de um conjunto inicial de classificadores é realizada uma seleção para um subconjunto de "classificadores 1", também chamados de candidatos[1]. Estes primeiros classificadores selecionados são aqueles entendidos como "fracos" porque reconhecem apenas uma parte das instâncias do problema, vez que possuem uma estrutura simples e com baixo poder discriminatório.

De acordo com Barsalou et al. (2008), na seleção estática, a região de competência[2] dos classificadores será definida na fase de treinamento, sendo definido apenas um subconjunto de classificadores para classificação de todas as amostras de teste. Já na seleção dinâmica, a região de competência será estipulada na fase de classificação, sendo que nesta, para cada nova amostra serão escolhidos os classificadores que possuem maior probabilidade de classificar corretamente a amostra. Em seguida, os classificadores serão agrupados em um subconjunto, que será responsável pela rotulação.

[1] Os classificadores candidatos são aqueles que são selecionados a partir da amostra de teste, em razão de seu melhor desempenho na classificação da amostra dentro de um conjunto de classificadores.

[2] A ideia da região de competência vem da avaliação da competência dos classificadores com melhor desempenho na classificação da amostra teste. Sendo considerada a região de competência, aquela em que os k-classificadores obtiverem a maior taxa de assertividade. É a base da abordagem de seleção de classificadores.

Por fim, os rótulos são usados para facilitar a pesquisa, exploração e localização de itens novos ou semelhantes, além da localização de outros usuários com interesses comuns e, para além destas experiências, podem se tornar a solução-chave para muitos problemas difíceis do campo da pesquisa relacionados à área musical, MIR, como por exemplo, de detecção de gênero ou humor, descoberta musical ou recomendações. As pesquisas MIR estão em crescente aumento nos últimos anos, objetivando automatizar e prever rótulos a partir do áudio (SHAO; CHENG; KANKANHALLI, 2019).

O presente trabalho busca averiguar se a utilização de seleção dinâmica de conjuntos de classificadores é capaz de realizar uma eficiente rotulação de canções diversas através da combinação de conjuntos de características complementares para as diferentes representações da canção, com o objetivo de alcançar melhores resultados.

1.1 Motivação

Para o problema acima apresentado, a utilização de conjuntos de classificadores fracos surge como uma solução, visto que os sistemas baseados em múltiplos classificadores frequentemente possuem ao menos um de seus elementos do conjunto inicial que classifica corretamente uma determinada amostra de testes. Isso ocorre devido às diferenças dos classificadores, que possibilita o reconhecimento de diferentes pontos do domínio de aplicação.

A tendência crescente em pesquisas na rotulação da música com a utilização de diversos conjuntos e combinação de características tem como motivação a utilização de seleção de conjunto de classificadores devido à dificuldade em selecionar e recombinar conjuntos de características que melhorem a classificação (KOERICH; POITEVIN, 2005). Para tanto, se faz necessária a investigação e exploração de técnicas que realizem a seleção dinâmica das diferentes fontes de informação, obtendo melhores resultados na rotulação de canções.

Para tanto, a proposta deste trabalho é a utilização do rótulo, do gênero e da emoção para auxiliar o campo de pesquisa do MIR, que enfrenta diversas dificuldades, sendo necessária a expansão de seu rol de conhecimento contextual sobre música. Em destaque, uma grande problemática a ser enfrentada pela MIR é referente ao desempenho de um classificador nas amostras teste, sendo necessária uma quantidade significativa de amostras de treinamento para alcance de um bom resultado, no entanto é dificultoso o acesso a uma base de dados que seja representativa e com todas as amostras possíveis. Neste contexto, é comum o uso de bases de dados pequenas, impossibilitando o correto treinamento, que gera classificadores pouco apurados. Por consequência, na categorização por emoção, pesquisas anteriores também confirmaram a existência de uma lacuna no campo da Music Emotion Recognition (MER) (AN; SUN; WANG, 2017)

Há diversas explicações em diferentes níveis de interpretação, mas podemos considerar que a música induz facilmente as emoções humanas, e essa relação tem fascinado os seres humanos desde a antiguidade (JUSLIN; VÄSTFJÄLL, 2008). No entanto, os mecanismos desse fascínio ainda não são muito bem compreendidos, o que corrobora com o difícil desenvolvimento das técnicas de rotulação.

1.2 Objetivo

Diante dos cenários apresentados, o objetivo deste trabalho consiste na rotulação automática das diferentes representações de gênero e emoção de músicas diversas, sendo utilizadas a base de músicas brasileiras (Brazilian Music Database) e a base de músicas latinas (Latin Music Database e Latin Music Mood Database), identificando como essas representações contribuem na classificação dos rótulos e validando se as técnicas de seleção dinâmica destas fontes de informação trazem melhores resultados. Para tanto, foram analisados conjuntos de características extraídas de representações distintas das canções, verificando de maneira comparativa se estas alcançam o mesmo desempenho nos diferentes rótulos, explorando as técnicas que realizam a seleção dinâmica de classificadores, para então estabelecer se os resultados obtidos são melhores do que aqueles obtidos com as abordagens tradicionais de classificação.

Assim, são determinados os seguintes objetivos: a criação de diferentes conjuntos de informações a partir do áudio, da imagem e da letra da música, ou seja, de diversos conjuntos de representações através da utilização da técnica de extração de características; a criação de subconjuntos de classificadores complementares; avaliação do desempenho das representações para diferentes rótulos; avaliação de desempenho em bases musicais distintas para as rotulações de gênero e emoção e, por fim, a avaliação do impacto de uso de técnicas de seleção dinâmica de classificadores.

1.3 Contribuição

Em destaque e de maior impacto, dado a inovação de seu uso, a principal contribuição científica fora a utilização das bases de emoção no contexto de classificação automática de sentimentos através das bases Latin Music Database Mood Database e Brazilian Music Database, considerando que esta última ainda não havia sido utilizada em outros trabalhos. As demais contribuições desenvolvidas neste trabalho são: comparativo do desempenho das representações com diferentes rótulos, verificando eficácia, se superior ou inferior; analise do desempenho de diferentes técnicas de seleção de subconjunto de classificadores para diferentes tipos de rótulos e, por fim, a utilização de método de extração de características obtidas através das letras das músicas, através de técnicas de redução de dimensionalidade

da palavra, calculando a sua relevância na música, sendo posteriormente associadas a geração de classificadores.

1.4 Organização do Trabalho

Este trabalho encontra-se organizado da seguinte forma: no capítulo 2 é apresentada a revisão bibliográfica, que envolve a bibliografia acerca de classificação automática de gêneros e sentimentos musicais; no capítulo 3 são apresentados os principais fundamentos teóricos inerentes às técnicas de aprendizado supervisionado, utilizadas para a realização dos experimentos; o capítulo 4 descreve a metodologia utilizada nos experimentos; o capítulo 5 apresenta os resultados obtidos para as representação do áudio, representação visual e a representação das letras, empregando seleção dinâmica de agrupamento de classificadores para para o reconhecimento de emoções e gêneros musicais; por fim, o capítulo 6 aponta as conclusões finais.

2 Revisão Bibliográfica

A revisão bibliográfica foi dividida em duas partes: primeiramente uma breve análise dos trabalhos presentes na literatura que relacionam classificação de emoções com as características musicais, e a segunda parte foi dedicada tanto aos trabalhos relacionados à problemática da classificação de gênero musical, quanto àqueles focados na recuperação de informações musicais.

2.1 Classificação por Emoção

Lu, Liu e Zhang (2006) propõem um framework hierárquico para detectar emoções através de características acústicas (intensidade, timbre e ritmo) do áudio, utilizando o modelo de emoção de Thayer (1989) e, através destas características, aplicam uma regra de divisão em dois grupos, onde são considerados os que possuem baixa intensidade de timbre e ritmo como contentamento e depressão; e em situação contrária são considerados de excitação e entusiasmo. Entretanto, nesse modelo não se mostra evidente o desempenho alcançado utilizando os algoritmos de classificação *Gaussian Mixture Models* (GMM) ou *Support Vector Machine* (SVM), relata somente que o uso do SVM obteve resultados melhores.

O banco de dados uspop2002[1], composto por 8752 canções de músicas pop, fora utilizado por Eerola, Lartillot e Toiviainen (2009) através da ferramenta MIRtoolbox, que extraiu diversas características, aplicando técnicas de redução de dimensionalidade com *Principal Component Analysis* (PCA), utilizando variados conjuntos de recursos, mas que também não evidencia os resultados obtidos.

Utilizando características do áudio em conjunto com letras, Laurier, Grivolla e Herrera (2008) criaram uma base com músicas da *last.fm* e letras da *LyricWiki*[2], na qual 17 indivíduos catalogaram 1000 músicas dividindo-as em 4 categorias: triste, feliz, irritado e descontraído. Utilizando uma distribuição normalizada entre estas classes binárias foram extraídas as características acústicas como timbre, ritmo e descritores temporais, e para letra fora extraída a similaridade entre as letras, formando vetores de atributos baseados na análise dimensional e semântica latente, obtendo um desempenho de 61,30% para a utilização de SVM.

Em experimento com um conjunto de dados contendo 288 canções, que foram avaliadas por 50 indivíduos, onde informavam que a canção transmitia a emoção positiva

[1] http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html
[2] http://lyricwiki.org

ou negativa e, realizando o experimento em um ambiente controlado, após a classificação da emoção, Huq, Bello e Rowe (2010) extraem um conjunto de recursos como timbre, ritmo e harmonia do áudio, que realizando a redução de dimensionalidade com PCA e considerando uma variedade de algoritmos em cada estágio do processo de treinamento, o resultado final alcançado obteve 69,70% de precisão, também utilizando o SVM.

Para Chen et al. (2014), alcançar uma precisão satisfatória no reconhecimento de emoções musicais é uma tarefa difícil, pois requer uma grande quantidade de canções rotuladas pelos usuários. Adota-se uma estrutura probabilística para a modelagem de emoções, com estimulação de valência e um método de adaptação baseado na regressão linear para personalizar um modelo de fundo, aplicando uma estratégia de vinculação de componentes, onde são agrupados diversos componentes em um grupo e cada grupo vinculado compartilha uma transformação linear. Esta estratégia alivia a complexidade do problema da Regressão Linear e, ao mesmo tempo, aumenta a capacidade de generalização, criando um conjunto de dados para a realização dos experimentos: a AMG240, contendo 240 músicas diferentes com 30s todas rotuladas por de 14 a 16 indivíduos. Além disso, fazem o uso o MIRToolbox para extrair 70 características, demonstrando os resultados de forma empírica. Desta maneira foram obtidos bons resultados, porém não houve qualquer comparação ou exposição de uma porcentagem de assertividade.

Utilizando o modelo de emoção bi-dimensional de Thayer (1989) e um banco de dados com 280 músicas populares com as quatro emoções básicas, Pouyanfar e Sameti (2014) usaram um classificador de dois níveis, aplicando uma abordagem de seleção de recursos, diminuindo as correlações entre eles. Obtiveram então, uma taxa de precisão de 72,14% com SVM simples para 87,27%, com nosso classificador hierárquico. Também utilizando o modelo de Thayer (1989), An, Sun e Wang (2017) classificaram, através de Redes Bayesianas, 3552 canções chinesas rotuladas. Com uma biblioteca *Python Jieba*, os mesmos autores extraíram das letras das músicas as características e segmentações, para depois serem selecionadas as palavras no dicionário de emoção, se obtendo, então, o conjunto de dados para o experimento. Nesse sentido, a assertividade foi de 68%. An, Sun e Wang (2017) ressaltam que dada a multiplicidade de sentidos de uma palavra, não foi possível utilizar um conjunto maior de canções.

A Tabela 1 apresenta de forma resumida os trabalhos descritos nesta seção, que realizaram rotulação de emoções a partir do sinal de áudio.

2.2 Classificação por Gênero Musical

Após a expansão do acesso à informação causado pela internet, se fez necessário o desenvolvimento de métodos computacionais para organizar grandes volumes de dados a partir de uma representação compacta. Inclusa neste rol de dados que passaram a ter acesso

Tabela 1 – Resumo dos trabalhos de classificação automática por emoções em músicas

Autores	Características	Classificador	Modelo	Bases	Assertividade
(LU; LIU; ZHANG, 2006)	Intensidade, timbre e conteúdo rítmico	GMM e SVM	Thayer	uspop2002	62,90%
(LAURIER; GRIVOLLA; HERRERA, 2008)	Características visuais e áudio	SVM	Russell	Base Própria mil músicas	61,30%
(HUQ; BELLO; ROWE, 2010)	Textura de timbre, conteúdo rítmico com PCA	SVM	Thayer	Base Própria 288 músicas	69,70%
(CHEN et al., 2014)	MFCC	Regressão Linear	Russell	AMG240	não informado
(POUYANFAR; SAMETI, 2014)	Textura de timbre, conteúdo rítmico e frequência de vibração	SVM e classificador hierárquico	Thayer	Base Própria 280 músicas	72,14% e 87,27%
(AN; SUN; WANG, 2017)	Segmentação da letra, e dicionário de emoção	Redes Bayesiana	Russell	Base Própria 3552 músicas chinesas	68,00%

facilitado, a música não se diferencia quanto a esta necessária organização automatizada e por isso, a caracterização e classificação por gêneros musicais tem se destacado, no campo de pesquisa da MIR, pois contribuem para a busca de conteúdos musicais organizados.

Pachet e Cazaly (2000) discutem a dificuldade de classificar por gênero de maneira automática, entendendo que a classificação por gênero requer uma hierarquia de categorias a serem mapeadas, no entanto não existe uma taxonomia bem definida sobre gêneros. Analisando as principais bases de músicas são encontrados mais de 70 tipos de gêneros comuns, sendo que os gêneros mais usuais como rock ou pop, são detentores de diferentes conjuntos taxonômicos, demonstrando que são utilizados tanto critérios quanto taxonomias diferentes para classificação dos gêneros musicais. Pachet e Cazaly (2000) por sua vez, argumentam como essa confusão semântica dentro de uma única taxonomia pode não ser confusa para um indivíduo, mas dificilmente será resolvida por um sistema automático.

Para Tzanetakis e Cook (2002), a classificação automática de gêneros musicais é uma tarefa de reconhecimento de padrões que tomou notoriedade a partir do trabalho por eles desenvolvido, sendo descrita por três características: o espectro sonoro (*timbral texture*), o padrão rítmico (*beat-related*) e a altura da nota (*pitch-related*). Nos experimentos, criaram a primeira base específica para a tarefa de reconhecimento de gêneros musicais, que é composta por 1000 músicas divididas em 10 gêneros musicais populares nos anos 90 (clássica, country, eletrônica, hip-hop, jazz, latinas, metal, pop, reggae e rock) denominada GTZAN. Sendo utilizado o *framework* MARSYAS para extração de atributos, o *Short Term Fourier Transform* (STFT) para representação do timbre, e o *Discrete Wavelet Transform* (DWT) para estrutura rítmica da música, que provê uma resolução de todas as frequências de forma uniforme e, utilizando classificadores GMM e *K-Nearest Neighbors* (k-NN), foi obtida uma taxa final de acerto de 61% de proximidade à percepção humana.

Com as extrações tradicionais utilizadas por Tzanetakis e Cook (2002), foram coletadas informações incompletas de sinais musicais, os recursos textuais do timbre foram usados para o reconhecimento de fala e calculados para cada quadro curto do sinal

sonoro, enquanto os ritmos foram computados em toda a música, ou seja, os extratores de timbre coletaram as estatísticas das informações locais dos sinais da música a partir de uma perspectiva global, mas não o suficiente para representar a informação completa da música. Deixando-se, então, de coletar conteúdo suficiente de ritmo e tom para fins de classificação (LI; OGIHARA; LI, 2003). Pelo motivo exposto, Li, Ogihara e Li (2003) efetuaram a extração de conteúdo com *Daubechies Wavelet Coefficient Histogram* (DWCH), que coleta informações locais e globais do sinal da música em conjunto com Histogramas de Coeficientes fornecidos pela ondulação, obtendo um desempenho de 78,5% em média na validação cruzada de dez vezes.

Gjerdingen e Perrott (2008), em trabalho focado na capacidade humana de reconhecimento de gêneros musicais, realizaram uma série de experimentos com indivíduos e, onde deveriam classificar uma música entre os gêneros mais populares (clássica, country, eletrônica, jazz, rock, pop), concluem que a classificação humana chegou a 70%. Com uma nova forma de extração de conteúdo e utilizando o classificador SVM, Li, Ogihara e Li (2003) obtiveram 72% de acerto na base GTZAN, ou seja, uma taxa superior à análise humana.

Koerich e Poitevin (2005) propõem a extração de vários vetores, de características relacionadas a diferentes partes da música. Assim, utilizando centroide espectral, as características relacionadas à batida são extraídas de três regiões diferentes de um mesmo sinal de música, formando um vetor e gerando classificadores individuais que, por sua vez, são combinados de forma a melhorar a taxa de acerto do classificador. A forma adotada demonstra que a combinação de classificadores distintos atinge resultados superiores e mais acurados que os obtidos a partir da aplicação de processo que utilize um único classificador (KOERICH; POITEVIN, 2005). Os autores ressaltam que apesar dos resultados serem semelhantes aos já existentes na literatura, de 72% utilizando SVM, o método foi prejudicado pelo banco de dados usado, o GTZAN, de Tzanetakis e Cook (2002), que contém apenas 30 segundos de áudio.

Pampalk et al. (2005) desenvolvem um trabalho que introduz o conceito de *artist filter*, estabelecendo que um conjunto de treinamento não poderia conter títulos de um mesmo artista no conjunto de teste e de treinamento simultaneamente, recomendação esta que visa evitar que classificadores fiquem eficientes em classificar artistas ao invés de gêneros musicais. Os resultados obtidos demonstraram grande redução das taxas de acertos, de 72% para 27% quando comparados a experimentos que não utilizavam o conceito e, a partir desse trabalho, o *artist filter* passou a ser utilizado na tentativa de produzir classificadores mais desenvolvidos e realistas nas tarefas de reconhecimento de gêneros.

Investigando os reais benefícios do *artist filter*, Flexer (2007) obteve resultados que demonstram que seu uso não apenas diminui a precisão da classificação de gêneros, mas também pode diminuir as diferenças na utilização de diferentes técnicas de exatidão e,

como consequência, sugere que todos os resultados de trabalhos que não utilizam a técnica devam ser revistos, passando a ser empregada em vários outros trabalhos futuros.

Ocorre que as bases musicais disponíveis publicamente apresentam algumas limitações para o desenvolvimento de trabalhos de recuperação de conteúdo. Por exemplo, o GTZAN disponibiliza apenas 30 segundos de cada música no formato *Pulse-Code Modulation* (PCM). Além disso, outras bases possuem poucas músicas e os gêneros são sempre os mais populares: rock, pop, clássica, assim como outros já citados.

Diante dessas dificuldades, Silla Jr., Kaestner e Koerich (2008b) apresentaram a *Latin Music Database* (LMD), que possui 3227 gravações musicais catalogadas em dez gêneros de música latino-americana em formato MP3 (Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja e Tango). O desenvolvimento desta base visava o desafio da classificação de gêneros oriundos de regiões culturalmente semelhantes, e teve a sua classificação baseada na percepção de professores profissionais com mais de dez anos de experiência no ensino de dança de salão e danças culturais brasileiras, que levaram em consideração a forma com que cada música é dançada.

Utilizando a LMD, com a criação de múltiplos vetores de características de três segmentos de tempo, começo, meio e fim, com trechos de 30 segundos cada, Silla Jr., Kaestner e Koerich (2008a) geraram classificadores individuais para cada segmento de tempo, combinando-os posteriormente para obtenção de um classificador final. Os melhores resultados obtidos tiveram um acerto de 65%, o que é 3% melhor do que quando utilizado o classificador individual. Em estudo posterior, Silla Jr., Kaestner e Koerich (2010), combinaram diferentes conjuntos de características extraídas dos sinais *Statistical Spectrum Descriptor* (SSD), *Rhythm Histogram* (RH) e *Inset-Onset Interval Histogram Coefficients* (IOIHC), obtendo uma melhor assertividade, de 89% na base LMD, sem o uso de *artist filter*.

Seyerlehner et al. (2010), explorando o conceito de *block-level features*, defendem que desta forma são coletadas mais informações temporais do que outras características, melhorando assim a classificação dos gêneros. Utilizando como base a GTZAN e ISMIR 2004 e o classificador SVM, obtiveram uma efetividade de 82,72% e 77,96%, respectivamente. Na sequência, os autores obtiveram na MIREX 2010, com os experimentos utilizando a LMD, uma taxa de acerto de 79,86%, utilizando o conceito de *artist filter*.

Lopes et al. (2010) utiliza um método para a seleção de instâncias de treinamento com base na precisão do classificador SVM. O método consiste em vetores que representam características de tempo curto e de baixo nível extraídos de sinais de áudio de músicas. Em um conjunto de dados de 900 músicas da LMD, os autores afirmam que em seu modelo de classificação a resultante é significativamente menor, permitindo uma classificação mais rápida se comparada aos dados de teste. É importante considerar que os autores utilizaram o conceito de *artist filter*.

Wu et al. (2011) experimentaram a extração de características acústicas por meio de imagens de espectrogramas geradas a partir do sinal que denominaram "características visuais", além de características acústicas de tempo curto como *Octave-based Spectral Contrast* (OSC) e *Mel-frequency cepstrum coefficients* (MFCCs), entre outras de tempo longo, com o filtro de Gabor aplicado para extrair traços de textura do espectrograma. Com as bases GTZAN e ISMIR 2004, obtiveram 86,1% para ambos os conjuntos utilizando classificador SVM.

Utilizando o mesmo conceito de extração de características no domínio visual, Costa et al. (2013) combinam diferentes descritores de textura, através de regras de fusão, para geração de classificadores. Utilizando a base LMD, descritor de textura *Local Binary Pattern* (LBP), divisão em escala Mel e classificadores SVM, obteve uma assertividade de 82,33% com *artist filter*. No mesmo trabalho propõem a utilização de seleção dinâmica de classificadores a partir de características extraídas do LBP e com o uso do *K-Nearest-ORAcles* (KNORA) para seleção de um subconjunto de classificadores, obtiveram uma assertividade de 83%, taxa que o autor ressalta só ter sido alcançada quando aplicado o zoneamento por escala Mel, questionando, portanto, o custo adicional de processamento do mesmo, uma vez que o resultado alcançado é próximo a experimentações anteriores.

Ainda na seleção dinâmica de classificadores, Vriesmann et al. (2015) apresentam um comparativo de experimentações empregando a *Local Class Accuracy* (LCA) e *Overall Local Accuracy* (OLA) ao método KNORA, além de sua combinação com k-NN. O método proposto obteve os melhores resultados quando comparado aos métodos originais, no entanto, não apresenta qualquer superioridade significativa no reconhecimento de gênero musical, sendo assim, os autores corroboram a necessidade de explorar a proposta mais a fundo, visando melhores resultados.

Com uma abordagem diferente da extração de características e classificação, Romano e Adami (2015) propõem a utilização de diferentes técnicas para preparação e geração do classificador *multiExpert*, utilizando extração por segmentação e a combinação de classificadores. Em seus experimentos, utilizaram-se de duas bases de dados, a *Last.FM* com 5 mil músicas divididas em 10 gêneros balanceados, e a base GTZAN, empregando dois classificadores: o paramétrico *Gaussian Mixture Model* (GMM) e o não paramétrico SVM. Em ambas as bases de dados, o SVM alcançou resultados estatisticamente superiores em relação ao GMM, a saber, 46,2% contra 40,2% na base *Last.FM* e 75,1% contra 61,5% na base GTZAN.

No contexto da utilização de classificação através de características visuais, Nanni et al. (2016) trazem uma abordagem que utiliza um conjunto de classificadores heterogêneos para maximizar o desempenho, criando uma técnica para classificação de imagens que ignora a informação espacial e a ocorrência de elementos no conteúdo. Nesse sentido, os descritores fazem uma combinação através da fusão dos classificadores SVM e subespaço

aleatório de AdaBoost. Avaliando o método proposto nas bases LMD, ISMIR 2004 e GTZAN, obteva-se o melhor resultado para a base LMD com precisão de 86,1%, e 81.6% para ISMIR 2004 e 77.0% GTZAN.

Costa, Oliveira e Silla Jr (2017) utilizam os mesmos recursos de características visuais, como espectrogramas, treinando Redes Neurais Convolutivas (CNN - *Convolutional Neural Networks*) e comparando os resultados obtidos com classificadores de SVM. Os experimentos foram realizados nos bancos de dados ISMIR 2004, LMD e com uma coleção de gravações de músicas étnico-africanas. Os experimentos mostram que a CNN se destaca em relação a outros classificadores em vários cenários, se revelando uma interessante alternativa para o reconhecimento de gêneros musicais. A combinação de CNN e *Robust Local Binary Pattern* superou os resultados obtidos na literatura com o banco de dados de músicas africanas. No caso utilizando o banco de dados LMD, foi alcançada uma assertividade de 92%, novamente o melhor resultado utilizando *artist filter* até o momento. No conjunto de dados do ISMIR 2004, foram apresentados resultados semelhantes aos encontrados na literatura e um desempenho melhor se comparado aos autores que usaram classificadores individuais.

Não havendo um conjunto de dados em larga escala publicamente disponível que inclua anotações de áudio, imagem, texto e multi-label, Oramas et al. (2017) gerou uma base de dados com 250 classes de gênero, denominada MuMU, contendo 147 mil faixas de áudio multimodal, com anotações de gêneros multi-tipos, que combina informações do conjunto de dados *Amazon Reviews* e do *Million Song Dataset* (MSD).

Empregando uma abordagem de aprendizagem profunda com CNN aplicado ao áudio, texto, dados de imagem e suas combinações, avaliando a MuMU sobre um conjunto de 135 mil músicas, Oramas et al. (2017) constatam que o efeito de diferentes parâmetros para CNN's na classificação de áudio constataram que o efeito de diferentes parâmetros superaram as abordagens tradicionais. A base de texto, entretanto, por possuir maiores características semânticas, supera as modalidades, já a classificação baseada em imagens rendeu o menor desempenho, no entanto quando combinada a outras bases auxiliou na melhora dos resultados. Pode-se concluir, portanto, que as abordagens multimodais aparentemente superaram as únicas. Contudo, os autores não informam a assertividade dos métodos empregados em comparativo com outros trabalhos.

A Tabela 2 apresenta de forma condensada os trabalhos descritos nesta sessão, que realizaram rotulação de gênero a partir do sinal ou da letra da canção.

Tabela 2 – Resumo dos trabalhos de classificação automática de gêneros musicais

Autores	Características	Classificador	Bases	Assertividade
(TZANETAKIS; COOK, 2002)	Textura de timbre, conteúdo rítmico e frequência de vibração	GMM e KNN	GTZAN	61,00%
(LI; OGIHARA; LI, 2003)	DWCH	SVM	GTZAN	72,00%
(KOERICH; POITEVIN, 2005)	Textura de timbre, conteúdo rítmico	MLP	GTZAN	65,00%
(FLEXER, 2007)	MFCC	GMM	ISMIR 2004	75,72% e 61,22% c/ artist filter
(SILLA JR.; KAESTNER; KOERICH, 2008a)	Textura de timbre, conteúdo rítmico e frequência de vibração	J48, 3NN, MLP, NB e SVM	LMB	65,00%
(SEYERLEHNER et al., 2010)	Block-level	SVM	GTZAN, ISMIR 2004	82,72% e 77,96%
(LOPES et al., 2010)	MARSYAS c/ Tempo curto, baixo nível	SVM	900 músicas da LMD	59,60 % c/ artist filter
(WU et al., 2011)	GSV e Filtro Gabor	SVM	GTZAN e ISMIR 2004	86,10%
(COSTA et al., 2013)	Descritor de textura LBP, divisão em escala Mel	SVM e KNORA	LMD e ISMIR 2004	83,00% c/ artist filter e 80,65%
(VRIESMANN et al., 2015)	Descritor de textura LBP	KNORA, LCA e OLA	LMD	70,00%, 70,22% e 66,33%
(ROMANO; ADAMI, 2015)	Extração por segmentação, multiExpert e combinação de classificadores	GMM e SVM	5 mil músicas da Last.FM e GTZAN	75,1% e 61,5%
Nanni et al. (2016)	Características Visuais, com descritores de textura	Fusão de classificadores AdaBoot c/ SVM	LMD, ISMIR 2004 e GTZAN	86,10%, 81,60% e 77,00%
Costa, Oliveira e Silla Jr (2017)	Características visuais, com descritores de textura	Redes Neurais Convolutivas e SVM	LMD, ISMIR 2004 e Músicas Africanas	92,00%, 85,90% e 73,00%
Oramas et al. (2017)	Características semânticas de letras, MFCC e características visuais	Redes Neurais Convolutivas	MuMu	Não informado

3 Fundamentação Teórica

Neste capítulo serão apresentados os principais conceitos acerca de representação da emoção, as diferentes perspectivas e as terminologias utilizados. Serão demonstrados também os princípios de aprendizado supervisionado e os principais algoritmos de classificação, assim como os conceitos de Sistemas de Múltiplos Classificadores, com foco em Seleção Dinâmica de Conjunto de Classificadores. Por fim, também serão detalhados os modelos de representações do sinal do áudio e o processamento para extração de recursos.

3.1 Definição de Emoção

As emoções são difíceis de definir e medir, no entanto para os fins deste trabalho, podemos entender emoção como um conceito cotidiano ("teoria popular"), mas também como uma teoria científica. A suposição comum é que existem emoções que nos trazem bem estar e outras que são ruins. Além disso, é comumente aceito que algumas pessoas são mais "emocionais" do que outras, como apontado por Juslin e Sloboda (2001).

Para Peter (2010), as emoções são uma série de alterações no estado do corpo que estão conectadas a imagens mentais que ativaram um determinado subsistema cerebral, por exemplo, o subsistema de processamento de música. Assim, as emoções envolvem reações fisiológicas, mas também são orientadas a objetos, provocando assim, uma categorização: "se a emoção é de medo, seu objeto deve ser visto como prejudicial" (MAYER et al., 2001). Já o humor pode ser considerado um estado emocional duradouro e leva em consideração sentimentos gerais. Humor e emoções podem ser considerados conceitos muito semelhantes em alguns casos, por exemplo, felicidade, tristeza e raiva podem ser vistos tanto como humor, quanto como emoções, no entanto, algumas emoções são consideradas transitórias, como a surpresa se a considerarmos como uma emoção.

Existe uma visão comum de que a principal função das emoções é guiar o comportamento e este comportamento, provocado pelas emoções, foi desenvolvido com relação à interação bem-sucedida com o meio ambiente, servindo funções que nem sempre são conscientes e raramente são intencionais (Lin; Chung; Chen, 2018).

No estudo de Zhang et al. (2010), a distinção entre a emoção percebida e a induzida é que a induzida pode se evidenciar pelas mudanças fisiológicas quando por exemplo, os batimentos cardíacos aceleram, ou até mesmo o desencadear de um choro ou o arrepio. No entanto para Konečni (2008), a emoção induzida, pode ser oposta à emoção que supostamente é expressa pela canção. Por exemplo, uma canção que expressa alegria, pode causar angústia ou tristeza a um indivíduo que tenha vivenciado um momento triste

ouvindo a música.

Assim, a distinção entre a indução e a percepção emocional é postulada como quantitativa, sendo a indução a possibilidade de a emoção ser explicada pelo mecanismo de condicionamento (JUSLIN; VÄSTFJÄLL, 2008). Em outras palavras, emoção induzida é mais complexa de avaliar, por considerar o contexto e o que a música lhe remete, pois como no exemplo, mesmo que a música seja de alegria, ela pode induzir a um estado de tristeza. Por isso, neste trabalho foi considerada a emoção percebida.

Para Juslin e Sloboda (2001), a variação da estrutura musical pode fazer com que os ouvintes percebam e identifiquem corretamente uma expressão emocional específica pretendida, no entanto, a percepção da emoção na expressão musical não levará necessariamente à mesma percepção emocional para todos.

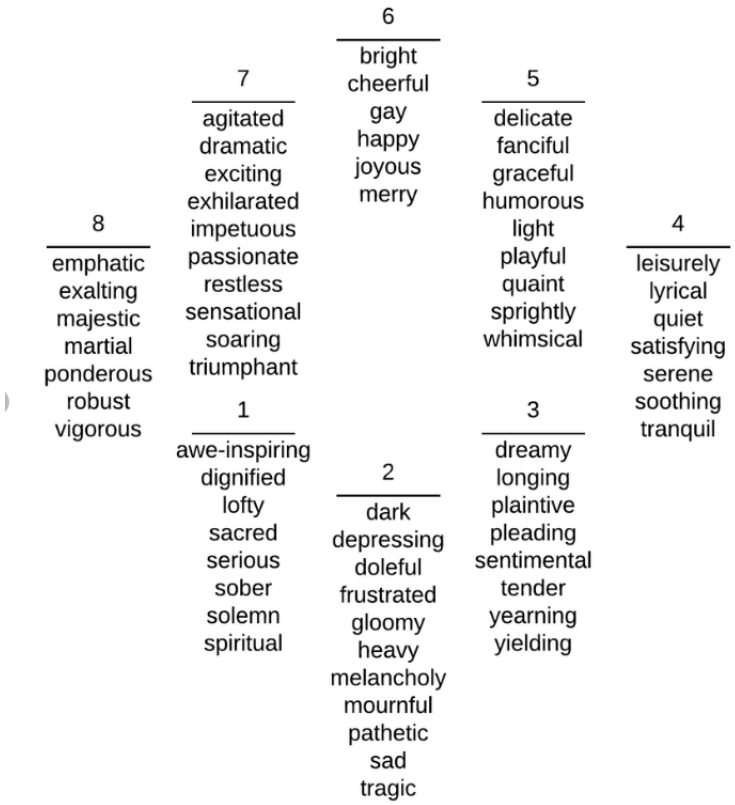
3.2 Representação da Emoção

Uma das primeiras perguntas que vem à mente quando pensamos em formalizar emoções é: como podemos representá-las? Inclusive, na literatura sobre psicologia da música existem dois principais paradigmas para representar emoções, o primeiro é a representação categórica que faz a distinção dentre as várias classes de emoção, a outra é a representação dimensional, que define um espaço emocional. Essa distinção é bastante geral, não se tratando apenas de emoções musicais, mas de estudos que foram projetados especificamente para testar e refinar esses modelos para a música (KIM et al., 2010). Detalhamos neste trabalho as principais teorias usando ambas as abordagens e explicitamos o caso especial das representações emocionais relacionadas à música.

A representação categórica visa dividir emoções em categorias onde cada emoção é rotulada com um ou vários adjetivos. O modelo mais canônico é o conceito de emoções básicas, onde várias categorias distintas são a base de todas as emoções possíveis. Muitos psicólogos propõem que seu conjunto de adjetivos emocionais seja aplicável à música e um dos trabalhos mais relevantes nesse domínio é o estudo de Hevner (1936) e seu círculo adjetivo, mostrado na Figura 1. A lista de adjetivos de Hevner é composta por 67 palavras organizadas em oito grupos e a partir deste estudo, cada *cluster* inclui adjetivos que têm um relacionamento próximo, a semelhança entre palavras do mesmo *cluster* permite trabalhar no nível do próprio *cluster*, reduzindo a taxonomia para oito categorias.

No modelo de Watson e Tellegen (1985), outra variação de modelo circumplexo – que apresenta as várias emoções distribuídas numa circunferência – as emoções distribuem-se na área de circunferência em torno de dois eixos ortogonais: um eixo designado de alta afetividade positiva versus baixa afetividade positiva; e o outro eixo designado de alta afetividade negativa versus baixa afetividade negativa (TELLEGEN; WATSON; CLARK, 1999).

Figura 1 – Modelo de Hevner (1936): sessenta e seis palavras dispostas em oito grupos.

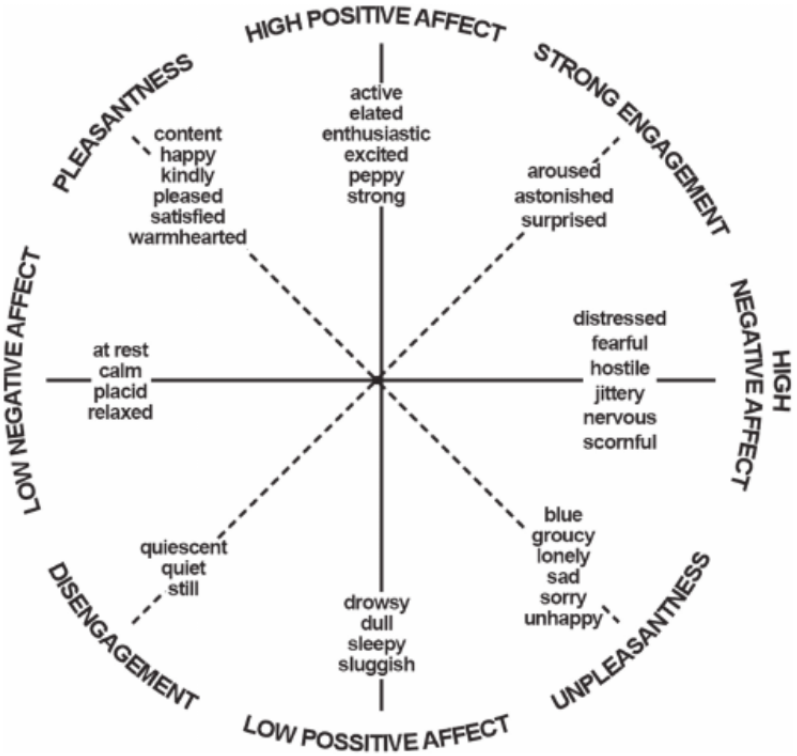


Fonte: (HEVNER, 1936)

No modelo circumplexo, a proximidade ou a distância entre as emoções representadas na circunferência pressupõem a semelhança ou a diferença entre elas. O último modelo apresentado define que as emoções estão menos positivamente correlacionadas quando estão afastadas em aproximadamente 90 graus entre si, pois neste grau de afastamento dois estados afetivos devem estar muito pouco ou nada correlacionados. Por sua vez, aos 180 graus de afastamento, os estados afetivos devem estar negativamente correlacionados (TELLEGEN; WATSON; CLARK, 1999). Na Figura 2 podemos observar a distribuição do modelo de Watson et al. (1999).

A extrema dificuldade em definir uma emoção também ocorre para as pessoas, havendo então uma necessidade de fazer relações entre outros tipos de emoção, as sobrepondo Russell (1980) e ao descobrir essa intercorrelação entre as emoções, notou-se a necessidade de classificá-las com o uso de Modelos Dimensionais, considerando assim emoções relacionadas em um limite determinado e, em certas vezes, suas ambiguidades. Dentro da psicologia existem alguns Modelos Dimensionais de Emoções, entre eles alguns dos utilizados são os de Russell (1980) e Thayer (1989), ambos definidos por dimensão, e

Figura 2 – Modelo circumplexo de Watson e Tellegen (1985)



Fonte: (WATSON et al., 1999)

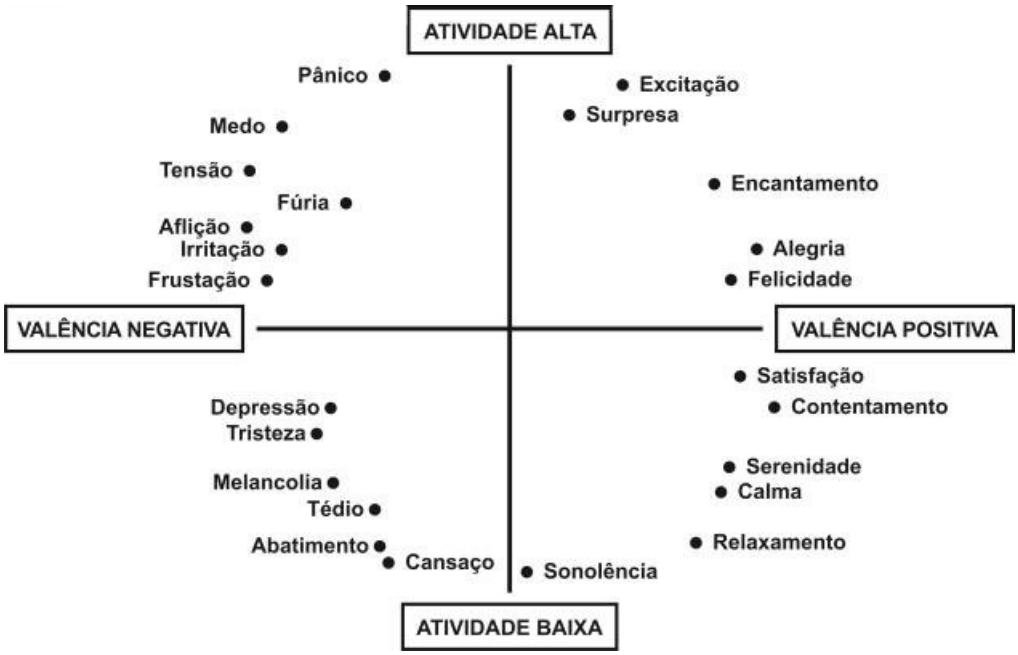
utilizados para representar as emoções básicas, conforme podemos observar na Figura 3.

No espaço bidimensional é abrangida a excitação (*arousal*), atividade de excitação da emoção e a valência (*valence*), positividade ou negatividade da emoção (RUSSELL, 1980). A representação de uma emoção ocorre, por exemplo, num ponto no canto superior direito, que tem alta valência (*valence*) e excitação (*arousal*), o que significa feliz (*happy*) com uma atividade alta, como animado (*excited*). Oposto a este, a parte inferior esquerda é indicada como negativa, com baixa atividade como entediado (*bored*) ou deprimido (*depressed*). Assim, o sistema é interpretado como a representação de uma emoção particular e única (KIM et al., 2010).

Alguns estudos propõem outras representações dimensionais, sendo o modelo de Thayer (1989) uma variante do modelo de Russell (1980), que aplica essa abordagem dimensional e desenvolve a ideia de um modelo de estresse energético. No entanto, todos os estudos apresentados se relacionam de alguma forma com o modelo apresentado anteriormente. No caso do espaço bidimensional de emoção proposto pela aplicação de música Gracenote[1], ocorre distribuindo 25 tipos de emoções entre o eixo bidimensional, conforme demonstrado na Figura 4. Como pode ser visto a partir da figura, alta valência e

[1] https://www.gracenote.com/auto/music-recognition-auto/

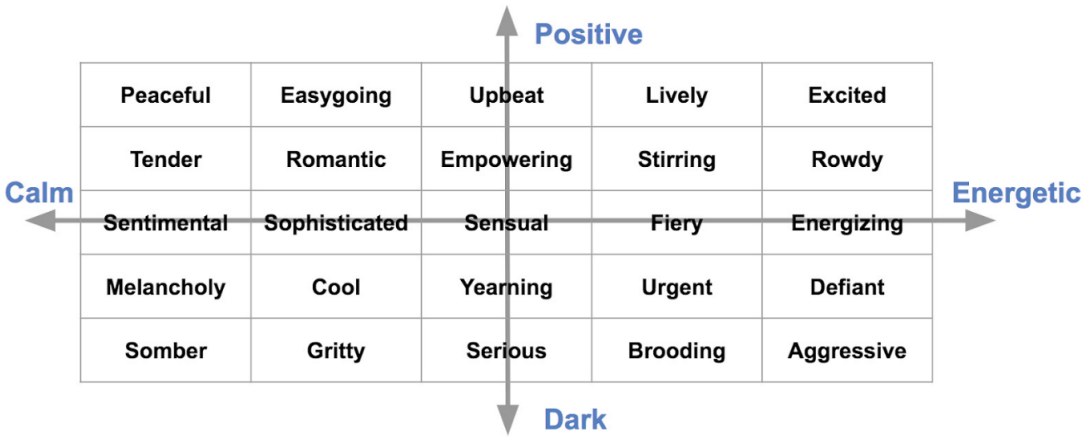
Figura 3 – Modelo de emoção de Russell - Representação dimensional, as emoções são classificadas ao longo do eixo. É importante destacar que em uma representação dimensional, as emoções são classificadas ao longo do eixo.



Fonte: (RUSSELL, 1980)

alta excitação no canto superior direito correspondem ao humor musical mais positivo e mais enérgico (*excited*). Desta mesma forma, outras emoções também podem ser encontradas nos eixos, dependendo da valência e da excitação.

Figura 4 – Representação gráfica do "espaço de emoção bidimensional" modelo de Thayers 2D.



Fonte: (TONG et al., 2018)

A principal vantagem de representar a emoção em uma forma dimensional é que qualquer emoção pode ser mapeada no espaço através das coordenadas X e Y. Uma crítica comum a essa abordagem é que emoções muito diferentes não apenas em termos de significado semântico, mas também em termos de mecanismos psicológicos e cognitivos envolvidos, podem estar próximas no espaço emocional. Por exemplo, observando a Figura 3, é possível identificar que a distância entre zangado (*sad*) e depressivo (*depressed*) é pequena, embora estas duas emoções sejam bem diferentes.

No reconhecimento de emoções esses modelos são utilizados tanto para reconhecimento quanto para classificação. No caso da música o reconhecimento é visto como um problema de múltiplas classes, pois é possível considerar a música inteira ou então por partes. Por exemplo, Lu, Liu e Zhang (2006) e Hu et al. (2008) selecionam apenas faixas que contenham uma emoção clara, para que haja concordância subjetiva.

3.3 Classificação

Nesta sessão serão abordados os conceitos básicos de aprendizagem supervisionada para a classificação de uma amostra de teste, ou seja, a rotulação da amostra dado o seu vetor de características. A classificação é um procedimento de aprendizagem baseado na teoria da aprendizagem estatística e pode ser entendido como o processo que atribui uma determinada classe C_i a um conjunto de características x, extraídas a partir de uma amostra a ser classificada. Para atribuição dessas categorias um sistema de classificação se baseia em etapas bem definidas:

- Pré-processamento dos dados a serem analisados;
- Extração de características;
- Classificação das amostras;
- Avaliação de resultados;

Na parte seguinte desta seção, são destacados os processos de extração de características, apresentando os algoritmos de classificação mais comuns, cobrindo uma grande variedade de classificadores que serão utilizados neste trabalho, as técnicas com multi-classificadores, a seleção dinâmica de subconjunto de classificadores e as principais formas de avaliação.

3.3.1 Extração de Características

Conforme aponta a literatura, a etapa de extração de características é fundamental para o desenvolvimento de sistemas de reconhecimento de padrão Scaringella, Zoia e Mlynek

(2006), assim como é conhecido também que a qualidade dos conjuntos de dados extraídos tem influência direta no processo de reconhecimento de padrões, pois, se executado de forma inadequada, a acurácia será seriamente prejudicada (MACIÀ; ORRIOLS-PUIG; BERNADÓ-MANSILLA, 2010). Na etapa de extração de características é produzida uma representação do conteúdo, por exemplo, através das ondas sonoras de uma música é possível extrair um vetor que representa a variação de timbre, que pode ser utilizado para treinar um classificador.

Neste trabalho, para reconhecimento dos rótulos musicais as características usadas são: acústica e domínio visual através dos espectrogramas e das letras.

3.3.1.1 Acústicas

Para o reconhecimento de emoção e gênero, algumas características acústicas, são de importante consideração no processo de extração, como a harmonia, o timbre, ritmo e batida.

A harmonia pode ser definida como uma sucessão de eventos percebidos como entidade única, por vezes referida como o elemento vertical da música, sendo a melodia o elemento horizontal. As análises melódica e harmônica são utilizadas há muito tempo pelos musicólogos para estudar estruturas musicais (SCARINGELLA; ZOIA; MLYNEK, 2006). O timbre é definido como a qualidade que permite distinguir entre sons de um mesmo nível e o volume, quando feitos por diferentes instrumentos musicais ou vozes (ZHANG; KUO, 1999), sendo considerado como a característica mais importante na diferenciação de classes de sons ambientais e também para a construção de um modelo adequado para a percepção do timbre (ZHANG; KUO, 1999).

A batida e a estrutura rítmica de uma música costumam ser bons indicadores do gênero, mas para Tzanetakis e Cook (2002) algumas características devem ser consideradas para a construção do vetor referente à rítmica, deve-se encontrar a batida principal da música e o seu período em BPM (batidas por minuto), relacionando-o à segunda batida mais forte, obtendo-se uma série de características em relação a primeira e a segunda batida, decompondo em várias bandas de frequência que são adicionadas a um histograma de batidas (COSTA; VALLE; KOERICH, 2004).

3.3.1.2 Domínio Visual

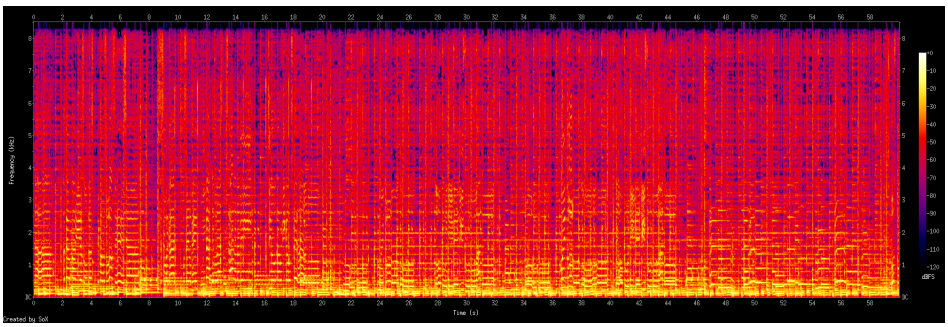
A definição de um conjunto de recursos visuais capazes de descrever imagens de maneira eficaz, de modo que os processos de reconhecimento de padrões possam ser aplicados, é uma tarefa complexa na análise de imagens. Uma maneira de resolver esse problema é introduzida no trabalho de Portilla e Simoncelli (2000), com o conceito de análise de imagens através de informação de conteúdo extraída somente do aspecto visual, possibilitando a manipulação de dados a partir de um conjunto de características extraídas

de uma determinada amostra. O reconhecimento de padrões baseado em textura segue o mesmo princípio, de comparação de semelhança, no qual a partir do espectrograma é obtido um vetor de características que representa a amostra.

Por apresentar bons resultados na investigação do recurso realizada por Costa et al. (2013), neste trabalho são utilizadas características para a representação do conteúdo das músicas, obtidas a partir de imagens através de espectrograma geradas a partir do sinal do áudio. Portanto o processamento desta representação utiliza o sinal de áudio de 60 segundos inicialmente extraído da música, convertido em uma imagem de espectrograma, o espectro de frequências variando com o tempo, que pode ser descrito por um gráfico com duas dimensões geométricas: o eixo horizontal representando o tempo e o eixo vertical representando a frequência. Uma terceira dimensão que descreve a amplitude do sinal em uma frequência específica de um determinado momento é representada pela intensidade de cada ponto na imagem (NANNI et al., 2016). Para geração de espectrograma, a *Transformada Discreta de Fourier* é calculada com um tamanho de janela de 1024 amostras, utilizando a função de janela *Hanning*, que possui boas propriedades de resolução de frequência e faixa dinâmica (NANNI et al., 2016).

A Figura 5 mostra espectrograma retirado de amostra de áudio.

Figura 5 – Exemplo de espectrograma de áudio



Como é possível notar, as Figuras 6, 7 e 8 apresentam diferenças significativas, com presença de linhas que parecem representar dimensões distintas e que podemos associar aos gêneros dos quais foram extraídas. O espectrograma 6 foi extraído de uma música de Bolero, nele é possível observar que as linhas horizontais são predominantes e que existe um espaçamento entre as intensidades, presumidamente relacionando as estruturas hadrônicas e as intensidades espaçadas, muito presentes nesse tipo de gênero musical. Na Figura 7, extraída de uma música de Axé, nota-se uma presença maior de linhas verticais, que estão relacionadas às batidas, comuns neste gênero. E na Figura 8, no gênero Salsa, é possível observar uma ondulação devido as suas características harmônicas e ritmo serem crescentes, havendo algumas pausas.

Essas diferenças apresentadas são caracterizadas como diferenças de textura, que

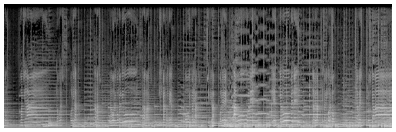


Figura 6 – Bolero

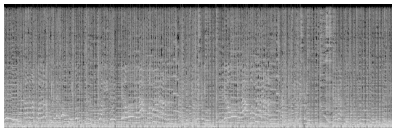


Figura 7 – Axé

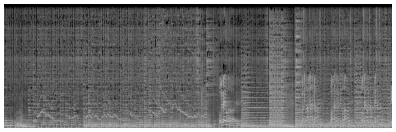


Figura 8 – Salsa

podem ser definidas como o aspecto de uma superfície, frequentemente relacionadas à distribuição de características visuais em uma determinada representação (WANG; HE, 1990). O conceito básico de textura será considerado uma distribuição entrelaçada das intensidades dos *pixels* (KARKANIS; GALOUSI; MAROULIS, 1999) e uma das formas de representar estas características é quantificar o conteúdo de texturas descritas em uma determinada região. Dentre as formas possíveis, é utilizada a representação estrutural, que também é a forma utilizada neste trabalho.

3.3.1.2.1 Representação Estrutural

O uso de técnicas estruturais com descritores de textura se justifica pelo fato deste ser o principal atributo perceptível na imagem de espectrograma, definindo a textura como composta de primitivos, também conhecidos como *texton*. Depois de identificar os *textons* que compõem a textura, duas classes de métodos podem ser consideradas para a extração de características. O primeiro utiliza descritores extraídos das *textons* para descrever a textura, enquanto o segundo considera regras para descrever a disposição espacial destes *textons* (SCHWARTZ; SIQUEIRA; PEDRINI, 2012).

A seguir está descrita a abordagem LBP, que obtiver os melhores resultados na classificação de gênero, conforme observado por Costa, Oliveira e Silla Jr (2017), por isso foram utilizados estes mesmos princípios para a representação dos rótulos neste trabalho.

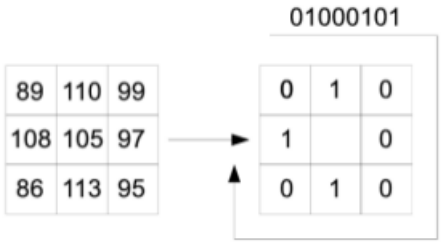
Local Binary Pattern - (LBP)

O *Local Binary Pattern* (LBP) é definido como uma medida de textura invariante em tons de cinza, derivada de uma definição geral de textura. Para cada pixel em uma imagem, um código binário é produzido pelo limiar de seu valor com o valor do pixel central (WU; SUN, 2009) e um histograma é criado para coletar as ocorrências de diferentes padrões binários previstos, gerando um vetor final de características corresponde à este histograma normalizado (OJALA; PIETIKÄINEN; HARWOOD, 1996). O LBP original é formado analisando o formato do bloco de imagem com a vizinhança 3X3, considerando um *pixel* central. Assim, o mesmo atribui 0 ou 1 como valor aos 8 *pixels* vizinhos conforme demonstrado na Equação 3.1:

$$\begin{cases} 0 & se \quad g_n\, n < g_c \\ 1 & se \quad g_n\, n \geq g_c \end{cases} \tag{3.1}$$

Considerando n um binário atribuído ao *pixel* vizinho, g_n o valor do nível de cinza do *pixel* vizinho e g_c o valor do nível de cinza do *pixel* central (LENC; KRÁL, 2014), os valores resultantes são da diferença entre seu valor e o valor do *pixel* central, multiplicado pelo peso da posição correspondente à porção da vizinhança, determinando assim parte da somatória para a obtenção do operador. Ao final, os valores resultantes são linearizados e concatenados em um vetor binário de 8 bits (LENC; KRÁL, 2014), conforme podemos observar na Figura 9.

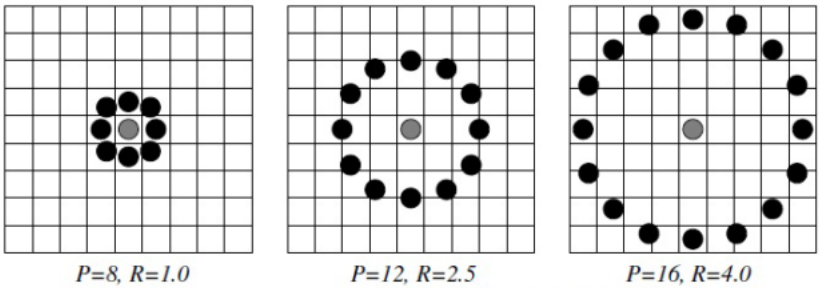
Figura 9 – Exemplo de extração original do LBP



Fonte: (LENC; KRÁL, 2014)

Por definição, o LBP é invariante para qualquer transformação monotônica da escala de cinza e é rápido para calcular, assim os operadores com diferentes parâmetros podem ser combinados para obter uma descrição multiescala das texturas (WU; SUN, 2009). Na Figura 10, são demonstradas algumas variações possíveis para LBP, sendo R corresponde à distância entre o *pixel* central e os vizinhos a serem tomados e P a quantidade de vizinhos a serem considerados:

Figura 10 – Exemplos de possíveis vizinhanças utilizadas em LBP



Fonte: (MAENPAA; PIETIKAINEN, 2005)

Entre os as variações de LBP utilizadas, a mais difundida será utilizada nos experimentos, que é a $LBP_{8,2}$. Este padrão considera oito vizinhos a uma distância de dois pixels a partir de cada pixel da imagem.

3.3.2 Através das Letras

Neste contexto descartamos a música instrumental e nos concentramos nas músicas com letras e com a ideia de investigar se as letras podem auxiliar na rotulação musical e se tornarem complementares ao uso do sinal de áudio como fonte de informação. Comumente utilizada para o processamento de texto, o método *Bag-of-Words* consiste na representação de cada documento como um vetor de palavras que ocorre no documento. Para definir a similaridade entre diferentes músicas, uma forma comumente usada nas tarefas de classificação de documentos é a utilização da representação das músicas como uma *bag* de palavras, ou seja, um conjunto de palavras ou termos usados em uma música e a sua frequência (SALTON, 1971).

3.3.2.1 Bag-of-Words

Utilizando a *bag* por frequência, aplicando a um modelo vetorial para classificar a palavras presente em uma coleção de palavras de um documento pela sua relevância e calculando o TF/IDF, é possível atribuir desta forma mais importância a uma palavra, ou aos termos que são frequentes em determinada música, mas menos frequentes no conjunto da coleção. TF (*Term Frequency*) significa em uma tradução livre “frequência do termo” e IDF (*Inverse Document Frequency*) para "frequência do documento inverso". O IDF é definido como:

$$idf(t) = log\frac{D}{|\{d \ : \ t \ \epsilon \ d|\}} \tag{3.2}$$

Onde $|D|$ é o número total de documentos e $|\{d : t \ \epsilon \ d|\}$ o número de documentos em que o termo t aparece. Note que isto é definido apenas para um *corpus* onde t aparece. Então, o peso TF/IDF de um termo t em um documento d é dado por:

$$tf - idf(t, d) = tf(t, d) \ \times \ idf(t) \tag{3.3}$$

3.3.2.2 N-Grans

Um conjunto de co-ocorrência de termos com comprimentos pré-definidos, é denominado *n-gram*, ou seja, uma sequência de palavras ou sílabas de comprimento N (CAVNAR; TRENKLE et al., 1994). Para Alemayehu e Willett (2003) uma vez que cada sequência é decomposta em pequenas partes, todos os erros presentes tendem a impactar apenas um número limitado de partes da palavra, possibilitando minimizar erros quando nos conjuntos de palavras existem idiomas distintos. Essa abordagem permite inferir os morfemas dividindo as palavras em uma sequência de caracteres de tamanho fixo "n"que é pré-definido de acordo com o contexto, por isso, geralmente essa técnica é utilizada em casos de documentos com idiomas distintos (TAVARES; LOPES; LIMA, 2007). Utilizando

a técnica proposta por Cavnar, Trenkle et al. (1994), para definir o início e o fim de uma palavra utiliza-se o caractere underline ("_"). Desta forma a palavra "AMOR" seria composta por:

bi-grams: _*A*, _*AM*, MO, *OR*_, *R*_

tri-grams: _*AM*,_*AMO*, *MOR*_, OR, *R*_

3.3.2.3 Stemming

Stemmer, utilizada para redução de ruídos na classificação utilizando as letras musicais é a *stemming*, que consiste na redução da dimensionalidade de uma palavra, obtendo o radical de todas as palavras, por exemplo, a palavra "computação" e "computador", o resultado do algoritmo poderia ser "comput" (PORTER; FEIG, 1980).

3.3.3 Classificação das Amostras

Classificação de amostras é o processo de atribuição de rótulos a objetos, também chamados de classes, onde esses objetos são descritos por conjuntos de medidas chamadas de atributos ou características (KUNCHEVA; RODRIGUEZ, 2007).

A classificação de uma amostra ocorre em duas fases distintas: fase de aprendizagem e fase de reconhecimento. Na primeira fase é construído um modelo classificador a partir de um algoritmo de aprendizado, em outras palavras, o algoritmo, através das características das amostras de treinamento, aprende a como separar e distinguir uma amostra. Na segunda fase, a de reconhecimento, após a aprendizagem o classificador é aplicado sobre novas amostras para determinar a respectiva classe. Diferentes algoritmos utilizam técnicas e paradigmas diferentes desenvolvidos para captação do conhecimento. As próximas subseções descrevem os principais algoritmos utilizados na fase experimental deste trabalho.

É importante ainda ressaltar que, antes da realizar as tarefas de classificação, todos os algoritmos foram otimizados utilizando um procedimento *grid-search*[2].

3.3.3.1 Decision Trees (J48, C4.5)

O algoritmo *Decision trees* (J48, C4.5), ou em português "árvore de decisão", divide o conjunto de dados de treinamento em subconjuntos com base em um valor de atributo de teste. Esse processo é repetido em cada subconjunto de maneira recursiva, criando um novo subconjunto dividido pelo atributo de forma recursiva. A árvore de decisão classifica uma nova amostra a partir do nó principal (raiz), sendo que cada nó descendente desse nó é um dos valores possíveis para esse atributo.

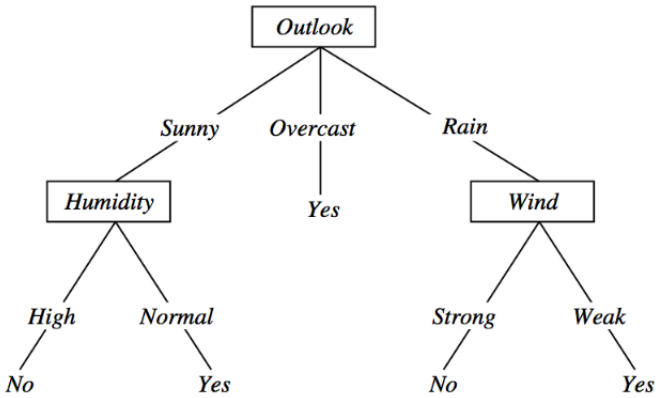
[2] Método utilizado para encontrar os parâmetros ideais de um modelo que resultam nas previsões mais "precisas".

O Algoritmo da árvore de decisão pode ser definido como um método para aproximar as funções-alvo de valor discreto, no qual a função aprendida é representada por uma árvore de decisão. Assim, podemos entender o algoritmo da árvore de decisão como um método para aproximar funções-alvo de valor discreto, onde a função aprendida é representada por uma árvore de decisão (MITCHELL, 1999).

Uma vez aprendidas, as funções-alvo podem ser facilmente implementadas através de um conjuntos de regras (se, então), usando então uma implementação da árvore de decisão C4.5 da biblioteca *Python Sklearn*, chamada J48 no software *Java Weka* 3 (SALZBERG, 1994). Para melhorar a arvore de decisão, são utilizados dois parâmetros principais: "C", o fator de confiança usado para a poda, isto é, limitando o crescimento da árvore, de 0,1 a 0,5 em 10 etapas e "M", o número mínimo de instâncias por folha, de 2 a 20.

Na Figura 11 mostramos um exemplo típico de uma árvore de decisão.

Figura 11 – Exemplo do algoritmo de árvore de decisão para classificar se é um bom dia para jogar tênis.

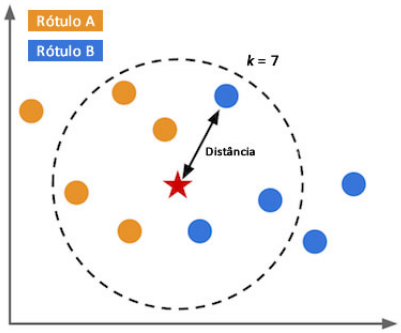


Fonte: (MITCHELL, 1999)

3.3.3.2 k-Nearest Neighbor - k-NN

O algoritmo k-NN, ou em tradução livre "k vizinhos mais próximos", é um dos algoritmos que provavelmente possui o classificador mais simples (FIX; JR, 1951). Trata-se de um algoritmo não paramétrico, sendo que a estrutura do seu modelo é determinada pelo conjunto de amostras de treinamento utilizado. Para cada nova amostra de teste, o algoritmo k-NN procura um número k de suas amostras de treinamento mais próximas para decidir sobre a escolha da classe, sendo então classificado de acordo com a classe mais comum dos k vizinhos mais próximos, considerando que k será um número inteiro positivo.

Figura 12 – Exemplo do algoritmo de k-NN para classificar uma nova amostra.



Fonte: (PACHECO, 2017)

Na Figura 12 é demonstrado o funcionamento do algoritmo, no qual se tem a classificação com uma amostra a partir de dois rótulos de classe, considerando $k = 7$. A nova amostra é representada por uma estrela e as demais amostras de treinamento são representadas pelas bolinhas azuis e amarelas. A variável k representa a n de vizinhos mais próximos, que serão utilizados para definir a qual classe a nova amostra pertence (PACHECO, 2017). Com isso, das sete amostras de treinamento mais próximas da nova amostra, 4 são do rótulo A e 3 do rótulo B. Portanto, como existem mais vizinhos do rótulo A , a nova amostra receberá o mesmo rótulo deles, ou seja, A.

3.3.3.3 Gaussian Mixture Models

Um *Gaussian Mixture Models* (GMM), modelo de mistura de gaussianas, é uma combinação linear de distribuições de probabilidades gaussianas, usado como método de agrupamento baseado em distribuições estatísticas específicas (PEEL; MCLACHLAN, 2000). Ao modelar a função de densidade de probabilidade de um conjunto de dados, o GMM automaticamente realiza um agrupamento do conjunto, discriminando qual componente da mistura gerou cada elemento.

O GMM se baseia na suposição de que cada elemento do conjunto se origina a partir de um componente da mistura com uma determinada probabilidade. Assim, ao inferir os parâmetros da mistura, essa probabilidade pode ser usada para associar cada elemento ao componente com maior probabilidade de tê-lo gerado. Na fase de treinamento, é utilizado o algoritmo *Expectation-Maximization* como parâmetro para cada classe, que são aprendidas e recalculadas em cada interação (REYNOLDS; QUATIERI; DUNN, 2000).

3.3.3.4 Support Vector Machines - SVM

As *Support Vector Machines* (SVMs), máquinas de vetores de suporte, constituem uma técnica de aprendizado para problemas de reconhecimento de padrão. A técnica foi introduzida por Schölkopf, Burges e Smola (1999) e é essencialmente uma abordagem geométrica para o problema da classificação.

A SVM foi originalmente desenvolvida para a classificação binária, na qual existem somente duas classes para decisão, no entanto a maioria dos cenários envolvendo reconhecimento são problemas de multi-classes. Com o objetivo de solucionar problemas envolvendo um maior número de classes possível, foram estendidos os critérios de avaliação da SVM original, assim o mecanismo se baseou na construção de modelos distintos associados aos grupos de amostras, combinados para descrição de subconjuntos de características. Cada um dos modelos construídos é um classificador binário associado a uma das classes envolvidas no sistema. A construção dos modelos é dividida em dois processos: treinamento e teste. O processo de treinamento consiste na utilização de um conjunto de vetores para a construção de modelos de cada uma das classes e o processo de teste na atribuição de uma classe para a amostra, com base nas características extraídas (VAPNIK, 1999).

A SVM presente neste trabalho, utilizar kernel com *Radial Basis Function* (RBF) e os parâmetros C (custo) e γ otimizados.

3.3.3.5 Logistic Regression

A regressão logística pode prever a probabilidade de ocorrência de um evento ajustando os dados a uma curva logística. É um modelo linear generalizado, usado para regressão binomial, que tenta modelar os dados na função logística:

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} \tag{3.4}$$

Sendo que z é definido por:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{3.5}$$

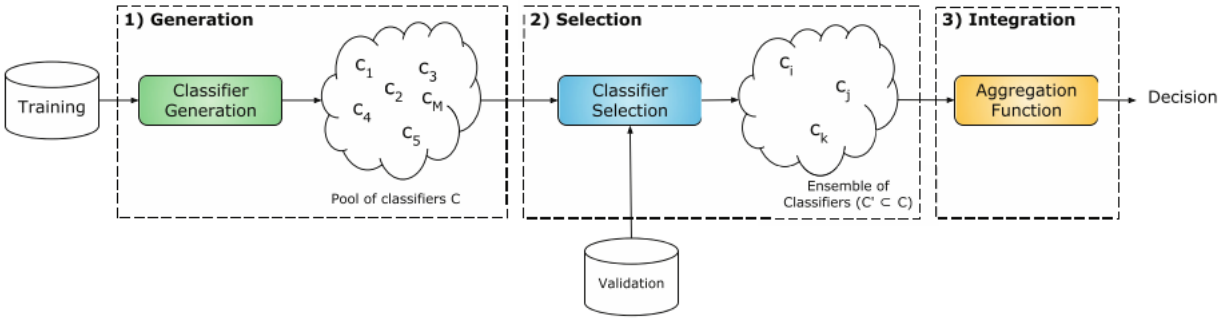
Como uma constante β_0 e β_1,β_2, ... β_k são coeficientes de regressão dos valores correspondentes: $\beta_1, \beta_2, \beta_3 ... \beta_k$. No contexto da classificação, a curva logística modela a relação entre um conjunto de variáveis e uma resposta binária expressa como probabilidade e este valor binário é a saída do classificador. Para tanto, fora utilizada a implementação da biblioteca *Python* Pedregosa et al. (2011), com parâmetros otimizados para multi-classes.

3.3.4 Sistemas de Múltiplos Classificadores

Os resultados das pesquisas mais recentes nos levam a uma conclusão: criar um classificador monolítico com o objetivo de cobrir toda a variabilidade inerente à maioria dos problemas de reconhecimento de padrões é algo impraticável (CAVALIN; SABOURIN; SUEN, 2013). Um dos principais problemas, segundo Dietterich (2000) é a construção de bons conjuntos de classificadores. De acordo com Hansen e Salamon (1990), existe uma condição necessária e suficiente para que conjuntos de classificadores tenham melhor desempenho que seus elementos individuais: os mesmos devem ser acurados, com um nível de precisão superior a 50%, e diversificados. Como normalmente esta condição se faz presente nos conjuntos, são observados acréscimos na assertividade dos classificadores.

Com isso em mente, muitos pesquisadores se concentraram nos Sistemas de Múltiplos Classificadores (SMC) e como consequência, novas soluções foram estudadas para todas as fases da classificação:

Figura 13 – (1) geração, (2) seleção e (3) integração



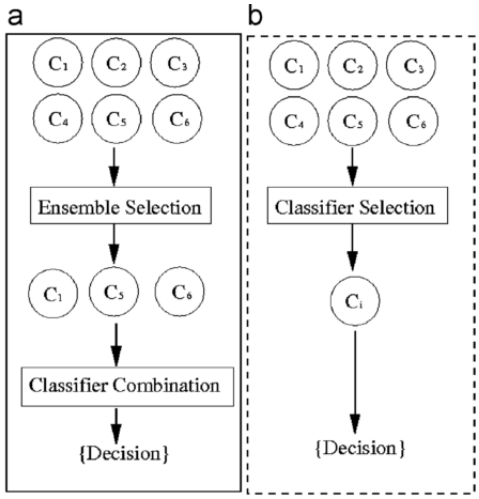
Fonte: (CRUZ; SABOURIN; CAVALCANTI, 2018)

Conforme apresentado na Figura 13, na primeira fase um conjunto de classificadores é gerado; na segunda fase, um subconjunto desses classificadores é selecionado; enquanto que na última fase, uma decisão final é feita com base na previsão, ou previsões, do(s) classificador(es) selecionado(s) (BRITTO; SABOURIN; OLIVEIRA, 2014).

Na primeira fase (*Generation*), são construídos os classificadores, isso pode ser feito de forma homogênea, consistindo em vários classificadores que utilizarão o mesmo algoritmo de aprendizado, ou de forma heterogênea, combinando algoritmos distintos para assim construir os classificadores. A ideia é gerar classificadores que cometam erros diferentes e, consequentemente, mostrar algum grau de complementaridade (KO; SABOURIN; BRITTO JR, 2008).

A segunda fase (*Selection*), é responsável por selecionar o(s) classificador(es), podendo ser realizada de forma dinâmica ou estática. Na seleção estática a classificação

Figura 14 – Os diferentes esquemas para seleção e combinação de classificadores: (a) seleção de conjuntos estáticos; (b) seleção de classificador dinâmico;



Fonte: (KO; SABOURIN; BRITTO JR, 2008)

de todas as amostras influencia na seleção do grupo escolhido, representada na Figura 14.a. No entanto, neste trabalho o foco está na seleção dinâmica, um conceito inicialmente introduzido por Ho, Hull e Srihari (1994); que consiste na escolha do classificador por base da instância de teste, visando minimizar os erros dos classificadores em comparação à seleção estática (TSOUMAKAS; PARTALAS; VLAHAVAS, 2008), representada na Figura 14.b. Na Seção 3.3.5, os métodos de seleção dinâmica de classificadores serão mais explorados.

A terceira fase (*Integration*) consiste na combinação dos classificadores, possibilitando obter conjuntos que tomem decisões mais acuradas (BRITTO; SABOURIN; OLIVEIRA, 2014) (KITTLER et al., 1998). Existem cenários em que a fase de integração não se faz necessária, como quando somente um classificador é selecionado na segunda fase (BRITTO; SABOURIN; OLIVEIRA, 2014).

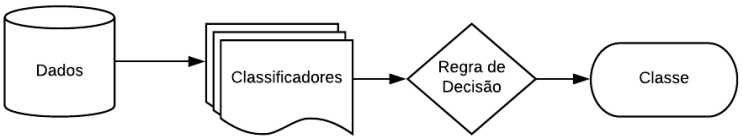
A combinação pode ocorrer através da fusão dos classificadores, isso quando todos os classificadores, dado algum critério, contribuem na decisão do rótulo para a instância, ou através da seleção do classificador do conjunto para atribuir o padrão da amostra (PONTI JR; P, 2011). Essa combinação só será efetiva quando os classificadores individuais forem acurados e variados, neste caso, utilizam-se diferentes espaços de características, diferentes conjuntos de aprendizagem ou variados classificadores, mudando a configuração ou o tipo de classificador (TUMER; GHOSH, 1996).

Os classificadores podem ser combinados seguindo três abordagens, conforme Lu (1996), Ponti Jr e P (2011): paralela, serial e híbrida.

A forma mais simples para combinação é a paralela, conforme podemos observar na Figura 15, que utiliza as regras: soma, produto do mínimo, máximo, média e mediana (GUNES et al., 2003). Também é utilizado o voto da maioria, voto ponderado, *edge count*, inferência bayesiana e teoria das crenças (GUNES et al., 2003).

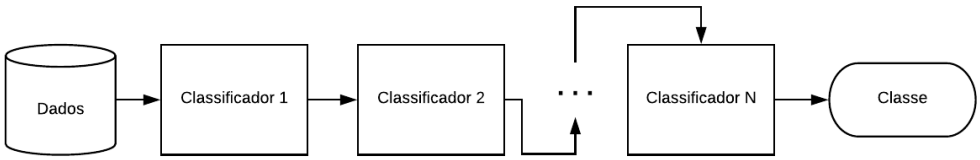
A forma mais simples de combinação é a paralela, conforme podemos observar na Figura 15, que utiliza as regras: soma, produto do mínimo, máximo, média e mediana (GUNES et al., 2003). Também é utilizado o voto da maioria, voto ponderado, *edge count*, inferência bayesiana e teoria das crenças (GUNES et al., 2003).

Figura 15 – Topologia paralela



Na forma serial, observada na Figura 16, os classificadores são colocados em forma crescente de complexidade, partindo do classificador mais simples, e os elementos rejeitados, a partir de um patamar previamente estabelecido, são submetidos ao classificador seguinte. A cada interação são reduzidos os números de instâncias e desta forma são reavaliados. O processo é realizado até que não ocorram mais rejeições (RANAWANA; PALADE, 2006).

Figura 16 – Topologia Serial



Buscando uma melhora no desempenho propõe-se uma topologia híbrida, combinando técnicas paralelas e seriais, oferecendo assim a possibilidade de checar os erros em cada interação e, por consequência, evitar a propagação equivocada, o que poderia ocorrer utilizando somente a forma serial.

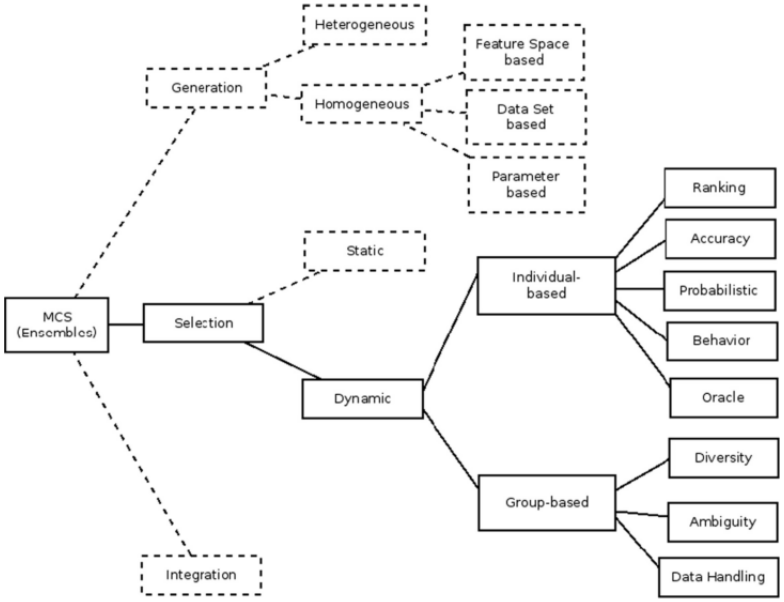
3.3.5 Seleção Dinâmica de Classificadores

A seleção dinâmica de conjunto de classificadores tem se destacado se comparada à seleção estática, isso ocorre devido a sua capacidade de formular uma decisão subsidiada em diferentes soluções de seleção, de acordo com o padrão de teste. Sabe-se que diferentes

padrões de teste são associados a diferentes dificuldades e assim, apresentar uma solução de seleção de classificadores para cada padrão de teste se mostra melhorado em relação à utilização da mesma solução para todos os padrões, portanto, a seleção dinâmica é menos propensa a erros que a seleção estática (KO; SABOURIN; BRITTO JR, 2008).

Os métodos de seleção dinâmica são baseados no princípio da região de competência, acredita-se que um classificador ou ensemble (subconjunto de classificadores) possa ser mais adequado (ou competente) para classificar um padrão de teste em uma região. Esses métodos estão subdivididos em dois grupos, sendo um baseado em classificadores individuais e outro baseado em grupo de classificadores. A divisão dos grupos que se subdividem, está melhor exemplificada na Figura 17.

Figura 17 – Taxonomia proposta para o contexto de Múltiplos classificadores para Seleção Dinâmica



Fonte: (BRITTO; SABOURIN; OLIVEIRA, 2014)

Nas seções a seguir são descritos os métodos de seleção dinâmica mais tradicionais na literatura.

3.3.5.1 Overall Local Accuracy - OLA

A ideia básica é estimar a precisão $\delta_{i,j}$ de cada classificador em regiões locais do espaço de recurso ao redor de uma amostra de teste e, em seguida, usar a decisão do classificador mais preciso, estimando como a porcentagem de amostras de treinamento da região será corretamente classificada (WOODS; KEGELMEYER; BOWYER, 1997).

Dada a equação 3.6, o classificador de melhor estimativa da taxa de acerto local é então selecionado para fazer a decisão final x_j.

$$\delta_{i,j} = \frac{1}{K} \sum_{k=1}^{k} P(\omega_l \mid x_k \, \epsilon \, \omega_l, c_i) \tag{3.6}$$

3.3.5.2 Local Class Accuracy - LCA

Este método é semelhante ao método OLA, tendo como a única diferença a precisão da taxa de acerto local, que é realizada com relação às classes de saída ω_l (ω_l, que é a classe atribuída por x_j e por c_i), para toda a região de competência conforme a equação 3.7 (WOODS; KEGELMEYER; BOWYER, 1997). O classificador que apresentar maior relação é selecionado para prever a classe da amostra x_j (WOODS; KEGELMEYER; BOWYER, 1997).

$$\delta_{i,j} = \frac{\sum_{x_k \epsilon \omega_l} P(\omega_l \mid x_k \; c_i)}{\sum_{k=1}^{1} P(\omega_l \mid x_k \; c_i)} \tag{3.7}$$

3.3.5.3 Multiple Classifier Behaviour - MCB

Semelhante à técnica LCA, *Multiple Classifier Behaviour* (MCB), tem como diferença a saída de cada classificador e é ponderada pela distância entre a amostra de teste e cada padrão na região de competência, sendo baseada no espaço de conhecimento de comportamento e na precisão local do classificador (HUANG; SUEN, 1995). Dada uma nova amostra de teste x_k, sua região de competência θ_j, é estimada, em seguida, os perfis de saída da amostra de teste e os da região de competência, no entanto somente o classificador que atingiu o nível de competência mais alto é selecionado para prever o rótulo da amostra de teste x_j. A similaridade entre o perfil de saída da amostra de teste x_k e a sua região de competência, $P(\omega_l \mid x_k)$, é calculada por:

$$\delta_{i,j} = \frac{\sum_{x_k \epsilon \omega_l} P(\omega_l \mid x_k \;).W_n}{\sum_{k=1}^{K} \sum_{x_k \epsilon \omega_l} P(\omega_l \mid x_k).W_n} \tag{3.8}$$

As amostras com similaridades inferiores a um limiar pré-definido são removidas da região de competência ω_l. Assim, o tamanho da região de competência é variável, pois também depende do grau de similaridade entre a amostra de consulta e aqueles em sua região de competência. Após selecionar todas as amostras similares, a competência do classificador base $\delta_{i,j}$ é estimada por sua precisão de classificação na região resultante de competência (CRUZ; SABOURIN; CAVALCANTI, 2018).

3.3.5.4 A Priori

O método proposto por Didaci et al. (2005) consiste em calcular a precisão do classificador c_i em θ_j, que pode ser ponderada pelas distâncias entre as amostras de treinamento na região local e a amostra de teste. É similar ao OLA, pois não usa a informação da classe atribuída pelo classificador ao padrão de teste, considerando a classe dos k vizinhos mais próximos de x_k, a acurácia de cada classificador dada a amostra x_j. As amostras mais próximas têm uma influência maior no cálculo do nível de aptidão $\delta_{i,j}$ da amostra x_j. A equação 3.9, demonstra o cálculo do nível de competência de $\delta_{i,j}$, usando este método:

$$\delta_{i,j} = \frac{\sum_{k=1}^{K} P(\omega_l \mid x_k \, \epsilon \, \omega_l, c_i) W_k}{\sum_{k=1}^{K} W_k} \tag{3.9}$$

O classificador com maior valor de $\delta_{i,j}$ é escolhido, caso, o classificar selecionado não tenha nível de acurácia significativamente melhor que o dos outros, todos os classificadores no conjunto serão combinados usando regra de voto majoritário.

3.3.5.5 A Posteriori

É semelhante ao *A Priori*, tendo como diferença que a classe designada pelo classificador c_i é considerada, calculando a relação entre o somatório da probabilidade dos vizinhos, selecionando aquele com maior valor de $\delta_{i,j}$, através da equação:

$$\delta_{i,j} = \frac{\sum_{x_k \epsilon \omega_l} P(\omega_l \mid x_k, c_i) W_k}{\sum_{k=1}^{K} P(\omega_l \mid x_k, c_i) W_k} \tag{3.10}$$

O classificador selecionado é usado apenas para prever o rótulo de x_j se seu nível de competência for significativamente melhor que o dos outros conjuntos de classificadores, ou seja, quando a diferença no nível de competência for maior do que um nível pré-definido. Caso contrário, todos os classificadores no conjunto serão combinados usando a regra de votação majoritária (CRUZ; SABOURIN; CAVALCANTI, 2018).

3.3.5.6 K-Nearest-Oracles - KNORA

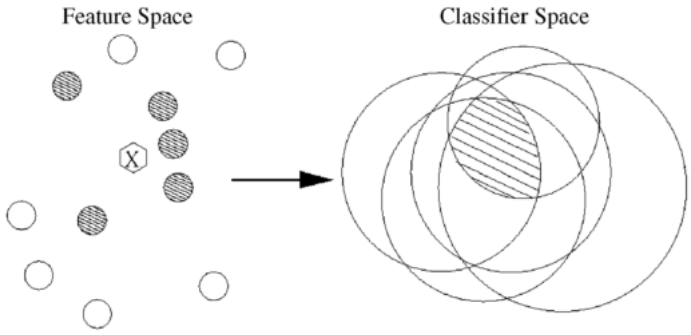
O método K-Nearest-Oracles (KNORA) apresentado por Ko, Sabourin e Britto Jr (2008), possui conceitos similares aos demais, OLA e LCA. Os métodos *A Priori* e *A Posteresi*, no que diz respeito à utilização de uma vizinhança dos padrões para cada instância de teste, podem especificar as estratégias de identificação do melhor subconjunto de classificadores, com aptidão para classificar corretamente uma dada amostra com base nas similaridades presentes em um conjunto de validação, empregando o conceito de "Oráculo", descobrindo assim quais classificadores atuam corretamente naquele conjunto

de testes, compondo uma mescla de classificadores mais adequados, o que aumenta as chances de sucesso (KUNCHEVA; RODRIGUEZ, 2007).

Para a aplicação do KNORA existem quatro diferentes variações do algoritmo: *KNORA-ELIMINATE, KNORA-UNION, KNORA-ELIMINATE-W e KNORA-UNIONW*. A seguir é detalhada cada variação:

KNORA-ELIMINATE dados K vizinhos x_j (com $1 \leq \mathrm{j} \leq \mathrm{K}$) de x_k, supondo um conjunto de classificadores C onde c_i é um classificador pertencente a este conjunto de classificadores corretos, é possível então supor que C' classifica corretamente todos os K padrões vizinhos da amostra de teste e que cada classe pertencente $\mathrm{c}_i \ \epsilon \ C'$ deve enviar uma votação no exemplo x_k. Conforme demonstrado na Figura 18. Caso nenhum classificador tenha desempenho satisfatório, todos os K-vizinhos mais próximos da amostra de teste, irão então diminuir o valor de K até que pelo menos um classificador classifique corretamente seus vizinhos (KO; SABOURIN; BRITTO JR, 2008).

Figura 18 – **KNORA-ELIMINATE**: utiliza apenas classificadores que rotulam corretamente todos os padrões K mais próximos. No lado esquerdo, o padrão de teste é apresentado como um hexágono, os pontos de dados de validação, como círculos e os 5 pontos de validação mais próximos estão escurecidos. No lado direito estão os classificadores usados, com a interseção de classificadores corretos, que estão escurecidos.

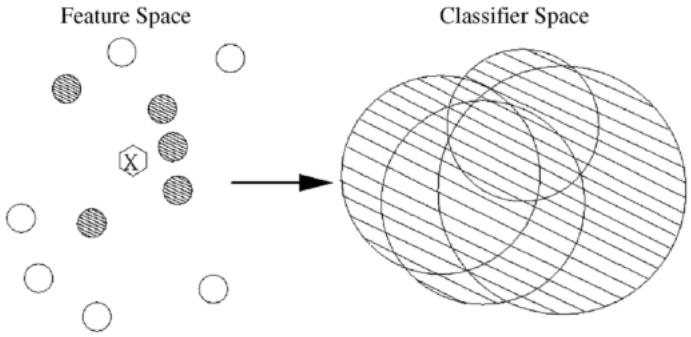


Fonte: (KO; SABOURIN; BRITTO JR, 2008)

KNORA-UNION dados K vizinhos x_j (com $1 \leq \mathrm{j} \leq \mathrm{K}$) de x_k, utiliza o conceito de "Oráculo" e também supõe um conjunto de classificadores igual ao KNORA-ELIMINATE, tendo a sua diferença na seleção do classificador que considera aquele que rotular corretamente ao menos em uma amostra da vizinhança. O classificador selecionado terá um peso na votação de acordo com o número de rotulações corretas (KO; SABOURIN; BRITTO JR, 2008). Portanto, supondo que o k-vizinho mais próximo tenha sido classificado corretamente por um conjunto de classificadores C(j), cada classificador $\mathrm{c}_i \ \epsilon \ C(j)$ pertencente a este conjunto submete-se a uma votação sobre a amostra x. Observe que, como todos os vizinhos K mais próximos são considerados, um classificador pode ter

mais de um voto se rotular corretamente mais de um vizinho. Quanto mais vizinhos um classificador rotular corretamente, mais votos este classificador terá para um padrão de teste (KO; SABOURIN; BRITTO JR, 2008). Demonstrado na Figura 19:

Figura 19 – **KNORA-UNION**: No lado esquerdo, o padrão de teste é apresentado como um hexágono, os pontos de dados de validação como círculos e os 5 pontos de validação mais próximos estão escurecidos. No lado direito, estão os classificadores usados, com a união dos classificadores corretos, que estão escurecidos.



Fonte: (KO; SABOURIN; BRITTO JR, 2008)

3.3.5.7 Dynamic Ensemble Selection Performance - DES-P

Proposta por Woloszynski et al. (2012), este método funciona da seguinte maneira: primeiro, selecionando todos os classificadores que atingiram um desempenho satisfatório, o desempenho local de um classificador base c_i é calculado, usando a região de competência θj, que é maior que o classificador aleatório (RC). A competência do classificador base é então calculada pela diferença entre a precisão do classificador base c_i, na região de competência θj (denotada por $\rho(c_i \mid \theta j)$, e o desempenho do classificador aleatório, ou seja, o modelo de classificação que escolhe aleatoriamente uma classe com probabilidades iguais. Para um problema de classificação com classes L, o desempenho do classificador aleatório é ($RC = \frac{1}{L}$), assim, o nível de competência $\delta_{i,j}$ nesta técnica é calculado de acordo com a equação 3.11.

$$\delta_{i,j} \;=\; \rho(c_i \mid \theta j) - \frac{1}{L} \tag{3.11}$$

Os classificadores de base com um valor positivo de $\delta_{i,j}$, isto é, que obtêm uma precisão local maior que o classificador aleatório, são selecionados para compor o conjunto c_i e se nenhum classificador base for selecionado, todo o conjunto será usado para classificação (CRUZ; SABOURIN; CAVALCANTI, 2018).

3.3.6 Avaliação de Resultados

.

Como a rotulação automática de gêneros e sentimentos musicais é uma problemática denominada de multi classes, e nas bases utilizadas essas classes estão desbalanceadas, será adotada para avaliar os resultados a medida de *F-Measure*, baseada nos valores de *Precision* e *Recall* que são obtidos através da matriz de confusão.

A matriz de confusão é uma tabela que auxilia na comparação de percentuais obtidos na classificação para cada uma das classes envolvidas. Regularmente utilizada para descrever os pontos críticos do desempenho de um sistema, esta tabela representa o total de amostras analisadas e o número de equívocos e acertos na classificação.

Na matriz de confusão os dados são organizados em uma tabulação cruzada entre a classe prevista pelo classificador e a classe real das amostras. Os valores constantes na diagonal principal indicam os acertos na classificação, enquanto os demais valores são referentes aos erros. Desde modo, um bom classificador é aquele que apresenta altos valores na diagonal principal e baixos valores nos demais elementos. Na Tabela 3 é demonstrado o exemplo de uma matriz de confusão de cão e gato:

Tabela 3 – Matriz de confusão

		Classe Predita	
		Cão	Gato
Classe Verdadeira	Cão	35	15
	Gato	10	40

Com base nas informações da matriz de confusão são calculadas as métricas utilizadas para avaliação dos classificadores:

- Acurácia (*accuracy*): é a medida geral para percentual de classificações corretas. É calculada tomando simplesmente a razão de amostras corretamente classificadas pelo número de amostras disponíveis (WITTEN et al., 2005).
- Revocação (*recall*): é a medida para quantos exemplos, de todos os exemplos existentes de uma classe, são efetivamente classificados na classe correta (WITTEN et al., 2005).
- Precisão (*precision*): é a proporção de exemplos que são verdadeiramente de uma classe sobre o total de exemplos classificados como aquela classe (WITTEN et al., 2005).
- Medida F (*F-Measure*): A medida *F-measure* (MP-F) é definida como a média ponderada harmônica dos valores de precisão (precision) e de revocação (*recall*) (WITTEN et al., 2005), ou seja:

$$F - measure = 2 \times Precision \times Recall/(Precision + Recall) \tag{3.12}$$

No processamento da validação cruzada os dados são fragmentados em dois subconjuntos, denominados base de treinamento e base de teste. Na primeira etapa um algoritmo de classificação é aplicado à base de treinamento. Com isso é obtido um modelo treinado, que representa o conhecimento extraído. Numa segunda etapa o modelo obtido é aplicado ao fragmento da base de testes. Como a base de testes também é previamente rotulada é possível medir a taxa de acerto do modelo, comparando o resultado obtido com a rotulação disponível na base de testes (HIRJI, 1999).

A técnica consiste em dividir a base de dados em n partes (*folds*) e é chamada de Validação Cruzada. Assim, n-1 partes são utilizadas para o treinamento e n é usada como base de testes (WOODS; KEGELMEYER; BOWYER, 1997). O processo é repetido n vezes, de forma que cada parte seja usada uma vez no conjunto de testes. Ao final, a correção total é calculada pela média dos resultados de cada etapa, obtendo-se assim uma estimativa da qualidade do modelo gerado e permitindo análises estatísticas (WITTEN; FRANK; HALL, 2011).

4 Metodologia

Este capítulo apresenta a metodologia empregada para a realização dos experimentos e está organizado da seguinte forma: bases de dados utilizadas; extração de características para as diferentes representações; combinação e geração do conjunto de classificadores; as técnicas e estratégias utilizadas para a seleção dinâmica, apresentando os detalhes de implementação.

4.1 Bases de Músicas Utilizadas

Nesta seção, são apresentadas as bases de dados que vão ser utilizadas para a aplicação do método proposto. A partir das músicas existentes nessas bases, são extraídas as características no domínio de áudio, do visual e das letras das músicas para futura classificação.

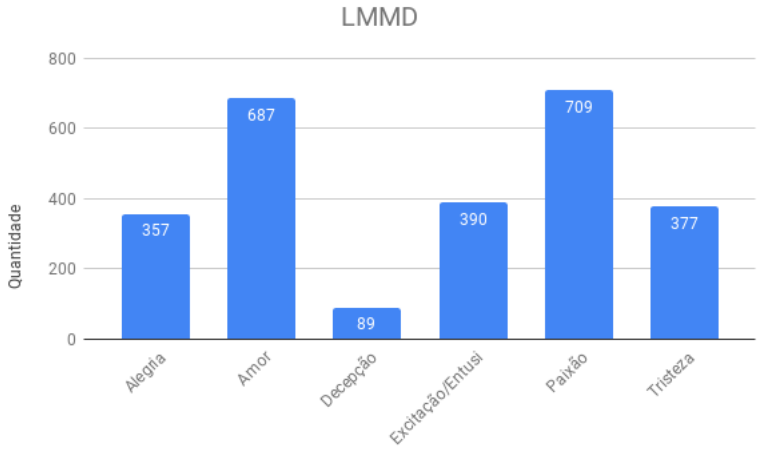
4.1.1 LMD

Criado por Silla Jr., Kaestner e Koerich (2008b), o LMD é uma base composta por 3.227 músicas subdivididas em 10 diferentes gêneros musicais latino-americanos: Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertanejo e Tango. É construída para ser utilizada em experimentos de classificação de gênero, tendo como critério de rotulação a percepção humana que leva em consideração a forma como elas são dançadas. Para a realização dos testes com gênero, o presente trabalho fez uso do *artist filter* com separação de amostra em *folds*.

Para não haver artistas repetidos, seguindo o definido pelo *artist filter*, na realização do treinamento não é possível aproveitar todas as canções, portanto são selecionadas 900 músicas, divididas em 3 *folds*, e igualmente entre as classes, contendo 30 músicas cada uma.

Para a emoção, é utilizada uma base derivada da LMD, criada por Santos e Silla (2015), a *Latin Music Mood Database* (LMMD), a qual atribui rótulos de sentimentos, seguindo o plano emocional fornecido pelo Modelo de Circunferência Afetiva de Watson e Tellegen, demonstrado no Capítulo 3.2, além das representações de áudio e letras contendo 2.609 canções rotuladas por especialista em alegria, amor, decepção, entusiasmo, paixão e tristeza, seguindo a distribuição demonstrada na Figura 20.

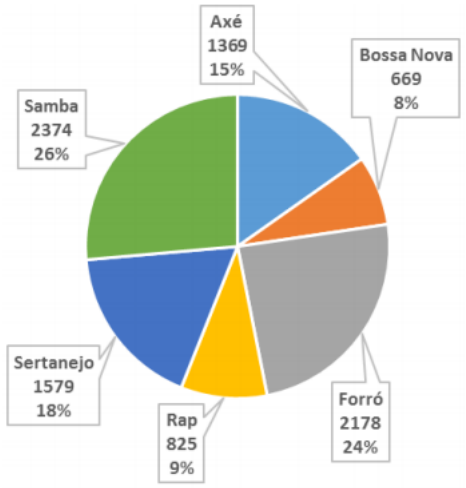
Figura 20 – Distribuição da LMMD



4.1.2 BRMD

A *Brazilian Music Database* (BRMD) é uma base derivada da GNMID, criada por Summers et al. (2016), contendo músicas de diferentes nacionalidades como Brasil, EUA, México e Colômbia, em que cada áudio contém as informações de país, data, track ID, artista, sendo 94% desta base classificada com o rótulo de humor. Posteriormente, Pereira e Silla (2017) criam uma base específica de gêneros típicos brasileiros como Axé, Bossa Nova, Forró, Rap, Sertanejo e Samba, denominada BRMD (PEREIRA; SILLA, 2017).

Figura 21 – Distribuição da BRMD por Gênero

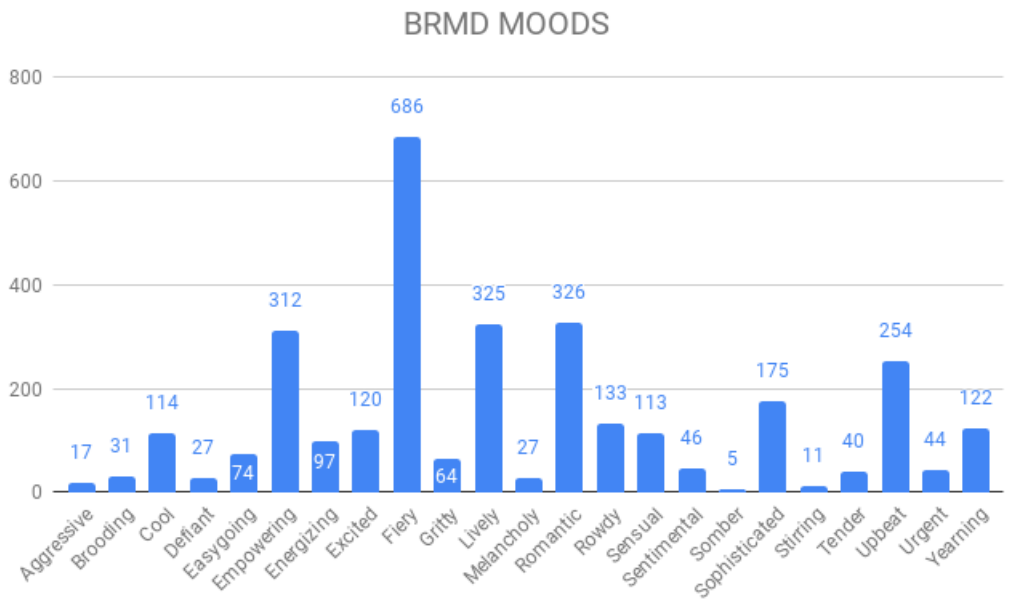


Fonte: (PEREIRA; SILLA, 2017)

Neste trabalho, também é utilizada a rotulação de sentimentos, com 3.169 canções

rotuladas, que seguem o modelo de Thayers 2D apresentado no Capítulo 33.2, divididas em 25 tipos de emoções e distribuídas conforme a Figura 22. A representação da BRMD em emoção será denominada neste trabalho como BRMD-MOOD quando necessária uma distinção das bases.

Figura 22 – Distribuição da BRMD por Emoção

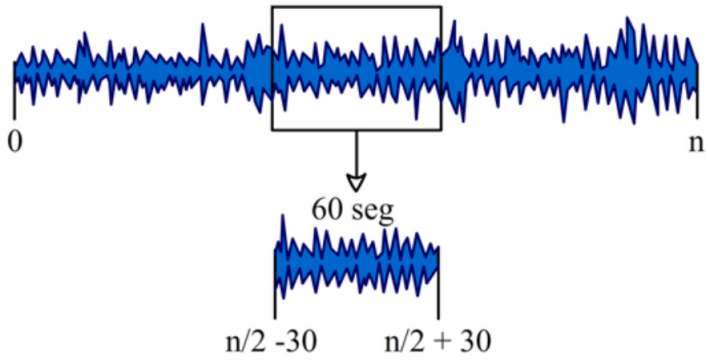


4.2 Segmentação do Áudio

Na segmentação, são estabelecidos critérios para o particionamento dos dados, tornando, assim, o volume de informações a ser trabalhado menor e mais adequado às necessidades, diminuindo, consequentemente, a quantidade de processamento. Neste trabalho, utilizamos a abordagem apresentada por George e Shamir (2014), na qual são utilizados segmentos únicos de 60s em vez de três segmentos de regiões distintas do sinal, conforme utilizado por Silla Jr., Kaestner e Koerich (2010) e Costa et al. (2013).

Na segmentação realizada, foram extraídos trechos de 60 segundos, selecionados conforme demonstrado na Figura 23. No caso de amostras com tempo inferior ao definido, fora utilizado todo o sinal (COSTA; OLIVEIRA; SILLA JR, 2017). Posteriormente, os áudios foram divididos em 3 pastas, contendo 30 músicas de tamanho igual. A divisão foi feita usando *artist filter* (FLEXER, 2007), que coloca as partes musicais de um artista específico exclusivamente em um único conjunto de áudios.

Figura 23 – Segmentação do sinal do áudio



Fonte: (COSTA et al., 2013)

4.3 Extração de Características

A etapa de extração de características se fundamenta no processo de representação de uma amostra através das características que a identificam e que sejam suficientes para alimentar o processo de aprendizagem do classificador, com posterior atribuição de um resultado final. Para o reconhecimento de emoção e gêneros musicais, se mostra possível a extração dessas características diretamente do sinal, ou seja, as caraterísticas acústicas, através de suas representações visuais na forma de espectrogramas, que são características no domínio visual. Também é possível, a utilização das letras das músicas, criando assim representações distintas para o treinamento dos classificadores.

4.3.1 Representação do Áudio

Frequentemente utilizado para processamento de representações acústicas, o *Statistical Spectrum Descriptor* (SSD) foi originalmente proposto por Lidy e Rauber (2005), oferecendo características distintas para os descritores de timbre, variação de frequência, melodia e harmonia, e ritmo, sendo estas as três principais características de baixo nível utilizadas em trabalhos de reconhecimento automático de gêneros musicais (TZANETAKIS; COOK, 2002). A escolha por SSD se dá com base nos resultados obtidos por Silla Jr., Kaestner e Koerich (2010).

Complementar ao SSD, outra representação do timbre, é o MFCCs, que é a distribuição temporal do espectro de sinais de áudio, sendo responsável em grande parte pela percepção do timbre, também extraída pela biblioteca LibROSA[1], um pacote *Python* para análise de música e áudio.

[1] https://librosa.github.io/librosa/index.html

Outras representações, como *Rhythm Pattern* (RP) e *Rhythm Histogram* (RH), são utilizadas extraindo as características do ritmo através da biblioteca em Python desenvolvida pela *Vienna University of Technology* (LIDY; PÖLZLBAUER; RAUBER, 2005).

4.3.2 Representação Visual

Além disso, para esse trabalho, com o propósito de gerar diversos classificadores de representações distintas para compor o *pool* de classificadores, foram extraídas características visuais gerando espectrogramas utilizando *Sound eXchange* (SoX) 14.4.2 (BAGWELL; KLAUER, 2010). O eixo vertical corresponde à frequência e à intensidade do brilho de cada pixel da imagem, de modo a representar a amplitude do sinal, ficando representada a variação do sinal no decorrer do tempo.

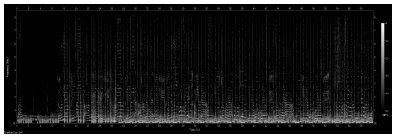


Figura 24 – Axe -50 dBFS

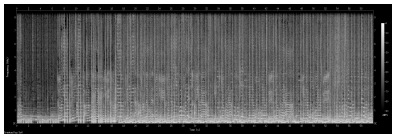


Figura 25 – Axe -90 dBFS

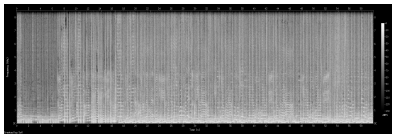


Figura 26 – Axe -130 dBFS

Os espectrogramas foram gerados variando o limite inferior definido em -50, -70, -90, -110 e -130 dBFS (o limite superior é sempre 0 dBFS). Isso define a quantidade de informações que está presente no espectrograma, ou seja, o seu contraste. A diminuição de informações, ou do limite inferior, há um aumento efetivo do contraste do espectrograma e vice-versa. As Figuras 24, 25 e 26 demonstram a variação obtida nas imagens geradas com limites inferiores distintos para uma música do gênero axé já em escala de cinza, se adequando da melhor maneira aos processos de extração de características.

As características são obtidas com utilização do $\text{LBP}_{8.2}$ e extração global de características, que, conforme observado por Costa, Oliveira e Silla Jr (2017), são os que obtiveram melhores resultados na classificação de gênero, utilizando, portanto, os mesmos princípios para a representação dos rótulos utilizados neste trabalho.

4.3.3 Representação das Letras

Para a realização dos experimentos utilizando a representação das letras das músicas, é necessário criar um conjunto de palavras usadas nas músicas considerando a sua frequência e calculando sua relevância na música através do cálculo do TF/IDF. Desta forma, é atribuído um valor maior para as palavras mais importantes, limitando a *bag* a um conjunto de 1500 palavras. Em Alemayehu e Willett (2003), é demonstrado que essa técnica TF/IDF, em conjunto com a *Stemming*, obtém bons resultados, além de ser um procedimento muito rápido. Nesta mesma abordagem, além da utilização TF/IDF com

Stemming, utilizam-se *n-grams* extraídos dos *bigrams* (para n=2), *trigrams* (para n=3) e *quadgrams* (para n=4) de todas as letras de músicas das bases da LMMD e BRMD. Não foi possível realizar o experimento com a LMD para gênero devido ao problema na base de dados de letras.

4.3.3.1 Pré Processamento das Letras

Antes da criação do conjunto de palavras que representa uma música, é necessário realizar um pré-processamento e remoção de *StopWords*. Nesta etapa são removidos caracteres especiais, acentos, quebras de palavras com hifens, números e padronização para minúsculo.

4.3.3.2 Remoção de StopWords

As *StopWords* são palavras de preposições ou artigos, que carregam significado semântico irrelevante ao texto. Estas palavras são normalmente presentes em uma lista definida por idiomas (EL-KHAIR, 2006). Neste trabalho, foram utilizadas duas lista de *StopWords* em português e espanhol.

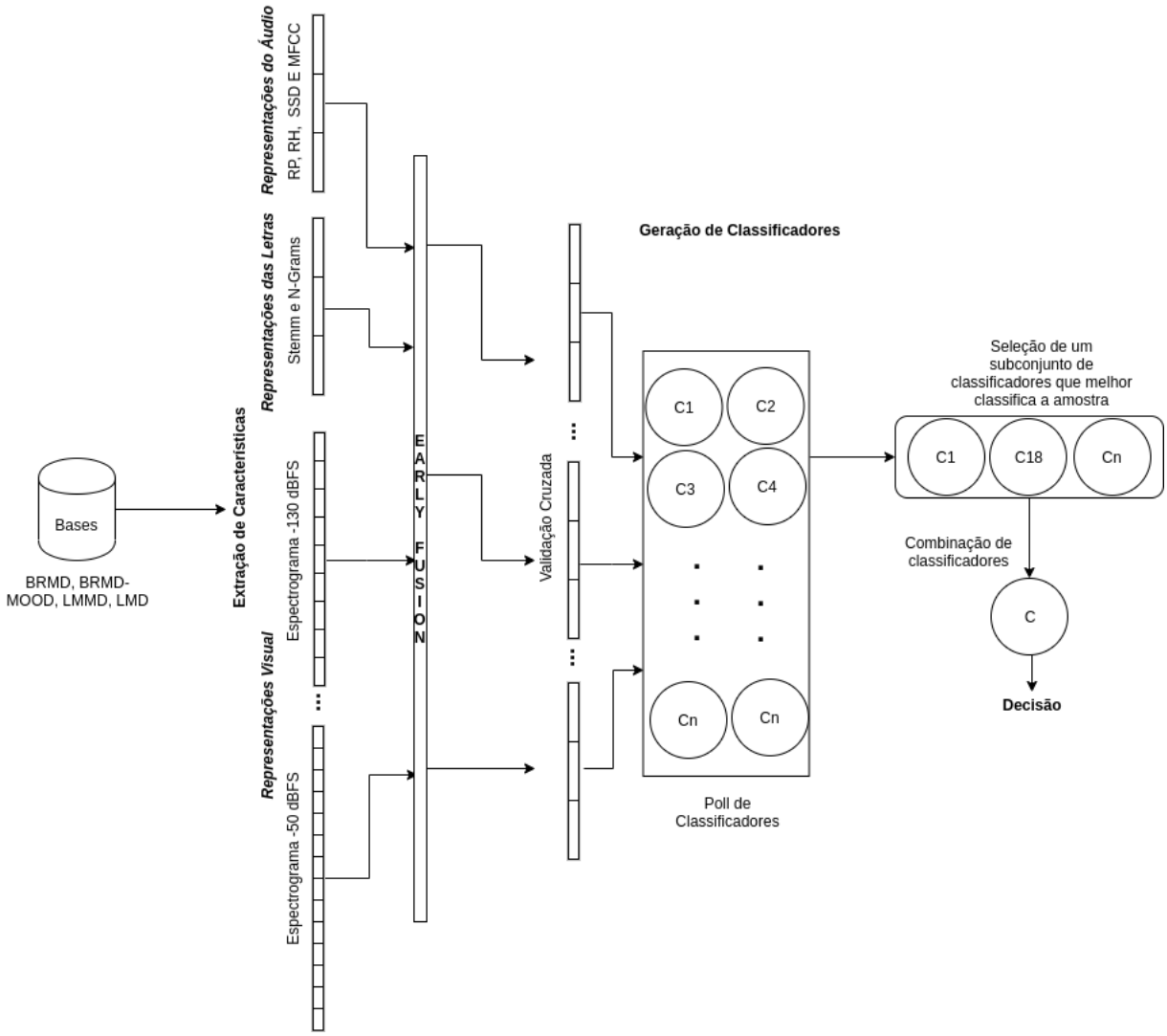
4.4 Seleção de Dinâmica de Conjunto de Classificadores

Os sistemas de múltiplos classificadores foram uma alternativa encontrada para a melhora do desempenho de sistemas monolítico (KUNCHEVA; WHITAKER, 2003) (KITTLER et al., 1998) (Jain; Duin; Jianchang Mao, 2000). A ideia consiste em adotar diversos classificadores para atenuar a variância observada em sistemas individuais de classificação. A estratégia apresentada nesta sessão visa construir um conjunto heterogêneo de classificadores com base nos algoritmos de classificação utilizados nos experimentos da classificação monolítica, de forma que o conjunto obtido possa superar as dificuldades em se adotar um classificador individual.

Além da geração do pool de classificadores, o processo de rotulação da amostra de teste na estratégia de seleção dinâmica é baseado nas características da amostra. Para tanto, são usadas informações sobre a complexidade dos dados e acurácias sobre o conjunto de instâncias presentes na vizinhança da nova amostra. A Figura 27 apresenta uma visão geral do SMC proposto.

Para o processo de geração dos classificadores que compõem o *pool*, foi adotado o conceito de validação cruzada, o qual consiste na metodologia de confrontamento de conjuntos de testes alternados contra os demais conjuntos utilizados para o treinamento. Os motivos para adotar essa técnica se dão pela escassez de dados, caracterizando uma situação em que é priorizado reservar uma maior quantidade de dados para construção de um bom modelo e que os dados restantes podem se mostrar insuficientes para um

Figura 27 – Visão Geral do Método Proposto com SMC.



conjunto de testes representativos. Na validação cruzada, o conjunto de dados é dividido em conjuntos denominados *folds*, nos quais um dos conjuntos é utilizado para teste e os demais para a construção do modelo. O processo é repetido de forma que todos os *folds* tenham sido utilizados como conjunto de teste e treino. A utilização desse conceito permite obter uma melhor estimativa de probabilidades menos sensível a variabilidade dos conjuntos.

A utilização de diferentes classificadores oferece a possibilidade de que uma determinada amostra no sistema de classificação seja interpretada de maneiras distintas em uma mesma região, criando-se então um *pool* heterogênico, de acordo com o comportamento de cada um dos classificadores envolvidos. A representação das características de uma amostra, quando submetida à classificadores diferentes, produz resultados diferentes, com base no desempenho individual dos classificadores (BRITTO; SABOURIN; OLIVEIRA,

2014).

Com a utilização de diferentes classificadores, oferece a possibilidade de que uma determinada amostra no sistema de classificação seja interpretada de maneiras distintas em uma mesma região, criando-se então um *pool* heterogênico, de acordo com o comportamento de cada um dos classificadores envolvidos. A representação das características de uma amostra quando submetida à classificadores diferentes produz resultados diferentes, com base no desempenho individual do classificadores (BRITTO; SABOURIN; OLIVEIRA, 2014).

Essa variedade de interpretações baseia-se no conceito de complementaridade dos resultados, segundo o qual dois classificadores são ditos complementares quando apresentam erros diferentes para um mesmo conjunto de teste. Esta diferença de interpretação é importante para a etapa de seleção de um subconjunto que será utilizado para definir a classe de uma amostra de teste, possibilitando um melhor desempenho.

A base de dados LMD é utilizada com 3 *folds*, sendo que cada *fold* contém 300 amostras, com 30 representações de cada gênero utilizando *artist filter*. Para a validação da seleção dinâmica, além dos *folds* de treinamento, um outro conjunto de amostras para validação é necessário. Para tanto, é utilizado um conjunto de amostras contendo 400 canções, com 40 representações de cada gênero.

Para a base de dados BRMD, também foram utilizadas 3 *folds*, com 1801 amostras cada e 274 canções representando cada gênero. Para validação da seleção dinâmica de conjunto de classificadores, é utilizado um 4° *fold*. As bases LMMD e BRMD-MOOD são utilizadas com 5 *folds*, separadas de forma a ter a mesma quantidade de amostras para cada tipo de emoção em cada *fold*, reservando um *fold* específico para validação, não sendo utilizada a técnica de *artist filter*.

4.5 Considerações

Neste capítulo, foi apresentado o sistema de múltiplos classificadores com seleção dinâmica de subconjunto proposto, que se baseia em medidas de complexidade e acurácia para efetuar o processo de geração dos classificadores e determinar o processo de classificação de uma nova instância. Detalhou-se como foram utilizadas e divididas cada base de dados das etapas envolvidas no processo de classificação. A descrição dos experimentos e os resultados obtidos pelas metodologias propostas podem ser encontrados no próximo Capítulo 5.

5 Resultados

Este capítulo vem apresentar e debater os resultados obtidos com os métodos propostos no Capítulo 4. Através da classificação monolítica, primeiramente fora avaliada cada representação e a contribuição desta representação por categoria. Em sequência foram avaliadas as representações combinadas através de um método de fusão das características extraídas de diferentes vetores, o método *early fusion*, mas integradas em um único vetor para treinamento (NIAZ; MERIALDO, 2013), com a utilização de classificação monolítica e da seleção dinâmica de classificadores, sendo realizadas as considerações e comparação de resultados ao final de cada seção. Quanto à realização dos experimentos, estes foram divididos de forma a criar uma base de comparação. Os primeiros experimentos a utilizarem algoritmos de classificação monolítica foram realizados em cada representação de forma individual, em seguida as representações foram combinadas, sendo realizados os experimentos com os mesmos classificadores e com múltiplos classificadores, utilizando o método de seleção dinâmica de classificadores. Por fim, todas as representações foram combinadas, criando um único vetor com áudio, imagem e letra, que avaliou o desempenho de todos os classificadores.

Na seção 5.1, primeiramente serão apresentados os resultados obtidos com a utilização da base de dados de sentimentos LMMD, que contém 3.609 canções rotuladas em 6 diferentes tipos de emoções. Já na seção 5.2 serão apresentados valores das taxas de reconhecimento obtidos para a base de dados BRMD-MOOD, que contêm 25 rótulos de emoção e 3.169 canções. A seção 5.3 apresentará os resultados dos experimentos com gêneros da base de dados LMD, que contêm 900 canções divididas em 3 *folds*. Por fim, a seção 5.4 apresentará os resultados experimentais para a base de dados BRMD, rotulada em 6 categorias de gênero.

5.1 LMMD

5.1.1 Classificação Utilizando Representações do Áudio

O desempenho dos classificadores com cada representação do áudio estão apresentados na Tabela 4, onde cada célula corresponde a média ponderada *f-measure* e o desvio padrão individual dos classificadores. As precisões obtidas utilizando classificadores baseados nas características do áudio são consideradas satisfatórias se comparadas com as demais. É importante observar que a precisão dos descritores SSD, RP e MFCC são melhores se classificados pelo algoritmo SVM, com média ponderada de 33,4%, 38,4% e 35,0% respectivamente. O descritor RH foi melhor classificado pelo algoritmo KNN,

ficando com acurácia de 31,4%.

Tabela 4 – Resultado dos classificadores por cada representação do áudio na LMMD. Em negrito o melhor classificador para cada representação.

Classificadores	**SSD**	**RP**	**RH**	**MFCC**
Decision Tree	27,4 ± 0,018	26,3 ± 0,009	25,1 ± 0,017	28,1 ± 0,020
k-NN	29,8 ± 0,017	34,4 ± 0,023	**31,4 ± 0,020**	31,3 ± 0,018
Gaussian NB	25,3± 0,034	24,9 ± 0,017	19,0 ± 0,005	29,3 ± 0,023
SVM	**33,4 ± 0,029**	**38,4 ± 0,022**	30,9 ± 0,022	**35,6 ± 0,035**
Logistic Regression	31,6 ± 0,025	31,7 ± 0,012	30,7 ± 0,028	30,1 ± 0,017

5.1.1.1 Contribuição

A partir da análise de todos os descritores, é possível observar na Tabela 5 que os piores resultados obtidos são na categoria "decepção", sendo este um reflexo da baixa quantidade de representações que esta categoria possui. Os descritores MFCC e RH não parecem contribuir para a classificação da categoria "alegria", enquanto o descritor RP é o que melhor contribui para a classificação das amostras, com a precisão mais alta para "paixão" e "excitação/entusiasmo", de 44,6% e 51,0% respectivamente. Todos os resultados demonstrados são uma média ponderada *f-measure* dos melhores classificadores por representação.

Tabela 5 – Resultado das representações do áudio por categoria para LMMD.

	SSD	**RP**	**RH**	**MFCC**
Alegria	26,0%	29,4%	24,4%	22,2%
Paixão	38,2%	44,6%	**39,4%**	41,4%
Decepção	16,0%	6,0%	14,4%	6,6%
Excitado/Entusiasmo	**43,6%**	**51,0%**	32,8%	**48,4%**
Amor	33,4%	38,4%	32,2%	35,0%
Tristeza	23,6%	22,8%	23,0%	27,8%

5.1.2 Combinando as Representações do Áudio

Aqui são apresentados os resultados obtidos através da utilização da técnica de *early fusion* para as representações de áudio. Na Tabela 6 é demostrado o desempenho de cada algoritmo de classificação, sendo que cada célula representa a média ponderada *f-measure* e o desvio padrão de desempenho do classificador.

Neste caso, o algoritmo SVM obteve o melhor desempenho, com 54,94%, resultado superior ao apresentado pelas representações individuais. No geral, ao combinar as representações esse algoritmo obteve uma leve melhora na precisão dos classificadores, sendo possível considerar que as representações do áudio, quando agregadas, melhoram a precisão média dos classificadores.

Tabela 6 – Resultado das representações de áudio combinadas com o método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	35,61 ± 0,008
k-NN	46,05 ± 0,012
Gaussian NB	26,77 ± 0,018
SVM	**54,94 ± 0,011**
Logistic Regression	39,29 ± 0,012

5.1.3 Seleção Dinâmica de Classificadores com as Representações do Áudio

Os resultados aqui demonstrados foram obtidos pela utilização do método de fusão *early fusion*, com características das representações do áudio, aplicando a seleção dinâmica de classificadores. Os algoritmos k-NN, *Decision trees*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, que compõem o conjunto de classificadores heterogêneos, com 30 classificadores, sendo 5 de cada algoritmo, foram previamente treinados com validação cruzada de 5 vezes, utilizando 70% da base de dados. O restante da base foi reservada para aplicar os algoritmos de seleção dinâmica de conjunto de classificadores para a realização da rotulação da amostra de teste.

Os resultados expostos na Tabela 7 apresentam os desempenhos dos algoritmos de seleção dinâmica com utilização da LMMD com os rótulos de sentimento, obtendo um oráculo de 89,39%. Os algoritmos KNORA-U e DES-P obtiveram os melhores resultados, com precisão de 66% e 65% respectivamente. Para o KNORA-U, os resultados obtidos foram melhores quando analisados de forma individual, por exemplo, a categoria "amor" e "excitação/entusiasmo" tiveram o desempenho igual ao DES-P, mas nas outras categorias mesmo com uma amostragem menor, o KNORA-U obteve um melhor resultado devido a construção da vizinhança, que é realizada pelo cálculo da distância euclidiana entre os vetores e, quando não encontrado um classificador base, realiza a união dos classificadores que obtiveram algum resultado para a amostra de teste, apresentando, desta forma, uma acurácia maior.

Tabela 7 – Resultado da seleção dinâmica utilizando as representações do áudio da LMMD. Em negrito fora destacado o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Alegria	***55,0%***	29,0%	***50,0%***	32,0%	33,0%	30,0%	8,0%
Amor	**72,0%**	61,0%	**72,0%**	54,0%	55,0%	51,0%	38,0%
Decepção	**73,0%**	43,0%	**73,0%**	25,0%	60,0%	40,0%	17,0%
Excitado/Entusiasmo	65,0%	**67,0%**	65,0%	50,0%	61,0%	47,0%	32,0%
Paixão	**67,0%**	52,0%	**67,0%**	47,0%	47,0%	44,0%	38,0%
Tristeza	**69,0%**	67,0%	62,0%	55,0%	57,0%	50,0%	39,0%
F-Measure	**66%**	**54%**	**66%**	**47%**	**56%**	**45%**	**32%**
Acurácia	**66,66%**	**54,54%**	**65,90%**	**47,72%**	**56,81%**	**51,51%**	**37,12%**

5.1.4 Classificação Utilizando Representações Visuais

Os classificadores, para a realização dos experimentos, foram otimizados para cada um dos limites de amplitudes do LBP, que são -50, -70, -90, -110 e -130 dBFS, extraídos utilizando $LBP_{8,2}$ e a extração global de características. O resultado alcançado para cada algoritmo de classificação é demonstrado na Tabela 8, onde cada célula representa a média ponderada do algoritmo e o seu desvio padrão.

O algoritmo *Logistic Regression*, com parâmetros de tolerância de parada e penalização ajustados, obteve os melhores resultados na maioria dos testes para as amplitudes -50, -70 e -110 dBFS, sendo superado pelo algoritmo SVM somente com amplitude de -130 dBFS.

Tabela 8 – Resultado dos classificadores por cada representação visual na LMMD. Em negrito, consta o melhor classificador para cada representação.

Classificadores	**LBP-50**	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Decision Tree	27,2 ± 0,013	28,4 ± 0,010	30,1 ± 0,014	26,4 ± 0,017	27,8 ± 0,017
k-NN	30,9 ± 0,013	30,1 ± 0,021	31,1 ± 0,016	31,7 ± 0,025	32,9 ± 0,012
Gaussian NB	19,6 ± 0,020	27,3 ± 0,019	29,0 ± 0,014	25,6 ± 0,010	25,9 ± 0,016
SVM	32,2 ± 0,027	33,5 ± 0,026	34,4 ± 0,023	33,1 ± 0,03848	**35,9 ± 0,019**
Logistic Regression	**33,6 ± 0,016**	**37,0 ± 0,028**	**37,0 ± 0,025**	**34,9 ± 0,004**	34,3 ± 0,007

5.1.4.1 Contribuição

Avaliando a contribuição das representações da imagem para cada categoria da base de dados LMMD, executando a validação cruzada 5 vezes e com a utilização do melhor classificador de cada descritor da imagem, temos os dados contidos na Tabela 9. Os resultados apresentados na tabela correspondem à média ponderada *f-measure* do classificador por categoria.

Tabela 9 – Resultado das representações da imagem por categoria para LMMD.

	LBP-50	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Alegria	18,4%	24,4%	21,6%	19,8%	30,2%
Paixão	39,4%	41,0%	48,2%	42,0%	39,0%
Decepção	0,0%	5,0%	9,2%	8,0%	12,6%
Excitação/Entusiasmo	**49,0%**	**55,8%**	**53,0%**	**52,6%**	**46,4%**
Amor	33,4%	35,8%	30,4%	31,0%	32,4%
Tristeza	14,6%	15,4%	12,4%	16,6%	23,6%

Para a categoria “decepção”, as diferentes representações resultaram em pouca contribuição para a classificação, obtendo um resultado bem inferior, muito em razão da baixa presença de amostras para treinamento na base de dados. As categorias “alegria” e “tristeza” também obtiveram resultados inferiores com LBP de baixa amplitude, alcançando os piores resultados com -50 dBFS, obtendo os resultados de 18,4% e 14,6% respectivamente.

No geral as categorias apresentaram melhores resultados com as amplitudes médias, ou seja, as da faixa de -70 a -110 dBFS.

5.1.5 Combinando as Representações Visuais

Com as representações visuais combinadas através de método de *early fusion*, foi possível observar que no desempenho dos algoritmos não houve diferença significativa nos resultados. Estes resultados estão disponibilizados na Tabela 10, onde cada célula representa a média ponderada. O melhor resultado apresentado é de 35,84% para o algoritmo *Logistic Regression*, inferior ao alcançado com as amplitudes -70 dBFS e -90 dBFS de forma individual com 59 atributos, portanto ao criar um vetor maior os algoritmos não obtiveram uma melhora no resultado, como o que fora observado com o áudio.

Tabela 10 – Resultado das representações das imagens da LMMD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	28,24 ± 0,018
k-NN	33,04 ± 0,022
Gaussian NB	28,80 ± 0,014
SVM	34,78 ± 0,022
Logistic Regression	**35,84 ± 0,035**

5.1.6 Seleção Dinâmica de Classificadores com as Representações Visuais

Os experimentos foram realizados a partir das características visuais extraídas conforme explicado na sessão 4.3.2, e combinadas através do método *early fusion*. O conjunto de classificadores é formado pelos algoritmos k-NN, *Decision trees*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, que foram treinados com 70 da base de dados, contendo 5 representações de cada um destes algoritmos. Este treinamento dos algoritmos ocorreu através da validação cruzada de 5 vezes. O restante da base de dados fora utilizada para aplicação dos algoritmos de seleção dinâmica nos conjuntos de teste, obtendo os resultados demonstrados na Tabela 11, com oráculo de 97,72% para esse conjunto inicial de classificadores.

É possível observar que o algoritmo KNORA-U obteve os melhores resultados com acurácia de 39,75% e *f-measure* de 37%, com o algoritmo DES-P alcançando um resultado semelhante, com acurácia de 39,25% e *f-measure* de 37%, sendo que, se comparados os dois algoritmos, somente o sentimento "tristeza" obteve desempenho superior com DES-P. Analisando os classificadores por categoria, "decepção" não é classificada pelos algoritmos OLA e MCB, além de ter um baixo desempenho com os demais classificadores. Enquanto a categoria "excitação/entusiasmo" é a que apresenta ter melhores padrões em suas imagens,

Tabela 11 – Resultado da seleção dinâmica utilizando as representações das imagens da LMMD. Em negrito o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Alegria	18,0%	**23,0%**	16,0%	16,0%	15,0%	15,0%	10,0%
Amor	**47,0%**	41,0%	45,0%	36,0%	30,0%	40,0%	36,0%
Decepção	18,0%	19,0%	14,0%	0,0%	0,0%	13,0%	**23,0%**
Excitação/Entusiasmo	**61,0%**	49,0%	60,0%	53,0%	49,0%	50,0%	54,0%
Paixão	**38,0%**	27,0%	**38,0%**	32,0%	33,0%	31,0%	34,0%
Tristeza	22,0%	25,0%	25,0%	26,0%	19,0%	**34,0%**	20,0%
F-Measure	**37%**	**33%**	**37%**	**32%**	**29%**	**34%**	**31%**
Acurácia	**39,75%**	**33,25%**	**39,25%**	**32,00%**	**31,50%**	**33,75%**	**35,00%**

com um bom desempenho entre os algoritmos de seleção de classificadores, tendo o melhor resultado com KNORA-U, de 61,0% de *f-measure*.

5.1.7 Classificação Utilizando Representações das Letras

Nos experimentos são utilizadas as letras das músicas da base de dados LMMD, com remoção de *stopwords* em português e espanhol. Os resultados apresentados na tabela abaixo estão organizados de forma a demonstrar o desempenho alcançado por cada algoritmo de classificação, após a validação cruzada de 5 vezes, onde cada célula representa a média ponderada da *f-measure* e o desvio padrão, devido à desbalanceada das categorias da base de dados.

Na Tabela 12 a primeira coluna demonstra os resultados dos experimentos utilizando a técnica de *stemming*, nela podemos observar que o melhor desempenho é do algoritmo SVM, com 44,6%. Já o algoritmo k-NN alcançou o pior desempenho com a mesma técnica, resultando em 27,90% de assertividade, quase 20% a menos que o melhor classificador.

Tabela 12 – Resultado dos classificadores por cada representação das letras com LMMD. Em negrito o melhor classificador para cada representação.

Classificadores	**Stemm**	**2-gram**	**3-gram**	**4-gram**
Decision Tree	36,6 ± 0,015	32,7 ± 0,024	32,8 ± 0,039	33,8 ± 0,037
k-NN	27,8 ± 0,022	24,5 ± 0,032	24,7 ± 0,026	24,2 ± 0,029
Gaussian NB	31,6 ± 0,015	35,4 ± 0,021	34,9 ± 0,020	35,0 ± 0,020
SVM	**44,6 ± 0,035**	28,0 ± 0,019	29,0 ± 0,000	26,0 ± 0,027
Logistic Regression	41,6 ± 0,033	**41,5 ± 0,035**	**41,4 ± 0,04125**	**41,8 ± 0,040**

Utilizando palavras com dimensionalidades reduzidas, com a técnica de *n-grams*, os resultados apresentaram uma pequena diferença quando aumentado o número do valor de N. Neste aspecto, o algoritmo *Decision trees* obteve o melhor desempenho com a representação de 4-grams e precisão de 33,80%, enquanto que os algoritmos K-NN e SVM apresentaram dificuldades em precisar a emoção utilizando a mesma técnica. O algoritmo *Logistic Regression* trouxe o melhor resultado com todas as representações, obtendo pouca variação de desempenho entre os *n-grams*, alcançando sua melhor assertividade com 4-grams, de 41,80%.

5.1.7.1 Contribuição

Esta seção avalia a contribuição da representação da letra por categorias de emoção da base de dados LMMD. Os resultados apresentados na Tabela 13 foram obtidos selecionando o melhor classificador geral e executando a validação cruzada 5 vezes. Na tabela cada célula representa a média ponderada *f-measure*.

O primeiro experimento utilizou a técnica de redução da dimensionalidade de uma palavra *stemming*, sendo representado na primeira coluna da Tabela 13, neste, os melhores resultados obtidos foram nas categorias "paixão" e "excitação/entusiasmo", com 50,2% e 56,2% respectivamente. As outras categorias obtiveram resultados bem inferiores com 30,0% para "alegria", 29,2% para "decepção" e 26,4% para "tristeza".

Tabela 13 – Contribuição das letras da LMMD por categoria.

	Stemm	**2-gram**	**3-gram**	**4-gram**
Alegria	30,0%	29,6%	30,8%	31,6%
Paixão	50,2%	47,8%	47,2%	47,6%
Decepção	29,2%	0,0%	0,0%	0,0%
Excitação/Entusiasmo	**56,2%**	**58,8%**	**59,6%**	**59,6%**
Amor	45,2%	41,4%	41,6%	41,8%
Tristeza	26,4%	9,2%	9,2%	9,2%

No segundo experimento utilizando *n-grams*, foi possível observar que é insignificante o aumento na precisão conforme é incrementado o valor de N. Nota-se também que todos os algoritmos de classificação falharam na categorização de "decepção" em todas as variações de N. O melhor resultado entre os *n-grams* foi de 59,6% com 4-grams para "excitação/entusiasmo". Já a categoria "tristeza" obteve os piores resultados com este método, restando com uma assertividade de 9,2%.

5.1.8 Combinando as Representações das Letras

Na Tabela 14 é demonstrado o desempenho dos algoritmos de classificação através da média ponderada e o desvio padrão das representações das letras da base de dados da LMMD, que tiveram suas características combinadas através de *early fusion*, ou seja, as características extraídas com o método de *stemming* foram combinadas com aquelas extraídas utilizando o método *n-gram*, relacionando os três níveis *bigrams* (para n=2), *trigrams* (para n=3) e *quadgrams* (para n=4) das letras com *stemming*. O desempenho apresentado, ao realizar a junção destas características e a validação cruzada de 5 vezes, foi inferior ao obtido de forma individual. O melhor resultado apresentado quando combinados foi com o algoritmo SVM, de 39,97%, desta forma não há vantagem em realizar a fusão das representações.

Tabela 14 – Resultado das representações das letras da LMMD combinadas com método *early fusion.*

Classificadores	**Desempenho**
Decision Tree	33,72 ± 0,018
k-NN	25,08 ± 0,044
Gaussian NB	39,02 ± 0,010
SVM	**39,97 ± 0,006**
Logistic Regression	37,05 ± 0,029

5.1.9 Seleção Dinâmica de Classificadores com as Representações das Letras

Utilizando as representações das letras da base de dados LMMD, foram extraídas as características destas letras através do método de *n-grams*, os *bigrams* (para n=2), *trigrams* (para n=3) e *quadgrams* (para n=4), além de realizar o fusão das características com *early fusion* entres os diferentes níveis de "N" juntamente à representação, utilizando *stemming*, para realização dos experimentos, que demonstram a utilização de um *pool* com 30 classificadores heterogêneos, sendo 5 classificadores dos algoritmos k-NN, *Decision trees*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, onde cada classificador é treinado com um *fold* distinto da validação cruzada, criando uma maior diversidade entre os classificadores. Importante ressaltar que antes do *early fusion* é realizado o TF/IDF das representações e que os classificadores foram treinados com 70% da base, sendo reservados 30% para aplicar os algoritmos de classificação dinâmica.

Os resultados demonstrados na Tabela 15 apresentam os desempenhos dos algoritmos de seleção dinâmica utilizando as letras da LMMD com rótulos de sentimento, obtendo um oráculo de 94,84%. O algoritmo que melhor classifica as amostras, considerando o desbalanceamento das representações, é obtido pelo algoritmo MCB com *f-measure* de 35% e acurácia de 35,45%. Este algoritmo tem como característica avaliar o nível de competência, que é definido pelo k-NN de cada classificador individual, considerando a precisão local do classificador base. O conjunto contendo os k-vizinhos mais próximos é filtrado com base na similaridade da amostra consultada.

Tabela 15 – Resultados da seleção dinâmica utilizando as representações das letras da LMMD. Em negrito está destacado o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Alegria	32,0%	24,0%	**33,0%**	30,0%	22,0%	23,0%	18,0%
Amor	**44,0%**	42,0%	38,0%	41,0%	40,0%	43,0%	**45,0%**
Decepção	0,0%	**13,0%**	0,0%	0,0%	12,0%	0,0%	0,0%
Excitação/Entusiasmo	36,0%	52,0%	41,0%	44,0%	**57,0%**	47,0%	41,0%
Paixão	42,0%	39,0%	42,0%	36,0%	39,0%	**44,0%**	9,0%
Tristeza	0,0%	7,0%	8,0%	7,0%	**15,0%**	9,0%	0,0%
F-Measure	**32%**	**34%**	**32%**	**32%**	**35%**	**34%**	**23%**
Acurácia	**36,66%**	**35,75%**	**36,06%**	**34,24**	**35,45%**	**33,93%**	**31,81%**

Analisando as categorias individuais, é possível observar que os classificadores fa-

lham na categoria "decepção". O mesmo comportamento é visto com a categoria "tristeza", que tem seu melhor desempenho no alcance de 15,0% com o classificador MCB. O melhor resultado entre os sentimentos com o algoritmo MCB, é da categoria "excitação/entusiasmo", com precisão de 57,0%.

5.1.10 Seleção dinâmica de Classificadores com Todas as Representação

Neste experimento, todas as características das representações áudio, visual e letras, foram juntadas através do método *early fusion*. Para este experimento fora utilizado um conjunto de classificadores heterogêneo, contendo 30 classificadores do tipo k-NN, *Decision trees*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, em quantidades iguais, ou seja, 5 de cada tipo acima indicado.

Os classificadores foram previamente treinados com validação cruzada de 5 vezes, utilizando 70% da base de dados, enquanto o restante da base foi reservado para aplicar os algoritmos de seleção dinâmica de conjuntos de classificadores para a realização da rotulação da amostra de teste. A Tabela 16 apresenta os resultados desses algoritmos de seleção dinâmica, onde cada célula representa a média aritmética *f-measure* para cada uma das categorias da LMMD. Os resultados obtidos para o oráculo foram de 97,72% nesse conjunto inicial de classificadores.

Analisando o desempenho dos algoritmos, pode-se observar que o DES-P apresenta o melhor resultado, com uma acurácia de 77,23% e com média aritmética de 77%. Outro algoritmo com resultados expressivos é o KNORA-U com *f-measure* de 73% e acurácia de 74,24%.

Tabela 16 – Resultados da seleção dinâmica utilizando as representações do áudio da LMMD. Em negrito está destacado o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Alegria	**70,0%**	63,0%	67,0%	48,0%	54,0%	59,0%	29,0%
Amor	**74,0%**	70,0%	**76,0%**	51,0%	64,0%	65,0%	48,0%
Decepção	0,0%	0,0%	**67,0%**	0,0%	0,0%	0,0%	0,0%
Excitação/Entusiasmo	**74,0%**	58,0%	**72,0%**	51,0%	60,0%	59,0%	52,0%
Paixão	80,0%	70,0%	**85,0%**	65,0%	76,0%	71,0%	60,0%
Tristeza	69,0%	54,0%	**75,0%**	62,0%	69,0%	54,0%	36,0%
F-Measure	**73%**	**64%**	**77%**	**55%**	**65%**	**62%**	**47%**
Acurácia	**74,24%**	**65,10%**	**77,27%**	**55,30%**	**64,39%**	**65,15%**	**53,03%**

Analisando as categorias individuais, nota-se que o sentimento de "decepção" só é classificado pelo algoritmo DES-P devido à característica que este algoritmo possui de selecionar classificadores base que obtenham algum desempenho na região de competência correspondente àquela categoria. Desta forma, mesmo com poucas representações, acaba obtendo uma melhor performance. A categoria "excitação/entusiasmo" trouxe um bom desempenho com o algoritmo KNORA-U devido a quantidade de representações, uma vez que o KNORA-U seleciona todos os classificadores que alcançaram um resultado correto

pelo menos em uma amostra pertencente a região de competência, agregando os resultados para obter a decisão final. O melhor resultado obtido é com a categoria "paixão", devido ao bom desempenho observado nos classificadores simples, que compõe o conjunto de classificadores para treinamento.

5.1.11 Considerações

Nesta seção, será apresentado um resumo dos resultados obtidos para a base de dados LMMD. Na Tabela 17 está demonstrada a média ponderada de todos os algoritmos de classificação por cada representação combinada (*early fusion*), destacando os melhores resultados da classificação monolítica e da seleção dinâmica de classificadores.

A fim de comparar os experimentos, apesar de algumas representações apresentarem melhores resultados com características específicas, esses foram realizados somente com as representações combinadas, na intenção de verificar os classificadores treinados com características distintas, gerando uma melhora nos resultados com a utilização do *pool* heterogêneo de seleção dinâmica.

Tabela 17 – Comparativo de todos os resultados das representações da base de dados LMMD combinadas.

	Representações combinadas			
Classificadores	**Audio**	**Visual**	**Letra**	**Todas**
Decision Tree	35,6%	28,2%	33,8%	42,3%
k-NN	46,0%	33,0%	25,0%	49,7%
Gaussian NB	26,7%	28,8%	39,0%	39,4%
SVM	**54,9%**	34,8%	**39,9%**	**68,3%**
Logistic Regression	39,3%	**35,8%**	37,0%	59,8%
KNORA-U	**66,7%**	**37,6%**	32,2%	73,9%
KNORA-E	54,9%	33,0%	34,0%	64,1%
DES-P	**66,4%**	**37,2%**	32,0%	**77,8%**
OLA	47%	32,0%	32,1%	55,6%
MCB	56,8%	29,1%	**35,0%**	65,9%
A-Priori	45,7%	34,1%	34%	62,9,4%
A-Posteriori	32,1%	31,9%	23,2%	47,3%

É possível observar um interessante resultado demonstrado com a seleção dinâmica quando utilizada a fusão de todas as representações, ou seja, com uma maior diversidade de informações, resultando em uma diferença de aproximadamente 10% em relação a outras representações ou mesmo utilizando um classificador monolítico, embora o valor do oráculo não tenha sido atingido em nenhuma das situações.

Destacando-se também, que diferentemente do apresentado com algoritmos monolíticos, que quando demasiadamente aumentado o tamanho do vetor de características resulta em uma diminuição na capacidade de precisar uma amostra de teste, na seleção

dinâmica ocorreu uma melhora, com destaque e para o algoritmo DES-P, com o melhor resultado, chegando a 77%. O melhor resultado com classificadores monolíticos ocorreu com a representação do áudio, de 66%.

5.2 BRMD - MOOD

5.2.1 Classificação Utilizando as Representações do Áudio

Utilizando as características extraídas do sinal do áudio, como a acústica, o timbre e o ritmo, através das representações SSD, RH, RP e MFCC das músicas da base de dados BRMD, foram treinados e avaliados os classificadores, através do processo de validação cruzada, onde a base de dados é dividida em *folds*, neste caso 5 partes, contando o mesmo número de amostras em cada um dos *folds* para cada um dos sentimentos analisados. Desta forma, cada parte é utilizada ou para treinamento ou para teste, repetindo n vezes, assim cada parte é utilizada ao menos uma vez.

Tabela 18 – Resultados dos classificadores por cada representação do áudio para BRMD-MOOD. Em negrito está destacado o melhor classificador para cada representação.

Classificadores	**SSD**	**RP**	**RH**	**MFCC**
Decision Tree	18,6 ± 0,014	17,4 ± 0,019	15,3 ± 0,020	18,3 ± 0,008
k-NN	21,7 ± 0,018	21,6 ± 0,008	20,0 ± 0,024	23,0 ± 0,01577
Gaussian NB	07,5 ± 0,012	12,5 ± 0,003	06,9 ± 0,008	16,9 ± 0,020
SVM	**25,7 ± 0,029**	**25,0 ± 0,020**	21,6 ± 0,015	**32,6 ± 0,035**
Logistic Regression	23,9 ± 0,027	23,0 ± 0,024	**22,3 ± 0,011**	25,6 ± 0,013

Conforme se pode observar na Tabela 18, o algoritmo SVM apresentou os melhores resultados nas representações SSD, RP e MFCC, com índices de 25,7%, 25,0% e 32,6% de *f-measure* respectivamente. O algoritmo *Logistic Regression*, obteve o melhor resultado para representação RH, 22,3%. Já o algoritmo *Gaussian NB* teve resultados irrisórios em todas as representações. Os resultados obtidos em geral são baixos, sendo possível observar que as diferentes representações são melhor classificadas por um tipo distinto de classificador, mas mesmo assim apresentam muita dificuldade em precisar uma categoria da BRMD-MOOD.

5.2.1.1 Contribuição

Na Tabela 19 são apresentados os resultados obtidos pela utilização do melhor classificador da representação, no qual cada célula descreve a média ponderada *f-measure* resultante para cada categoria. Devido a baixa quantidade de amostras para a categoria "*aggressive*", somente a representação SSD contribuiu para a classificação, com um resultado de 31,4%. As categorias "*brooding*", com 40,6%, "*fiery*", com 45,0% e "*empowering*", com 36,2%, apresentam todas um bom desempenho com a mesma representação. Já o

MFCC traz poucas contribuições, tendo o pior desempenho médio e não conseguindo classificar as categorias "*aggressive*", "*brooding*" e "*defiant*". Os descritores RP e RH têm bom desempenho para as categorias "*empowering*" e "*fiery*", já nas demais apresentam desempenhos semelhantes.

Tabela 19 – Resultado das representações do áudio por categoria para BRMD-MOOD.

	SSD	**RP**	**RH**	**MFCC**
Aggressive	31,4%	0,0%	0,0%	0,0%
Brooding	40,6%	27,6%	8,6%	0,0%
Cool	22,0%	14,6%	21,2%	23,6%
Defiant	16,6%	2,4%	0,0%	0,0%
Easygoing	12,0%	15,8%	7,2%	3,4%
Empowering	36,2%	**34,2%**	25,4%	42,2%
Energizing	9,6%	10,0%	7,0%	19,6%
Excited	16,4%	27,8%	18,4%	23,8%
Fiery	**45,0%**	31,2%	**31,2%**	**56,6%**
Gritty	5,6%	9,0%	3,6%	5,0%
Lively	19,2%	20,8%	9,0%	24,6%
Melancholy	14,6%	16,6%	1,4%	14,6%
Romantic	22,8%	20,0%	13,6%	20,0%
Rowdy	12,0%	15,8%	6,8%	16,6%
Sensual	22,6%	23,6%	8,2%	22,8%
Sentimental	21,0%	10,0%	3,6%	8,8%
Somber	15,0%	16,4%	8,2%	7,8%
Sophisticated	4,4%	5,6%	0,0%	18,2%
Stirring	3,8%	14,6%	0,0%	6,6%
Tender	9,2%	13,8%	0,0%	6,4%
Upbeat	5,0%	6,0%	3,2%	10,2%
Urgent	5,8%	9,0%	1,6%	4,8%
Yearning	0,0%	0,0%	0,0%	3,8%

5.2.2 Combinando as Representações do Áudio

Os resultados obtidos combinando as representações do áudio utilizando método de fusão *early fusion*, não apresentam melhoras expressivas em relação ao esperado, sendo possível observar na Tabela 20, que o melhor desempenho é do algoritmo SVM, com 34,32% de *f-measure*, um desempenho melhor em 2% se comparado com a representação MFCC, que, entre as outras representações, tem a melhor média ponderada com SVM.

Tabela 20 – Resultado das representações do áudio da BRMD-MOOD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	17,37 ± 0,022
k-NN	23,94 ± 0,007
Gaussian NB	15,88 ± 0,011
SVM	**34,32 ± 0,017**
Logistic Regression	25,57 ± 0,025

5.2.3 Seleção Dinâmica de Classificadores com as Representações do Áudio

Utiliza somente características das representações do áudio, ou seja, SSD, RH, RP e MFCC e, realizando *early fusion*, cria um único vetor para treinamento com um conjunto de 30 classificadores compostos pelos algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, que foram treinados com *folds* diferentes, criando um conjunto de classificadores com maior diversidade e características heterogêneas. Os classificadores foram previamente treinados com validação cruzada de 5 vezes, utilizando 70% da base de dados, o restante da base foi reservado para aplicar os algoritmos de seleção dinâmica de conjunto de classificadores, para a realização da rotulação da amostra de teste.

Tabela 21 – Resultados da seleção dinâmica utilizando as representações do áudio da BRMD-MOOD. Em negrito está destacado o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Aggressive	**40,0%**	0,0%	**40,0%**	0,0%	0,0%	0,0%	0,0%
Brooding	**37,0%**	33,0%	35,0%	0,0%	24,0%	0,0%	13,0%
Cool	0,0%	**27,0%**	0,0%	13,0%	0,0%	0,0%	0,0%
Defiant	**22,0%**	0,0%	**22,0%**	0,0%	0,0%	0,0%	0,0%
Easygoing	**51,0%**	17,0%	50,0%	0,0%	46,0%	0,0%	37,0%
Empowering	0,0%	46,0%	0,0%	**52,0%**	0,0%	41,0%	0,0%
Energizing	**46,0%**	0,0%	40,0%	0,0%	23,0%	0,0%	16,0%
Excited	**51,0%**	26,0%	48,0%	12,0%	47,0%	11,0%	42,0%
Fiery	50,0%	50,0%	**53,0%**	47,0%	42,0%	41,0%	0,0%
Gritty	26,0%	33,0%	24,0%	**38,0%**	20,0%	33,0%	21,0%
Lively	0,0%	17,0%	0,0%	22,0%	0,0%	**25,0%**	0,0%
Melancholy	39,0%	0,0%	**41,0%**	0,0%	33,0%	0,0%	20,0%
Romantic	26,0%	31,0%	26,0%	**36,0%**	21,0%	28,0%	7,0%
Rowdy	**37,0%**	24,0%	25,0%	16,0%	11,0%	13,0%	0,0%
Sensual	0,0%	22,0%	0,0%	10,0%	25,0%	17,0%	**29,0%**
Sentimental	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Somber	**12,0%**	0,0%	**12,0%**	0,0%	9,0%	0,0%	11,0%
Sophisticated	0,0%	**15,0%**	0,0%	10,0%	0,0%	14,0%	0,0%
Stirring	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Tender	16,0%	0,0%	20,0%	0,0%	16,0%	**22,0%**	8,0%
Upbeat	0,0%	9,0%	0,0%	17,0%	0,0%	**26,0%**	0,0%
Urgent	0,0%	0,0%	0,0%	0,0%	12,0%	**25,0%**	6,0%
Yearning	0,0%	**10,0%**	0,0%	0,0%	0,0%	9,0%	0,0%
F-Measure	**33,0%**	**28,0%**	**32,0%**	**27,0%**	**27,0%**	**25,0%**	**20,0%**
Acurácia	**38,56%**	**32,23%**	**37,19%**	**29,57%**	**30,30%**	**25,61%**	**25,43%**

Para a rotulação das 25 emoções da base de dados BRMD, utilizando os algoritmos de seleção dinâmica foi obtido um oráculo de 75,75%. Dos resultados demonstrados na Tabela 21, utilizando a média ponderada *f-measure*, nota-se que todos os classificadores falharam no reconhecimento das emoções "*sentimental*" e "*stirring*". O algoritmo KNORA-U obteve o melhor desempenho com acurácia de 38,56% e *f-measure* de 33%. O classificador *A-Posteriori* é o que tem pior desempenho com acurácia de 25,43% e *f-measure* de 20,0%. O algoritmo usa a propriedade de classificação correta, dado um classificador base para cada vizinho em relação a uma única classe, ponderando a influência de cada vizinho de

acordo com a distância euclidiana de uma amostra, no entanto se mostrou pouco eficiente para precisar as categorias através das características do áudio.

Analisando as categorias individualmente, a categoria "*empowering*", obteve o melhor resultado com o algoritmo OLA, de 52,0%. O algoritmo avalia o nível de competência, ou seja, a acurácia de todos os classificadores de modo individual, sendo que o mais competente para prever uma amostra é o selecionado. Outra categoria com bom desempenho foi a "*fiery*", com o algoritmo DES-P.

5.2.4 Classificação Utilizando Representações Visuais

As características são extraídas pelo método de extração global de características e pela utilização do $LBP_{8,2}$ que, conforme observado por Costa, Oliveira e Silla Jr (2017), são os que obtiveram alguns dos melhores resultados na classificação de gêneros, contendo o menor custo para extração, portanto utilizando os mesmos princípios para a representação dos rótulos utilizados neste trabalho. Os resultados estão divididos pela variação de espectrogramas, que são gerados com o limite inferior definido em -50 dBFS, -70 dBFS, -90 dBFS, -110 dBFS e -130 dBFS (o limite superior é sempre 0 dBFS).

O algoritmo *Logistic Regression* obteve os melhores resultados com todos os descritores, conforme demonstrado na Tabela 22, muito pelas características dos dados da LBP, obtendo assertividade média de 32,2%, com amplitude de -90 dBFS. Outros algoritmos, SVM e KNN, obtiveram desempenho satisfatório com a amplitude de -90 dBFS, alcançando resultados de 29,5% e 23,8%, respectivamente. Os resultados mais baixos foram obtidos pela *Gaussiana NB*, que ainda em seu melhor desempenho com -130 dBFS, alcançou apenas 19,2%.

Tabela 22 – Resultado dos classificadores por cada representação visual na BRMD-MOOD. Em negrito se destaca o melhor classificador para cada representação.

Classificadores	**LBP-50**	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Decision Tree	20,1 ± 0,012	22,1 ± 0,018	21,5 ± 0,008	21,8 ± 0,021	22,2 ± 0,012
k-NN	22,4 ± 0,008	23,8 ± 0,025	23,8 ± 0,022	24,2 ± 0,018	22,0 ± 0,011
Gaussian NB	13,8 ± 0,007	12,3 ± 0,006	18,2 ± 0,033	18,1 ± 0,010	19,2 ± 0,009
SVM	28,1 ± 0,013	29,5 ± 0,025	27,7 ± 0,013	28,9 ± 0,016	27,9 ± 0,020
Logistic Regression	**31,9 ± 0,017**	**32,0 ± 0,012**	**32,2 ± 0,014**	**30,2 ± 0,021**	**31,3 ± 0,014**

5.2.4.1 Contribuição

Os resultados apresentados na Tabela 23, são referentes a base de dados BRMD e foram obtidos utilizando o melhor classificador de cada descritor, executando a validação cruzada 5 vezes, sendo que cada célula da tabela descreve a média ponderada *f-measure* para cada categoria. As categorias "*aggressive*", "*brooding*", "*defiant*", "*gritty*", "*melancholy*", "*urgent*", "*yearning*", apresentaram os piores resultado com -50, -70 e -90 dBFS, portanto essas amplitudes contribuíram pouco para o reconhecimento dessas categorias.

Tabela 23 – Resultado das representações da imagem por categoria para BRMD-MOOD.

	LBP-50	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Brooding	6,6%	0,0%	8,0%	0,0%	5,4%
Cool	30,0%	29,8%	25,4%	22,0%	26,8%
Defiant	0,0%	0,0%	18,0%	13,0%	15,2%
Easygoing	9,8%	2,4%	0,0%	12,4%	17,0%
Empowering	**34,6%**	**43,4%**	42,0%	**36,6%**	**34,0%**
Energizing	6,2%	8,8%	16,6%	9,2%	10,8%
Excited	20,8%	8,8%	9,6%	16,0%	17,8%
Fiery	**42,4%**	**52,0%**	**53,2%**	**41,0%**	**42,0%**
Gritty	1,6%	0,0%	0,0%	3,6%	6,2%
Lively	13,0%	17,4%	16,4%	11,6%	20,6%
Melancholy	9,6%	0,0%	0,0%	10,6%	11,6%
Romantic	22,6%	23,8%	**26,4%**	16,2%	18,4%
Rowdy	13,4%	12,2%	18,0%	11,0%	**26,2%**
Sensual	18,4%	**29,4%**	14,8%	14,6%	25,0%
Sentimental	16,8%	12,0%	9,4%	15,4%	19,6%
Somber	11,2%	15,8%	12,4%	14,0%	4,4%
Sophisticated	14,6%	5,2%	10,4%	5,0%	5,4%
Stirring	9,6%	10,0%	20,0%	4,6%	8,8%
Tender	14,8%	8,8%	13,0%	8,0%	11,8%
Upbeat	10,2%	4,8%	0,0%	3,6%	13,4%
Urgent	0,0%	0,0%	9,2%	4,2%	2,4%
Yearning	4,4%	6,8%	0,0%	0,0%	5,4%

Em contraponto, a categoria "*fiery*" apresenta os melhores resultados em todas as amplitudes, com a maior precisão, de 53,2%, na amplitude -90 dBFS. Outra categoria com bom desempenho foi a "*empowering*", com seus melhores resultados em 43,4% e 42,0% nas amplitudes -70 dBFS e -90 dBFS respectivamente. Para a categoria "*cool*", o melhor resultado encontrado foi com a amplitude de -50 dBFS, de 30,0%. Os resultados obtidos com LBP e utilizando classificadores otimizados, ao serem comparados com outras características obtiveram resultados satisfatórios. No geral os classificadores apresentam dificuldades em classificar uma amostra devido ao grande número de categorias com baixa representatividade, o que traz dificuldade ao treinamento.

5.2.5 Combinando as Representações Visuais

Na Tabela 24 é possível visualizar os resultados referentes à média ponderada, indicada em cada célula, e o desvio padrão do desempenho dos algoritmos de classificação para as representações visuais após fusão das amplitudes do LBP de -50, -70, -90, -110 e -130 dBFS, utilizando a técnica de fusão *early fusion*. Neste caso, o algoritmo *Logistic Regression*, obteve um desempenho de 32,13%, muito próximo aos demais níveis de LBP, concluindo que combinar as representações individuais não se mostra vantajoso, vez que inclui um processo a mais na classificação sem uma expressiva melhora nos resultados.

Tabela 24 – Resultado das representações visuais da BRMD-MOOD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	23,94 ± 0,026
kNN	27,38 ± 0,011
Gaussian NB	17,88 ± 0,017
SVM	30,97 ± 0,015
Logistic Regression	**32,13 ± 0,009**

5.2.6 Seleção Dinâmica de Classificadores com as Representações Visuais

Com as características das representações visuais utilizando $LBP_{8,2}$ e extração global de características e a fusão das variações das amplitudes do LBP entre -50, -70, -90, -110 e -130 dBFS, através da técnica de *early fusion*, é criando um conjunto de classificadores com os algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, contendo 5 classificadores de cada algoritmo. Utilizando 70% da base de dados, estes classificadores foram previamente treinados através da classificação cruzada de 5 vezes, onde cada classificador fora treinado com um *fold*, criando um conjunto de classificadores com maior diversidade e características heterogêneas. O restante da base foi reservado para aplicação dos algoritmos de seleção dinâmica de conjuntos de classificadores.

Os resultados alcançados pela utilização da base de dados BRMD, com 25 categorias de sentimento estão expostos na Tabela 25, obtendo um oráculo de 85% para este conjunto inicial de classificadores. Nota-se que todos os classificadores falharam no reconhecimento das emoções "*aggressive*", "*gritty*" e "*stirring*".

O algoritmo KNORA-U apresentou o melhor desempenho, com acurácia de 40% e *f-mensure* 36%. Já o algoritmo MCB é o que melhor classificou a categoria Sentimental, devido a sua característica de considerar, dentro do conjunto de classificadores, aqueles que possuem a melhor precisão local na região de competência, calculando o k-NN entre estes classificadores, selecionando então os k-vizinhos mais próximos da amostra de teste, que são usados para a decisão final.

5.2.7 Classificação Utilizando Representações das Letras

Neste experimento, foram utilizadas as letras da base de dados BRMD-MOOD com classificadores, com parâmetros de otimização e o dicionário de português para as remoções de *stopwords*, considerando se tratar de uma base exclusivamente de letras brasileiras. Os resultados apresentados estão organizados de forma a apresentar a média ponderada *f-measure*, e o desvio padrão dos experimentos, onde cada célula representa o desempenho dos algoritmos de classificação com a validação cruzada de 5 vezes. Na Tabela 26 se observa que o algoritmo SVM, com parâmetros ajustados para melhorar a

Tabela 25 – Resultados da seleção dinâmica utilizando as representações visuais da BRMD-MOOD. Em negrito, consta o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Brooding	**50,0%**	31,0%	**50,0%**	33,0%	20,0%	27,0%	13,0%
Cool	**32,0%**	24,0%	**32,0%**	24,0%	19,0%	18,0%	33,0%
Defiant	40,0%	**44,0%**	40,0%	25,0%	25,0%	20,0%	0,0%
Easygoing	**27,0%**	18,0%	25,0%	10,0%	21,0%	0,0%	9,0%
Empowering	47,0%	42,0%	**50,0%**	42,0%	46,0%	34,0%	36,0%
Energizing	9,0%	8,0%	17,0%	7,0%	0,0%	**25,0%**	10,0%
Excited	19,0%	**24,0%**	18,0%	20,0%	17,0%	16,0%	17,0%
Fiery	**58,0%**	53,0%	57,0%	53,0%	53,0%	49,0%	52,0%
Lively	22,0%	21,0%	19,0%	**25,0%**	21,0%	17,0%	17,0%
Melancholy	**40,0%**	**40,0%**	**40,0%**	20,0%	20,0%	0,0%	25,0%
Romantic	**36,0%**	32,0%	34,0%	31,0%	30,0%	32,0%	28,0%
Rowdy	26,0%	19,0%	25,0%	14,0%	24,0%	23,0%	**30,0%**
Sensual	**29,0%**	19,0%	27,0%	19,0%	18,0%	25,0%	0,0%
Sentimental	**67,0%**	60,0%	60,0%	44,0%	**67,0%**	36,0%	25,0%
Somber	0,0%	0,0%	0,0%	22,0%	17,0%	9,0%	5,0%
Sophisticated	**29,0%**	16,0%	34,0%	0,0%	0,0%	0,0%	0,0%
Tender	**25,0%**	0,0%	**25,0%**	13,0%	15,0%	20,0%	0,0%
Upbeat	**28,0%**	20,0%	27,0%	0,0%	0,0%	0,0%	0,0%
Urgent	40,0%	0,0%	**50,0%**	21,0%	24,0%	8,0%	0,0%
Yearning	**19,0%**	17,0%	**19,0%**	0,0%	0,0%	0,0%	0,0%
F-Measure	**36%**	**31%**	**36%**	**30%**	**30%**	**27%**	**25%**
Acurácia	**40,00%**	**33,18%**	**39,72%**	**31,36%**	**32,72%**	**29,09%**	**29,09%**

precisão, obteve o melhor resultado médio, de 25,6%. Os demais classificadores obtiveram desempenho inferior, sendo que k-NN obteve o pior resultado, com assertividade de 11.10

Tabela 26 – Resultado dos classificadores pelas representações Stemm e 2, 3 e 4 grams das Letras da BRMD-MOOD.

Classificadores	**Stemm**	**2-gram**	**3-gram**	**4-gram**
Decision Tree	16,4 ± 0,019	17,5 ± 0,011	20,7 ± 0,010	**23,2 ± 0,005**
k-NN	11,1 ± 0,022	12,7 ± 0,004	15,3 ± 0,033	13,0 ± 0,008
Gaussian NB	20,3 ± 0,029	18,3 ± 0,011	07,6 ± 0,005	04,1 ± 0,008
SVM	**25,6 ± 0,035**	21,6 ± 0,003	**21,6 ± 0,003**	21,6 ± 0,003
Logistic Regression	19,4 ± 0,019	**22,3 ± 0,013**	21,4 ± 0,01094	22,4 ± 0,006

Os experimentos utilizando palavras com dimensionalidades reduzidas em *n-grams* tiveram resultados que apresentaram uma melhora conforme o aumento do número de termos analisados. O melhor resultado médio obtido foi com a representação de 4-grams, com 23,20% para o algoritmo *Decision tree*. Já o algoritmo *Logistic Regression* tem os melhores resultados para 2-grams e 4-grams, apresentando 22,3% e 22,4% respectivamente. Os outros algoritmos apresentaram um baixo desempenho com as representações de letras.

5.2.7.1 Contribuição

Para os experimentos utilizando a BRMD-MOOD, os resultados estão demonstrados na Tabela 27, onde cada célula descreve a média ponderada e o desvio padrão utilizando o melhor classificador geral, após execução na validação cruzada 5 vezes.

Somente 4 categorias obtiveram assertividade acima de 20%, sendo "*easygoing*" (20,0%), "*excited*" (22,4%), "*fiery*" (24,6%) e "*livery*" (24,6%). As categorias "*aggressive*", "*defiant*", "*energizing*", "*rowdy*", "*sensual*", "*sentimental*", "*urgent*" e "*yearning*" não conseguiram ser classificadas por nenhum algoritmo, portanto foram removidas da tabela para facilitar a visualização. Para as demais categorias, podemos observar que os classificadores falharam, obtendo baixo desempenho.

Tabela 27 – Contribuição das letras da BRMD-MOOD por categoria.

	Stemm	**2-gram**	**3-gram-**	**4-gram**
Brooding	10,6%	0,0%	0,0%	0,0%
Cool	16,4%	**13,2%**	0,0%	0,0%
Easygoing	20,0%	0,0%	0,0%	0,0%
Empowering	13,4%	9,6%	0,0%	**8,4%**
Excited	**22,4%**	0,0%	0,0%	0,0%
Fiery	**24,6%**	**35,6%**	**35,6%**	**36,0%**
Gritty	16,0%	0,0%	0,0%	0,0%
Lively	**24,6%**	12,4%	0,0%	6,0%
Melancholy	9,0%	0,0%	0,0%	0,0%
Romantic	13,6%	10,8%	0,0%	7,2%
Somber	5,6%	0,0%	0,0%	0,0%
Sophisticated	4,6%	0,0%	0,0%	0,0%
Stirring	6,0%	0,0%	0,0%	0,0%
Tender	4,8%	2,8%	0,0%	0,0%
Upbeat	2,8%	2,6%	0,0%	0,0%

Utilizando a técnica de *n-grams*, os classificadores falharam ao rotular a maior parte das categorias. Podendo-se observar na Tabela 27, a categoria "*fiery* a melhor classificada, 36,0% com 4-grams. Podemos concluir então, que utilizando somente a técnica de *n-grams* há pouca contribuição para a identificação das categorias.

5.2.8 Combinando as Representações das Letras

Os resultados apresentados utilizando as representações das letras com *n-grams*, extraídos os *bigrams* (para n=2), *trigrams* (para n=3) e *quadgrams* (para n=4) e *stemming* da BRMD-MODD, combinadas com método *early fusion*, método para agregação das características extraídas das letras das músicas, não alcançaram desempenho satisfatório. Demonstrados na Tabela 28, os resultados obtidos pela média ponderada do desempenho dos classificadores são próximos aos obtidos pelas representações individuais, não apresentando vantagem em adicionar mais uma etapa no processo de classificação. O melhor resultado com *early fusion* é dos algoritmos SVM, com 26,36%.

Tabela 28 – Resultado das representações das letras da BRMD-MOOD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	15,97%
k-NN	14,43%
Gaussian NB	21,84%
SVM	**26,36**%
Logistic Regression	21,14%

5.2.9 Seleção Dinâmica de Classificadores com as Representações das Letras

Utilizando as representações das letras, extraídas com método n-grams com *bigrams* (para n=2), *trigrams* (para n=3) e *quadgrams* (para n=4) e realizado *early fusion* com as características extraídas com método *stemming*, serão realizados os experimentos utilizando 70% da base de dados para treinar os algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, que irão compor o *pool* de classificadores. Os outros 30% da base de dados, será utilizado para os algoritmos de seleção dinâmica. Ressaltando que antes de realizar o *early fusion* é realizado o TF/IDF das representações.

Tabela 29 – Resultados das seleções dinâmicas utilizando as representações das letras da BRMD-MOOD. Em negrito, o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Cool	4,0%	9,0%	8,0%	**11,0%**	8,0%	6,0%	0,0%
Empowering	7,0%	**18,0%**	9,0%	14,0%	16,0%	16,0%	17,0%
Energizing	0,0%	0,0%	0,0%	6,0%	**7,0%**	6,0%	0,0%
Excited	0,0%	0,0%	0,0%	0,0%	0,0%	**7,0%**	0,0%
Fiery	**36,0%**	31,0%	35,0%	31,0%	27,0%	31,0%	28,0%
Lively	8,0%	**19,0%**	11,0%	11,0%	11,0%	12,0%	5,0%
Romantic	3,0%	3,0%	4,0%	12,0%	**17,0%**	3,0%	0,0%
Tender	0,0%	3,0%	0,0%	3,0%	3,0%	0,0%	**4,0%**
Upbeat	**20,0%**	**20,0%**	0,0%	18,0%	15,0%	6,0%	0,0%
Urgent	0,0%	0,0%	0,0%	5,0%	0,0%	**18,0%**	0,0%
F-Measure	**10%**	**12%**	**10%**	**12%**	**12%**	**12%**	**9%**
Acurácia	**19,39%**	**16,38%**	**19,96%**	**14,87%**	**13,93%**	**15,25%**	**16,19%**

Na Tabela 29 são demonstrados os resultados dos classificadores de seleção dinâmica, utilizando as letras da base de dados BRMD para rotulação de emoção, com oráculo para esse conjunto inicial de classificadores alcançando 59,88% de taxa de reconhecimento. Todos os classificadores falharam no reconhecimento das emoções "*aggressive*", "*brooding*", "*cool*", "*excited*", "*fiery*", "*gritty*", "*lively*", "*melancholy*", "*romantic*", "*rowdy*", "*sensual*", "*stirring*", "*yearning*", motivo pelo qual, facilitando a visualização da tabela, todas essas categorias foram removidas.

O desempenho da representação das letras é bem inferior se comparado com o de outras representações, sendo o algoritmo KNORA-E o que obteve melhor desempenho, considerando o desbalanceamento das classes, alcançando uma acurácia de 16,38% e

f-measure de 12%. No geral os classificadores tiveram baixo desempenho.

5.2.10 Seleção Dinâmica de Classificadores com todas as Representação

Nestes experimentos foram utilizadas todas as características das representações: áudio, visual e letras, realizando a fusão das características extraídas de vetores diferentes através do método *early fusion* e criado um único vetor para treinamento. Os algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models* compõem o conjunto de classificadores heterogêneo de 30 classificadores, sendo 5 de cada algoritmo. Estes classificadores foram previamente treinadas com validação cruzada de 5 vezes, utilizando 70% da base de dados, enquanto o restante da base foi reservado para aplicação dos algoritmos de seleção dinâmica.

Na Tabela 30, estão presentes os resultados dos algoritmos de seleção dinâmica, onde cada célula representa a média ponderada por categoria para um oráculo de 91,47%.

Tabela 30 – Resultado da seleção dinâmica utilizando todas as representações da BRMD-MOOD. Em negrito, o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Aggressive	**67,0%**	**67,0%**	**67,0%**	0,0%	0,0%	0,0%	0,0%
Brooding	57,0%	57,0%	57,0%	**67,0%**	57,0%	40,0%	0,0%
Cool	71,0%	61,0%	**72,0%**	64,0%	59,0%	61,0%	14,0%
Defiant	**40,0%**	0,0%	**40,0%**	0,0%	0,0%	0,0%	0,0%
Easygoing	44,0%	40,0%	44,0%	52,0%	54,0%	**56,0%**	33,0%
Empowering	**68,0%**	57,0%	65,0%	47,0%	42,0%	38,0%	35,0%
Energizing	**62,0%**	33,0%	**62,0%**	8,0%	10,0%	18,0%	0,0%
Excited	**31,0%**	29,0%	**31,0%**	24,0%	24,0%	17,0%	8,0%
Fiery	**67,0%**	65,0%	64,0%	61,0%	62,0%	64,0%	46,0%
Gritty	40,0%	**47,0%**	40,0%	40,0%	44,0%	36,0%	0,0%
Lively	**49,0%**	35,0%	47,0%	35,0%	32,0%	37,0%	21,0%
Melancholy	**83,0%**	62,0%	**83,0%**	71,0%	71,0%	56,0%	20,0%
Romantic	58,0%	52,0%	**59,0%**	50,0%	51,0%	46,0%	36,0%
Rowdy	**67,0%**	59,0%	64,0%	33,0%	46,0%	39,0%	29,0%
Sensual	59,0%	51,0%	**61,0%**	51,0%	51,0%	48,0%	8,0%
Sentimental	**76,0%**	56,0%	70,0%	18,0%	55,0%	35,0%	25,0%
Somber	0,0%	0,0%	0,0%	**67,0%**	**67,0%**	**67,0%**	50,0%
Shisticated	**62,0%**	44,0%	60,0%	54,0%	50,0%	58,0%	17,0%
Stirring	0,0%	0,0%	0,0%	0,0%	**67,0%**	0,0%	0,0%
Tender	67,0%	57,0%	67,0%	**77,0%**	73,0%	57,0%	0,0%
Upbeat	58,0%	52,0%	**60,0%**	42,0%	45,0%	44,0%	21,0%
Urgent	44,0%	**57,0%**	50,0%	18,0%	18,0%	17,0%	0,0%
Yearning	47,0%	37,0%	46,0%	51,0%	57,0%	**58,0%**	39,0%
F-Measure	**61%**	**53%**	**58%**	**47%**	**49%**	**49%**	**28%**
Acurácia	**61,22%**	**53,40%**	**60,22%**	**50,14%**	**52,69%**	**50,99%**	**37,21%**

A categoria "*aggressive*", apresentou bom desempenho com algoritmos que utilizam propriedades das amostras do conjunto de validação presentes na sua região de vizinhança, identificando, desta forma, o melhor conjunto de classificadores. As categorias "*brooding*", "*sentimental*" e "*tender*", tiveram os melhores resultados com o algoritmo OLA, que realiza sua avaliação considerando a acurácia de cada classificador individualmente, selecionando o mais competente.

O classificador de seleção dinâmica A-Priori teve os melhores resultados com "*yearning*" (58%), "*somber*" (67%) e "*easygoing*" (56%) que, apesar de uma baixa representatividade para esses sentimentos, o algoritmo calculou o nível de competência dos k classificadores vizinhos, ponderando a influência de cada um e combinando os resultados através do voto majoritário, apresentando assim, um bom desempenho.

O algoritmo KNORA-U apresenta o melhor desempenho, com acurácia de 61,22% e *f-measure* de 61%, devido à forma como realiza a seleção do conjunto de classificadores para uma amostra. Isso ocorre porque quanto maior o número de classificadores que classifica corretamente uma amostra pertencente a uma região de competência, melhor será o resultado atingido.

5.2.11 Considerações

Esta seção apresenta um resumo do desempenho dos algoritmos, comparando os resultados obtidos utilizando a abordagem monolítica e a seleção dinâmica de classificadores. Na Tabela 31 estão destacados os melhores resultados em cada representação, observando que foram consideradas somente as representações combinadas através do método *early fusion*, apresentando, neste caso, uma melhora no desempenho ao comparar com as representações individuais ou similares, não causando, desta forma, prejuízo a comparação.

Tabela 31 – Comparativo de todos os resultados das representações da base de dados BRMD-MOOD combinadas.

	Representações			
Classificadores	**Audio**	**Visual**	**Letra**	**Todas**
Decision Tree	17,4%	23,9%	16,0%	42,6%
k-NN	23,9%	27,4%	14,4%	49,3%
Gaussian NB	15,9%	17,9%	21,8%	39,70,%
SVM	**34,3%**	31,0%	**26,4%**	**68,3%**
Logistic Regression	25,6%	**32,1%**	21,1%	59,3%
KNORA-U	**33,0%**	**36,0%**	10,0%	**61,0%**
KNORA-E	28,0%	31,0%	12,0%	53,0%
DES-P	32,0%	**36,0%**	10,0%	58,0%
OLA	27,0%	30,0%	12,0%	47,0%
MCB	27,0%	30,0%	12,0%	49,0%
A-Priori	25,0%	27,0%	12,0%	49,0%
A-Posteriori	20,0%	25,0%	9,0%	28,0%

Para a base de dados BRDM-MOOD, a adição do processo de seleção dinâmica de subconjuntos de classificadores não melhorou a assertividade, mantendo em geral um baixo desempenho, sendo que o melhor resultado foi com o algoritmo monolítico SVM, com todas as representações alcançando desempenho de 68,8%. Podemos observar também que

as representações de letras, quando combinadas com o aumento do número de atributos, diminuiu o desempenho da seleção dinâmica.

5.3 LMD

5.3.1 Classificação Utilizando Representações do Áudio

Nesta seção foram analisadas as representações do áudio para rotular gênero musical utilizando LMD com *artist filter*. Após a otimização de parâmetros independentes, cada classificador é treinado com a validação cruzada entre 3 *folds*, igualmente divididos com 30 músicas para cada categoria e posteriormente validados com um *fold* de validação contendo 400 canções.

Tabela 32 – Resultados dos classificadores com as representações do áudio da LMD para Gênero. Em negrito, o melhor classificador para cada representação.

Classificadores	**SSD**	**RP**	**RH**	**MFCC**
Decision Tree	50,3 ± 0,034	40,1 ± 0,027	30,7 ± 0,014	51,3 ±0,042
k-NN	66,9 ± 0,038	56,6 ± 0,044	42,8 ± 0,016	55,6 ± 0,034
Gaussian NB	58,0 ± 0,044	62,2 ± 0,019	47,6 ± 0,048	65,8 ± 0,028
SVM	**76,62 ± 0,046**	72,1 ± 0,025	**52,4 ± 0,039**	**78,5 ± 0,016**
Logistic Regression	**76,6 ± 0,050**	**75,4 ± 0,026**	50,5 ± 0,034	75,8 ± 0,034

Na Tabela 32 observamos que a classificação pelas representações do áudio, obteve bom desempenho, em especial com SSD e MFCC que obtiveram, nessa ordem, 76,62% e 78,50% de média ponderada com o algoritmo SVM. A representação RH não contribui na definição dos gêneros, sendo o seu melhor resultado com a SVM de 52,40% de precisão.

5.3.1.1 Contribuição

Na rotulação de gênero, devido a um padrão nos áudios analisados, os algoritmos apresentaram melhores resultados, como podemos ver na Tabela 33, onde cada célula representa a média ponderada do desempenho do classificador. Os gêneros como “merengue”, “salsa” e “tango”, são melhor classificados com a representação RP, sendo que “merengue” obteve 90,8% de assertividade. Outras representações que obtiveram resultados expressivos classificados foram “axé”, “pagode e “tango, tendo resultados acima de 80%. A representação MFCC obteve uma assertividade de 96,0% para “tango” e a representação SSD, melhor classificou o gênero “bachata” com 88,6%.

5.3.2 Combinando as Representações do Áudio

Nesta seção, estão demonstrados os resultados dos experimentos combinados das representações do áudio da base de dados LMD. Utilizando o método de *early fusion* entre as representações do áudio SSD, RH, RP e MFCC, foi criado um único vetor

Tabela 33 – Resultado das representações do áudio por categoria para LMD.

	SSD	**RP**	**RH**	**MFCC**
Axé	67,20%	57,20%	19,20%	**82,80%**
Bachata	**88,60%**	87,80%	70,40%	77,80%
Bolero	**74,20%**	67,60%	59,80%	61,40%
Forró	61,20%	70,00%	44,40%	**78,40%**
Gaúcha	**68,60%**	60,60%	35,80%	61,20%
Merengue	88,40%	**90,80%**	70,60%	88,00%
Pagote	77,40%	75,00%	52,40%	**85,40%**
Salsa	72,20%	**84,40%**	62,00%	79,80%
Sertanejo	73,60%	57,00%	19,20%	**76,60%**
Tango	88,60%	92,40%	74,80%	**96,00%**

representativo, sendo que os classificadores foram treinados com validação cruzada em 3 *folds* e posteriormente validados com um *fold* específico de validação. Na Tabela 44 estão os resultados da média ponderada *f-measure* do desempenho dos classificadores.

Os algoritmos Logistic Regression e SVM, obtiveram desempenhos próximos com 84,99% e 85,44% respectivamente, o que significa uma melhora de 8% em média se comparado com o desempenho das representações individuais. Os resultados obtidos não puderam ser comparados com os presentes na literatura, vez que não utilizam as mesmas representações e extratores.

Tabela 34 – Resultado das representações do áudio da LMD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	55,50%
k-NN	63,40%
Gaussian NB	68,22%
SVM	84,88%
Logistic Regression	**85,44%**

5.3.3 Seleção Dinâmica de Classificadores com as Representações do Áudio

Para este experimento, também fora realizada a fusão com método *early fusion* entre as representações do áudio SSD, RH, RP e MFCC, transformando-as em um único vetor de características. Na Tabela 35, estão apresentados o desempenho dos algoritmos de seleção dinâmica, onde cada célula representa a média ponderada para um oráculo de 95,45% de taxa de reconhecimento para este conjunto inicial de classificadores.

Os algoritmos que compõem o conjunto de classificadores são: k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, sendo 5 representações de cada algoritmo treinadas com a validação cruzada de 3 folds, utilizando um *fold* de validação

Tabela 35 – Resultados da seleção dinâmica utilizando as representações do áudio da LMD. Em negrito o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Axé	35,0%	43,0%	35,0%	38,0%	36,0%	35,0%	30,0%
Bachata	**100,0%**	**92,0%**	**100,0%**	86,0%	**100,0%**	91,0%	**96,0%**
Bolero	67,0%	64,0%	67,0%	67,0%	58,0%	67,0%	44,0%
Forró	**96,0%**	87,0%	92,0%	76,0%	93,0%	67,0%	52,0%
Gaucha	61,0%	56,0%	61,0%	40,0%	46,0%	42,0%	25,0%
Merengue	87,0%	83,0%	83,0%	87,0%	**100,0%**	77,0%	90,0%
Pagode	89,0%	84,0%	86,0%	76,0%	80,0%	69,0%	71,0%
Salsa	67,0%	73,0%	67,0%	82,0%	75,0%	69,0%	43,0%
Sertaneja	37,0%	37,0%	40,0%	60,0%	56,0%	50,0%	24,0%
Tango	**97,0%**	91,0%	**97,0%**	**94,0%**	94,0%	90,0%	90,0%
F-Measure	**74%%**	**72%**	**74%**	**70%**	**74%**	**65%**	**57%**
Acurácia	**75,75%%**	**73,48%**	**75,00%**	**70,45%**	**72,72%**	**69,69%**	**57,57%**

para os algoritmos de seleção dinâmica. O algoritmo KNORA-U obteve o melhor resultado, com 74% de *f-measure* e 75,75% de acurácia, quando analisada cada categoria "bachata", "forró" e "tango", esse algoritmo tem os melhores resultados, obtendo 100%, 96% e 97% respectivamente. A categoria "merengue" obteve 100% de acerto com o classificador MCB. Já as representações do áudio têm desempenho bastante discreto quando comparadas com os resultados obtidos com as representações visuais.

5.3.4 Classificação Utilizando Representações Visuais

Neste experimento foram utilizados os algoritmos de classificação otimizados para cada limite de amplitude, obtendo os resultados, demonstrados na Tabela 36, com a utilização da LMD para rotulação de gênero através das representações da imagem, ou seja, das imagens geradas a partir do espectrograma do áudio com amplitudes de -50, -70, -90, -110 e -130 dBFS, utilizando $LBP_{8,2}$ e extração global de características.

Tabela 36 – Resultado dos classificadores por cada representação visual na LMD com gênero. Em negrito, o melhor classificador para cada representação.

Classificadores	**LBP-50**	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Decision Tree	49,6 ± 0,042	55,8 ± 0,027	60,3 ± 0,063	58,6 ± 0,042	55,8 ± 0,031
k-NN	63,5 ± 0,030	66,1 ± 0,010	69,0 ± 0,052	67,0 ± 0,038	66,5 ± 0,035
Gaussian NB	39,8 ± 0,029	60,0 ± 0,066	63,5 ± 0,032	60,8 ± 0,023	61,1 ± 0,040
SVM	**82,3 ± 0,027**	80,8 ± 0,033	81,0 ± 0,023	82,3 ± 0,037	80,5 ± 0,013
Logistic Regression	**82,3 ± 0,014**	**82,1 ± 0,022**	**83,6 ± 0,052**	**85,3 ± 0,020**	**84,6 ± 0,034**

Observa-se que o algoritmo com melhor desempenho é o *Logistic Regression* em todas as amplitudes, obtendo o melhor resultado com 85,33% na amplitude de -110 dBFS. Os algoritmos *Decision Tree* e *Gaussian NB* obtiveram desempenho inferior se comparados aos demais algoritmos.

É dever destacar que o algoritmo SVM também apresenta bom desempenho para a definição de gênero, com resultado médio acima de 80%, tendo o melhor desempenho

atingindo 82,33% na amplitude -50 dBFS. Podemos destacar também a variação dos resultados conforme a alteração da amplitude, pois alguns algoritmos, como *Logistic Regression* e SVM, obtiveram uma melhora na assertividade, conforme o aumento da amplitude, no entanto, quando em direção aos extremos houve perda na precisão.

5.3.4.1 Contribuição

Podemos observar uma melhora nos resultados dos classificadores na rotulação de gênero devido a uma presença maior no padrão das imagens geradas. Dos resultados apresentados na Tabela 37, cada célula descreve a média ponderada por categoria. Utilizando o melhor classificador de cada amplitude, realizando o treinamento com validação cruzada em 3 *folds* e posteriormente validando os resultados com um *fold* de validação contendo 40 representação de cada categoria.

Tabela 37 – Resultado das representações da imagem por categoria para LMD.

	LBP-50	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Axé	75,20%	73,60%	78,60%	78,20%	77,20%
Bachata	87,40%	**91,20%**	89,40%	**95,40%**	89,40%
Bolero	79,40%	83,20%	84,20%	85,20%	85,60%
Forró	82,80%	77,60%	83,40%	84,40%	83,60%
Gaúcha	77,00%	69,40%	78,40%	77,20%	82,00%
Merengue	**91,20%**	90,40%	**90,60%**	92,00%	**91,00%**
Pagote	77,60%	77,80%	80,40%	80,20%	80,80%
Salsa	80,60%	86,40%	86,80%	91,60%	88,20%
Sertanejo	78,60%	74,60%	74,00%	72,00%	71,20%
Tango	87,80%	**93,20%**	89,80%	**94,20%**	**95,20%**

Podemos verificar que um destes níveis tem melhor resultado com um tipo diferente de gênero, por exemplo, a amplitude de -50 dBFS teve uma precisão de 91,20% para a categoria “merengue”, enquanto a amplitude de -110 dBFS contribui com um maior número de categorias, com precisão de 95,40% para “bachata”, 85,20% para “forró”, 94,20% para “tango”, 92,00% para “merengue” e 92,60% para “salsa”, sendo também a representação que melhor as classifica. A categoria “axé”, teve o pior desempenho médio, sendo melhor classificada pela amplitude de -90 dBFS, com 78,60%.

5.3.5 Combinando as Representações Visuais

Nestes experimentos foram utilizadas todas os espectrogramas com as variações de amplitudes realizando a extração da LBP, realizando a fusão com método *early fusion* das características criando um único vetor representativo para cada canção analisada. Os resultados apresentados na Tabela 38, onde cada célula corresponde a média ponderada *f-measure*, foram obtidos no experimento com classificadores treinados através da validação

cruzada em 3 *folds* e avaliados através de um *fold* específico de validação, contendo 40 representações de cada gênero da base de dados LMD.

Tabela 38 – Resultado das representações das imagens da LMD combinados com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	62,00 +/ 0,028
k-NN	74,11 +/- 0,030
Gaussian NB	65,33 +/- 0,032
SVM	86,11 +/- 0,029
Logistic Regression	**89,00 +/- 0,019**

Comparando com os trabalhos presentes na literatura, que utilizaram a mesma abordagem de transformação do sinal do áudio em imagem, como realizam Costa et al. (2013) e Nanni et al. (2016), ressaltando que estes não utilizaram a composição das amplitudes dos LBPs, os resultados apresentados no presente trabalho são superiores. Os resultados obtidos combinando os LBPs foram moderadamente melhores que os obtidos de forma individual, com aumento de 4 a 5%. Na mesma proporção, é o melhor desempenho demonstrado neste trabalho.

5.3.6 Seleção Dinâmica de Classificadores com as Representações Visuais

Neste experimento, utilizando as representações visuais com a fusão através do método *early fusion* das variações das amplitudes do LBP entre -50, -70, -90, -110 e -130 dBFS, foi criando um conjunto de classificadores com os algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, sendo 5 representações de cada algoritmo previamente treinado, realizando a validação cruzada com 3 *folds* e aplicando os algoritmos de seleção dinâmica de classificadores para classificar uma dada amostra de teste utilizando um *fold* de validação.

Tabela 39 – Resultados da seleção dinâmica utilizando as representações visuais da LMD.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Axé	83,0%	89,0%	86,0%	76,0%	79,0%	89,0%	77,0%
Bachata	89,0%	**94,0%**	89,0%	89,0%	84,0%	**94,0%**	**94,0%**
Bolero	85,0%	85,0%	85,0%	72,0%	77,0%	79,0%	67,0%
Forró	64,0%	64,0%	62,0%	62,0%	62,0%	64,0%	65,0%
Gaucha	53,0%	38,0%	57,0%	40,0%	30,0%	53,0%	22,0%
Merengue	84,0%	**100,0%**	84,0%	84,0%	89,0%	**95,0%**	84,0%
Pagode	89,0%	86,0%	89,0%	85,0%	86,0%	91,0%	91,0%
Salsa	**94,0%**	**94,0%**	90,0%	88,0%	**94,0%**	91,0%	91,0%
Sertaneja	62,0%	74,0%	64,0%	53,0%	62,0%	57,0%	32,0%
Tango	**97,0%**	**94,0%**	**97,0%**	**97,0%**	**94,0%**	**100,0%**	**97,0%**
F-Measure	**81%**	**83%**	**81%**	**76%**	**77%**	**82%**	**74%**
Acurácia	**81,81%**	**83,33%**	**81,81%**	**76,51%**	**76,51%**	**80,30%**	**74,24%**

Os resultados demonstrados na Tabela 39, da média ponderada F-Measure, para um oráculo para esse conjunto inicial de classificadores de 96,21% de taxa de reconhecimento, são semelhantes aos obtidos no trabalho de Costa et al. (2013) utilizando os algoritmos KNORA-U e KNORA-E. No entanto, no presente trabalho, utilizada a fusão dos níveis de LBP, podemos observar que KNORA-E, apresenta o melhor resultado com *f-measure* de 83% e acurácia de 83,33%.

Este algoritmo tem por característica procurar um classificador que classifique corretamente todas as amostras pertencentes a região de competência da amostra de teste. No caso de nenhum classificador atingir com precisão perfeita, o tamanho da região de competência é reduzido, eliminando os vizinhos mais distantes e reavaliando o desempenho dos classificadores. As saídas do conjunto selecionado de classificadores são combinadas usando o esquema de votação majoritária, se nenhum classificador base for selecionado, todo o conjunto será usado para classificação.

Analisando as categorias, é possível observar que "merengue", "bachata", "salsa" e "tango" são as que apresentam melhor desempenho, com 100% e 94% de precisão respectivamente, com KNORA-U. A categoria Tango obteve 100% de assertividade pelo algoritmo *A-Priori*.

5.3.7 Seleção Dinâmica de Classificadores com Todas as Representação

Nesta seção, são analisados os algoritmos de classificação dinâmica depois de realizada a fusão, pelo método *early fusion*, com todas as características das representações de áudio e visual da base de dados LMD, ou seja, agrupando as representações anteriormente indicadas. Salientando que nesta categoria não temos a representação das letras.

Com um oráculo de 90,98% para o conjunto inicial de classificadores heterogêneos, composto pelos classificadores k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, onde 5 representações de cada algoritmo são treinadas com a validação cruzada em 3 *folds*, foi utilizado um *fold* específico para validação para os algoritmos de seleção dinâmica.

Na Tabela 40, são demonstrados os resultados obtidos da média ponderada *f-measure*, que por sua vez são bastante frustrantes, o desempenho com todas as representações é menor que aquele alcançado utilizando somente uma representação. O algoritmo KNORA-E tem o melhor desempenho, com 75% de *f-measure* e 75% de acurácia.

5.3.8 Considerações

O melhor resultado obtido para a base de dados LMD foi utilizando a composição das representações visuais com o classificador Logistic Regression, obtendo um desempenho de 89%, o que é superior a alguns resultados encontrados na literatura, no entanto é inferior

Tabela 40 – Resultado da seleção dinâmica utilizando todas as representações da LMD.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Axé	83,0%	57,0%	40,0%	58,0%	56,0%	42,0%	82,0%
Bachata	44,0%	**94,0%**	85,0%	94,0%	**91,0%**	**94,0%**	**97,0%**
Bolero	87,0%	62,0%	33,0%	39,0%	48,0%	45,0%	24,0%
Forró	76,0%	67,0%	77,0%	72,0%	64,0%	81,0%	81,0%
Gaúcha	**92,0%**	72,0%	54,0%	61,0%	62,0%	72,0%	67,0%
Merengue	73,0%	**96,0%**	80,0%	**100,0%**	88,0%	88,0%	91,0%
Pagode	36,0%	77,0%	65,0%	65,0%	62,0%	80,0%	68,0%
Salsa	71,0%	59,0%	53,0%	75,0%	62,0%	71,0%	38,0%
Sertaneja	25,0%	62,0%	36,0%	54,0%	59,0%	42,0%	40,0%
Tango	**93,0%**	90,0%	**90,0%**	87,0%	90,0%	81,0%	90,0%
F-Measure	**68%**	**75%**	**63%%**	**71%**	**70%**	**71%**	**69%**
Acurácia	**68,18%**	**75%**	**62,87%**	**70,45%**	**71,21%**	**71,96%**	**68,18%**

ao obtido por Costa, Oliveira e Silla Jr (2017) utilizando redes neurais convolutivas, que alcançaram 92,00% de assertividade. De maneira geral, a combinação das representações aumentou o desempenho dos classificadores, mas a utilização da seleção dinâmica de classificadores se demonstrou pouco eficiente quando aumentado o tamanho do vetor de características.

Tabela 41 – Comparativo de todos os resultados das representações da base de dados LMD combinadas.

	Representações		
Classificadores	**Audio**	**Visual**	**Todas**
Decision Tree	55,5%	62.0	42.6%
k-NN	63,4%	74.1	49.3%
Gaussian NB	68,2%	65.3	39.7%
SVM	**84,8%**	86.1	**68.3%**
Logistic Regression	**85,4%**	**89.0**	59.3%
KNORA-U	**75,7%**	81,8%	68%
KNORA-E	72,8%	**83,3%**	**75,0%**
DES-P	74,0%	81,8%	63,1%
OLA	70,4%	76,5%	71.4%
MCB	74,7%	77,5%	70,4%
A-Priori	65,6%	82,3%	71,2%
A-Posteriori	57,5%	74,2%	69,1%

5.4 BRMD - Gênero

5.4.1 Classificação Utilizando Representações do Áudio

Para a representação do áudio com a rotulação por gêneros da base de dados BRMD é utilizada a validação cruzada em 4 *folds*. Conforme demonstrado na Tabela 42, com a média ponderada *f-measure* o melhor resultado alcançado foi com o algoritmo SVM em todas as representações, sendo que as representações SSD e RP tiveram desempenhos

próximos, com 71,33% e 71,06% respectivamente. Com o melhor desempenho, a representação MFCC obteve 75,71% de assertividade. A representação RH, assim como na base LMD, obteve os piores resultados, alcançando apenas 55,46%.

Tabela 42 – Resultados dos classificadores das representações do áudio da BRMD para Gênero. Em negrito, consta o melhor classificador para cada representação.

Classificadores	**SSD**	**RP**	**RH**	**MFCC**
Decision Tree	48,6 ± 0,008	43,4 ± 0,012	36,2 ± 0,002	49,2 ± 0,007
k-NN	56,4 ± 0,017	57,3 ± 0,013	48,2 ± 0,007	59,0 ± 0,008
Gaussian NB	35,2 ± 0,010	40,9 ± 0,011	32,6 ± 0,007	43,4 ± 0,018
SVM	**71,3 ± 0,009**	**71,0 ± 0,006**	**55,4 ± 0,014**	**75,7 ± 0,009**
Logistic Regression	66,7 ± 0,008	60,69± 0,006	45,6 ± 0,004	69,5 ± 0,007

5.4.2 Contribuição das Representações do Áudio

Analisando a base de dados BRMD, demonstrada na Tabela 43 a média ponderada *f-measure*, nota-se que a representação MFCC é a que melhor classificou todas as categorias, com exceção da "bossa Nova", que obteve desempenho superior na representação RP. A representação RH teve o pior aproveitamento médio.

Tabela 43 – Resultado das representações do áudio por categoria para BRMD.

	SSD	**RP**	**RH**	**MFCC**
Samba	59,8%	60,0%	35,8%	68,8%
Sertanejo	70,6%	64,2%	39,2%	73,2%
Rap	**73,8%**	74,0%	55,2%	**79,8%**
Forró	67,0%	67,6%	31,6%	68,2%
Bossa Nova	76,2%	**79,8%**	**58,8%**	78,4%
Axé	72,8%	68,2%	46,0%	76,4%
Samba	47,2%	48,0%	28,4%	54,6%

5.4.3 Combinando as Representações do Áudio

Nestes experimentos fora realizada a fusão das características do áudio, criando um único vetor representativo da amostra através da técnica de fusão *early fusion*. Os classificadores foram otimizados com parâmetros independentes e treinados com a validação cruzada de 4 *folds*. Observa-se que os resultados apresentados na Tabela 44, referentes à média ponderada do desempenho dos classificadores indicam uma pequena melhora quando combinadas às representações do áudio, com o melhor resultado alcançando 79,70% para o algoritmo SVM.

5.4.4 Seleção dinâmica de Classificadores com as Representações do Áudio

Os resultados apresentados na Tabela 45 são referentes a realização da fusão das características através do método *early fusion* entre as representações do áudio, criando um

Tabela 44 – Resultado das representações do áudio da BRMD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	51,33 +/- 0,015
k-NN	61,32 +/- 0,013
Gaussian NB	43,50 +/- 0,007
SVM	**79,70 +/- 0,010**
Logistic Regression	70,97 +/- 0,005

único vetor representativo e com um conjunto de 30 classificadores heterogênico compostos pelos algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, onde após a validação cruzada em 3 folds, utilizando um *fold* para validação dos algoritmos de seleção dinâmica, para esse conjunto inicial de classificadores obteve-se um oráculo de 98,21% (limite superior de taxa de reconhecimento).

Tabela 45 – Resultados da seleção dinâmica utilizando as representações do áudio com a BRMD. Em negrito, é apresentado o melhor resultado para cada categoria.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Samba	67,0%	65,0%	67,0%	55,0%	63,0%	71,0%	35,0%
Sertanejo	84,0%	71,0%	78,0%	67,0%	73,0%	77,0%	65,0%
Rap	**86,0%**	**81,0%**	**86,0%**	77,0%	**79,0%**	**83,0%**	68,0%
Forró	72,0%	67,0%	68,0%	60,0%	63,0%	70,0%	36,0%
Bossa Nova	85,0%	79,0%	83,0%	**78,0%**	75,0%	83,0%	**69,0%**
Axé	76,0%	70,0%	74,0%	63,0%	69,0%	74,0%	62,0%
F-Measure	**80%**	**75%**	**78%**	**70%**	**72%**	**78%**	**60%**
Acurácia	**80,25%**	**74,67%**	**78,40%**	**69,40%**	**73,21%**	**74,67%**	**62,01%**

O algoritmo KNORA-U alcançou o melhor resultado, com 80% da *f-measure* e 89,25% de acurácia. Observa-se que o algoritmo só não teve bom desempenho, com a categoria "samba", alcançando 67%, enquanto o algoritmo *A-Posteriori* foi o classificador com melhor desempenho nesta categoria, com 71%.

Os resultados obtidos nas outras representações, “rap” e “bossa nova”, são bastante expressivos, visto que são as categorias com a menor representatividade na base de dados.

5.4.5 Classificação Utilizando Representações Visuais

Nos experimentos aqui discutidos foram utilizadas as imagens, através da extração das características por $LBP_{8,2}$ e da extração global de características para rotulação de gênero utilizando a BRMD, divididas em 4 *folds* para a validação cruzada. Conforme podemos observar a média ponderada *f-measure* na Tabela 46, os resultados foram divididos pela variação de espectrograma, que são gerados com o limite inferior definido em -50 dBFS, -70 dBFS, -90 dBFS, -110 dBFS e -130 dBFS (o limite superior é sempre 0 dBFS).

Tabela 46 – Resultado dos classificadores por cada representação visual na BRMD com gênero. Em negrito, consta o melhor classificador para cada representação.

Classificadores	**LBP-50**	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Decision Tree	49,4 ± 0,007	53,1 ± 0,011	54,1 ± 0,015	52,52 ± 0,007	53,48 ± 0,013
k-NN	55,0 ± 0,009	59,0 ± 0,015	61,3 ± 0,013	59,1 ± 0,007	60,4 ± 0,004
Gaussian NB	28,5 ± 0,012	44,0 ± 0,006	47,0 ± 0,014	47,7 ± 0,011	48,2 ± 0,018
SVM	**72,8 ± 0,005**	**73,5 ± 0,012**	**73,9 ± 0,008**	**72,7 ± 0,003**	**72,5 ± 0,007**
Logistic Regression	70,37± 0,005	71,2 ± 0,010	71,2 ± 0,016	71,2 ± 0,003	71,2 ± 0,007

O desempenho dos classificadores, através da média ponderada que as representações com amplitude -70 e -90 dBFS obtiveram, foram os melhores resultados com algoritmo SVM, com uma precisão de 73,54% e 73,97% respectivamente. Outro algoritmo, que merece destaque é o *Logistic Regression* com resultado de 71,22%, variando somente o desvio padrão, para as amplitudes -70, -90 e -130 dBFS.

5.4.6 Contribuição das Representações Visuais

Os resultados dos classificadores treinados através da validação cruzada em 4 *folds*, com as representações visuais da base de dados BRMD, são apresentados na Tabela 47. Os valores apresentados em cada célula representam a média ponderada do desempenho do melhor classificador por amplitude.

Tabela 47 – Resultado das representações das imagens por categoria para BRMD.

	LBP-50	**LBP-70**	**LBP-90**	**LBP-110**	**LBP-130**
Samba	65,2%	64,4%	52,0%	63,8%	64,4%
Sertanejo	67,8%	64,6%	49,2%	62,0%	62,8%
Rap	75,8%	**79,4%**	**64,2%**	**79,0%**	**78,0%**
Forró	69,6%	67,2%	55,0%	66,2%	67,6%
Bossa Nova	**78,4%**	78,4%	63,0%	77,6%	76,6%
Axé	70,6%	72,0%	60,2%	72,4%	72,6%
Samba	51,8%	51,6%	39,4%	51,4%	51,4%

A amplitude de -50 dBFS é a que apresenta maior precisão na maior parte das categorias, sendo seu melhor resultado neste aspecto para Bossa Nova, com 78,40%. Observa-se também, que as amplitudes de -70, -110 e -130 dBFS contribuem positivamente para a catalogação de “rap” e “axé”, ficando com a maior assertividade, 79,40%, com amplitude de -70 dBFS para “rap”.

5.4.7 Combinando as Representações Visuais

Nesta sessão são apresentados os resultados da fusão das representações visuais da base de dados BRMD, ou seja, utilizando a técnica de *early fusion* entre os diferentes níveis de LBP, tal qual na Seção 5.4.5. Na Tabela 48 é demonstrada a média ponderada do desempenho dos algoritmos.

Tabela 48 – Resultado das representações visuais da BRMD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	54,89 +/ 0,019
k-NN	61,39 +/- 0,006
Gaussian NB	49,56 +/- 0,018
SVM	75,56 +/- 0,014
Logistic Regression	**76,41 +/- 0,008**

Observa-se que somente os algoritmos SVM e *Logistic Regression* apresentaram uma pequena melhora no desempenho com a combinação das representações.

5.4.8 Seleção dinâmica de Classificadores com as Representações Visuais

Com as características das representações visuais extraídas utilizando $LBP_{8,2}$ e extração global de características, variando as amplitudes entre -50, -70, -90, -110 e -130 dBFS na criação daquelas que representam as canções categorizadas por gênero da BRMD, foi criando um conjunto de classificadores heterogêneo, com algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, previamente treinados realizando a validação cruzada de 3 vezes e aplicando os algoritmos de seleção dinâmica de classificadores em um *fold* específico para validação.

Na Tabela 49 são demonstrados os resultados para esse conjunto inicial de classificadores, obtendo um oráculo de 87,00% (limite superior de taxa de reconhecimento).

Observa-se que são bem próximos ao máximo do desempenho da seleção dinâmica sugerido pelo oráculo.

Tabela 49 – Resultados da seleção dinâmica utilizando as representações visuais da BRMD.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Samba	71,0%	68,0%	69,0%	54,0%	59,0%	69,0%	43,0%
Sertanejo	76,0%	63,0%	76,0%	55,0%	55,0%	64,0%	57,0%
Rap	81,0%	77,0%	81,0%	71,0%	77,0%	80,0%	**74,0%**
Forró	64,0%	59,0%	62,0%	51,0%	53,0%	60,0%	48,0%
Bossa Nova	**86,0%**	**82,0%**	**83,0%**	**73,0%**	**79,0%**	**80,0%**	72,0%
Axé	76,0%	72,0%	76,0%	69,0%	72,0%	72,0%	65,0%
F-Measure	**78%**	**74%**	**77%**	**66%**	**70%**	**74%**	**64%**
Acurácia	**78,57%**	**74,18%**	**77,27%**	**65,90%**	**71,91%**	**76,62%**	**66,39%**

O melhor resultado atingido foi com o algoritmo KNORA-U, com 78% de média ponderada e 78,57% de acurácia. Os resultados foram próximos ao atingido utilizando somente as características do áudio. Os maiores resultados por categoria são da "bossa nova" e do "rap", que são categorias que não possuem muitas representações e que apresentam bom desempenho com o algoritmo KNORA-U, com 86,0% e 81,0% respectivamente.

5.4.9 Classificação Utilizando Representações das Letras

Nos experimentos utilizando as letras da base de dados BRMD com rotulação para gênero e com classificadores com parâmetros de otimização, os resultados são melhores do que os vistos na rotulação de emoção, o que ocorre devido a uma maior representatividade por categoria. Nos experimentos foi utilizado somente o dicionário português para remoção de *stopwords*, visto que a base é exclusivamente de letras brasileiras. Os resultados estão apresentados na Tabela 50, e foram obtidos após a validação cruzada em 4 *folds*, cada célula com representação média ponderada *f-measure* e desvio médio de cada algoritmo em uma determinada representação.

Tabela 50 – Resultado dos classificadores por cada representação das letras da BRMD por gênero. Em negrito consta o melhor classificador para cada representação.

Classificadores	**Stemm**	**2-gram**	**3-gram**	**4-gram**
Decision Tree	47,72 ± 0,016	36,48 ± 0,013	**33,09 ± 0,008**	**30,81 ± 0,006**
k-NN	43,83 ± 0,014	37,76 ± 0,048	28,91 ± 0,032	29,00 ± 0,031
Gaussian NB	46,46 ± 0,020	33,05 ± 0,008	24,28 ± 0,013	14,45 ± 0,010
SVM	**65,49 ± 0,011**	26,54 ± 0,004	26,54 ± 0,004	26,54 ± 0,004
Logistic Regression	60,32 ± 0,010	**40,33 ± 0,007**	32,93 ± 0,013	30,73 ± 0,009

O melhor desempenho com a técnica de *stemming* é com algoritmo SVM, obtendo 65,49% de assertividade, os demais algoritmos não tiveram tão bom desempenho. Utilizando palavras com dimensionalidades reduzidas com a técnica de *n-grams*, é evidenciada a diminuição da precisão quando aumentado o valor de N. O algoritmo *Logistic Regression* teve seu melhor resultado com a representação de 2-grams com precisão de 40,33%. No geral os algoritmos não tiveram um bom desempenho com as representações das letras.

5.4.10 Contribuição das Representações das Letras

Na Tabela 51 são demonstrados os resultados por categoria, onde cada célula representa a média ponderada. Podemos verificar que *stemming* apresenta melhor desempenho na grande maioria das categorias, com melhor resultado para "forró", com 80,2% de assertividade, somente "axé" apresentou melhor desempenho com a representação n-gram no caso com 2-grams, resultando 59,8%.

5.4.11 Combinando as Representações das Letras

Nesta seção são apresentados os resultados dos classificadores quando combinadas através do método de fusão *early fusion*, as características extraídas com o método de *stemming* com aquelas extraídas utilizando o método n-gram, relacionando os três níveis bigrams (para n=2), trigrams (para n=3) e quadgrams (para n=4). Na Tabela 52 cada célula representa a média ponderada do desempenho dos algoritmos de classificação após a validação cruzada em 4 *folds*.

Tabela 51 – Contribuição das letras da BRMD por categoria.

	Stemm	**2-gram**	**3-gram**	**4-gram**
Samba	60,0%	33,2%	28,0%	17,8%
Sertanejo	58,2%	8,8%	21,8%	7,8%
Rap	66,2%	43,6%	35,2%	22,2%
Forró	**80,2%**	**59,8%**	16,4%	8,0%
Bossa Nova	67,6%	44,4%	**40,4%**	**43,0%**
Axé	24,8%	30,2%	23,6%	7,2%
Samba	47,4%	27,2%	22,6%	14,0%

Tabela 52 – Resultado das representações das letras da BRMD combinadas com método *early fusion*.

Classificadores	**Desempenho**
Decision Tree	34,14 +/- 0,023
k-NN	29,62 +/- 0,018
Gaussian NB	38,47 +/- 0,024
SVM	**47,97 +/- 0,025**
Logistic Regression	45,18 +/- 0,027

De maneira geral, os algoritmos apresentam um desempenho inferior àqueles alcançados pelas representações individuais das letras. No caso, devido a grande diferença das representações, a realização do passo de fusão não trás benefícios para a classificação. O algoritmo SVM é o que tem melhor desempenho, com 47,9%.

5.4.12 Seleção dinâmica de Classificadores com as Representações das Letras

Os experimentos desta seção foram realizados utilizando as representações das letras da BRMD, ou seja, as características extraídas de *n-grams* e *stemming* e realizando a fusão através do método *early fusion* entres as diferentes características e o TF/IDF das representações, criando um conjunto de classificadores heterogêneos, com os algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, sendo 5 representações de cada algoritmo treinadas com a validação cruzada, com 3 *folds*, utilizando um *fold* de validação para os algoritmos de seleção dinâmica.

A Tabela 53 demonstra o desempenho dos algoritmos de seleção dinâmica, com oráculo de 85,68% (limite superior de taxa de reconhecimento). O algoritmo KNORA-E apresenta o melhor resultado, utilizando a média ponderada *f-measure* de 26% e acurácia de 32,47%. Os resultados obtidos são bem ruins se comparados ao que não utiliza a seleção dinâmica de classificadores e abaixo da máxima sugerida pelo oráculo. Portanto, a representação da letra pouco contribuiu na identificação dos gêneros. Nota-se também que todos os conjuntos de classificadores falharam na classificação do gênero “sertanejo”.

Tabela 53 – Resultados da seleção dinâmica utilizando as representações de letras da BRMD.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Samba	20,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Sertanejo	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Rap	36,0%	36,0%	37,0%	30,0%	34,0%	39,0%	34,0%
Forro	**46,0%**	**56,0%**	**54,0%**	30,0%	**42,0%**	**51,0%**	**42,0%**
Bossa Nova	39,0%	39,0%	38,0%	**38,0%**	41,0%	41,0%	39,0%
Axé	0,0%	0,0%	0,0%	0,0%	3,0%	0,0%	7,0%
F-Measure	**23%**	**26%**	**25%**	**21%**	**25%**	**26%**	**25%**
Acurácia	**22,64%**	**32,47%**	**31,83%**	**27,99%**	**31,83%**	**30,55%**	**30,55%**

5.4.13 Seleção dinâmica de Classificadores com Todas as Representação

Os resultados presentes na Tabela 54 são dos algoritmos de seleção dinâmica com oráculo para esse conjunto inicial de classificadores de 99,48% (limite superior de taxa de reconhecimento). Os algoritmos de seleção dinâmica utilizaram um conjunto de classificadores heterogêneos com os algoritmos k-NN, *Decision tree*, SVM, *Logistic Regression* e *Gaussian Mixture Models*, sendo 5 representações de cada algoritmo e todos treinados com a validação cruzada com 3 *folds*, utilizando um *fold* de validação para os algoritmos de seleção dinâmica. O algoritmo que apresenta melhor resultado é o KNORA-U com *f-measure* de 89% e acurácia de 88,96%. Analisando as categorias, a "sertanejo" destaca-se por possuir maior precisão, com 93,0% de *f-measure*, com algoritmo DES-P. As categorias "rap" e "axé" são melhores classificadas pelo KNORA-U.

Tabela 54 – Resultados da seleção dinâmica utilizando todas as representações com BRMD.

	KNORA-U	**KNORA-E**	**DES-P**	**OLA**	**MCB**	**A-Priori**	**A-Posteriori**
Samba	84,0%	78,0%	84,0%	65,0%	72,0%	78,0%	71,0%
Sertanejo	92,0%	79,0%	**93,0%**	64,0%	71,0%	82,0%	67,0%
Rap	**91,0%**	**85,0%**	90,0%	**79,0%**	83,0%	87,0%	69,0%
Forró	89,0%	83,0%	89,0%	67,0%	75,0%	85,0%	39,0%
Bossa Nova	89,0%	84,0%	**90,0%**	78,0%	77,0%	**86,0%**	71,0%
Axé	88,0%	84,0%	87,0%	74,0%	**84,0%**	82,0%	**72,0%**
F-Measure	**89%**	**83%**	**88%**	**73%**	**78%**	**84%**	**67%**
Acurácia	**88,96%**	**82,75%**	**88,44%**	**73,62%**	**75,34%**	**83,62%**	**66,55%**

5.4.14 Considerações

Os resultados exibidos na Tabela 55 com a base de dados BMRD para a rotulação de gênero, demonstram que utilizando a seleção dinâmica de classificadores há uma melhora nos resultados, isso quando utilizadas todas as representações. As representações individuais, não apresentaram melhora adicional ao passo da criação do *pool* de classificadores, ou seja, com a utilização do método de seleção dinâmica. Utilizando somente o algoritmo SVM foi possível obter resultados bem próximos aos melhores resultados utilizando seleção dinâmica, quando analisadas as representações individuais. Os resultados foram ligeiramente melhores que os apresentados por (PEREIRA; SILLA, 2017).

Tabela 55 – Comparativo de todos os resultados das representações da base de dados BRMD combinadas.

	Representações			
Classificadores	**Audio**	**Visual**	**Letra**	**Todas**
Decision Tree	51.33%	54.89%	34.14%	42.6%
k-NN	61.32%	61.39%	29.62%	49.3%
Gaussian NB	43.50%	49.56%	38.47%	39.7%
SVM	**79.70%**	75.56%	**47.97%**	**68.3%**
Logistic Regression	70.97%	76.41%	45.18%	59.3%
KNORA-U	**80,4%**	**78,0%**	**48,9%**	**89,0%**
KNORA-E	74,7%	74,0%	42,6%	83,0%
DES-P	78,4%	77,0%	48,1%	**88,0%**
OLA	69,5%	66,0%	41,8%	73,0%
MCB	73,2%	70,0%	43,0%	78,0%
A-Priori	74,7%	74,0%	38,2%	84,0%
A-Posteriori	62,0%	64,0%	30,7%	67,0%

6 Conclusão

Este trabalho se desenvolveu em torno da exploração de diferentes representações, como áudios, imagens e letras, utilizando a seleção dinâmica de classificadores com o propósito de classificar canções com diferentes rótulos (gêneros e sentimentos) em bases de músicas brasileiras e latinas.

Como representações do áudio foram extraídas as características acústicas, de timbre e de ritmo. Para representação visual foi utilizado o descritor $LBP_{8,2}$ e a extração global de características. Para a extração das características das letras das canções utilizou-se dos métodos *stemming* e *n-grams*, criando vetores através do cálculo da relevância da palavra na música.

Sob uma perspectiva geral, os resultados dos experimentos utilizando a rotulação em gênero atingiram um melhor desempenho, principalmente devido ao grande desbalanceamento das categorias de sentimento se comparada uma em relação à outra, o que dificultou o treinamento dos classificadores.

O uso de diferentes representações mostrará que em um determinado domínio uma representação se sobressai a outra, como por exemplo, a representação visual, que obteve resultados interessantes para as bases de dados rotuladas em gênero, o que ocorreu devido à estrutura harmônica das canções que se refletem nas imagens geradas, de onde, posteriormente, foram extraídas as características. O mesmo comportamento não é observado com a classificação de emoção.

Na classificação por emoção observa-se que as representações de áudio e letras apresentam um melhor desempenho, como observado com a base de dados LMMD. Além disso, estas representações foram melhores que as visuais. Destaca-se, neste caso, que a representação de áudio obteve 54,9% com algoritmo SVM e 66,0% com KNORA-U de média ponderada.

A base de dados BRMD-MOOD é a que apresenta maior dificuldade para treinamento dos classificadores, falhando ao rotular diversas emoções devido ao grande desbalanceamento das amostras de treinamento. No entanto é possível identificar uma melhoria em seu resultado quando utilizado o classificador SVM com todas as representações combinadas ao método *early fusion*.

Para os algoritmos de seleção dinâmica de grupos de classificadores, destacam-se os algoritmos KNORA's e DES-P, que apresentam os melhores resultados quando são utilizadas todas as representações combinadas. Embora para a base de dados LMD, o melhor resultado tenha sido obtido através dos algoritmos SVM, com a representação

visual.

De maneira geral, há um aumento da taxa de reconhecimento pela utilização de várias representações, ou seja, trabalhando com várias características e criando conjuntos distintos para a utilização de técnicas de seleção dinâmica de grupo de classificadores. O ponto negativo disto é que se torna necessário adicionar etapas para a criação das diferentes representações, além de criar uma grande quantidade de classificadores para a obtenção de uma melhoria não muito expressiva.

Para finalizar, para um trabalho futuro, se propõe a investigação da utilização de técnicas multi classes, trabalhando com gênero e sentimento para a representação da mesma classe, comparando esta técnica com as de aprendizagem profunda.

Referências

ALEMAYEHU, Nega; WILLETT, Peter. The effectiveness of stemming for information retrieval in amharic. *Program*, MCB UP Ltd, v. 37, n. 4, p. 254–259, 2003. Citado 2 vezes nas páginas 39 e 58.

AN, Yunjing; SUN, Shutao; WANG, Shujuan. Naive bayes classifiers for music emotion classification based on lyrics. *16th International Conference on Computer and Information Science (ICIS)*, n. 1, p. 635–638, 2017. Citado 3 vezes nas páginas 18, 22 e 23.

BAGWELL, Chris; KLAUER, U. Sox-sound exchange. *Online Website*, 2010. Citado na página 58.

BARSALOU, Lawrence W; SANTOS, Ava; SIMMONS, W Kyle; WILSON, Christine D. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, p. 245–283, 2008. Citado na página 17.

BRITTO, Alceu S.; SABOURIN, Robert; OLIVEIRA, Luiz E.S. Dynamic selection of classifiers - A comprehensive review. *Pattern Recognition*, v. 47, n. 11, p. 3665–3680, 2014. Citado 4 vezes nas páginas 44, 45, 47 e 61.

CAVALIN, Paulo R.; SABOURIN, Robert; SUEN, Ching Y. Dynamic selection approaches for multiple classifier systems. *Neural Computing and Applications*, v. 22, n. 3-4, p. 673–688, 2013. Citado na página 44.

CAVNAR, William B; TRENKLE, John M et al. N-gram-based text categorization. v. 161175, p. 161–175, 1994. Citado 2 vezes nas páginas 39 e 40.

CHEN, Yu An; WANG, Ju Chiang; YANG, Yi Hsuan; CHEN, Homer. Linear regression-based adaptation of music emotion recognition models for personalization. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, 2014. p. 2149–2153. Disponível em: <http://dblp.uni-trier.de/db/conf/icassp/icassp2014.html#ChenWYC14>. Citado 2 vezes nas páginas 22 e 23.

COSTA, Carlos Humberto Lopes; VALLE, Jaime Dalla; KOERICH, Alessandro L. Automatic classification of audio data. In: IEEE. *Systems, Man and Cybernetics, 2004 IEEE International Conference on*. 2004. v. 1, p. 562–567. Disponível em: <http://dblp.uni-trier.de/db/conf/smc/smc2004-1.html#CostaVK04>. Citado na página 35.

COSTA, Y.; OLIVEIRA, L.; KOERICH, A.; GOUYON, F. Music genre recognition based on visual features with dynamic ensemble of classifiers selection. *2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP)*, p. 55–58, 2013. Disponível em: <http://ieeexplore.ieee.org/document/6623448/>. Citado 7 vezes nas páginas 26, 28, 36, 56, 57, 87 e 88.

COSTA, Yandre MG; OLIVEIRA, Luiz S; SILLA JR, Carlos N. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied soft*

computing, Elsevier B.V., v. 52, p. 28–38, 2017. Citado 7 vezes nas páginas 27, 28, 37, 56, 58, 75 e 89.

CRUZ, Rafael M.O.; SABOURIN, Robert; CAVALCANTI, George D.C. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, Elsevier B.V., v. 41, p. 195–216, 2018. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2017.09.010>. Citado 4 vezes nas páginas 44, 48, 49 e 51.

DIDACI, Luca; GIACINTO, Giorgio; ROLI, Fabio; MARCIALIS, Gian Luca. A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern Recognition*, Elsevier, v. 38, n. 11, p. 2188–2191, 2005. Citado na página 49.

DIETTERICH, Thomas G. Ensemble methods in machine learning. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 1–15, 2000. Citado na página 44.

EEROLA, Tuomas; LARTILLOT, Olivier; TOIVIAINEN, Petri. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. ISMIR 2009 - 10th International Conference on Music Information Retrieval, p. 621–626, 2009. Disponível em: <http://dblp.uni-trier.de/db/conf/ismir/ismir2009.html#EerolaLT09>. Citado na página 21.

EL-KHAIR, Ibrahim Abu. Effects of stop words elimination for arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences*, v. 4, n. 3, p. 119–133, 2006. Citado na página 59.

FIX, Evelyn; JR, Joseph L Hodges. *Discriminatory analysis-nonparametric discrimination: consistency properties*. [S.l.], 1951. Citado na página 41.

FLEXER, Arthur. A closer look on artist filters for musical genre classification. *World*, v. 19, n. 122, p. 16–7, 2007. Citado 3 vezes nas páginas 24, 28 e 56.

GEORGE, Joe; SHAMIR, Lior. Computer analysis of similarities between albums in popular music. *Pattern Recognition Letters*, Elsevier, v. 45, p. 78–84, 2014. Citado na página 56.

GJERDINGEN, Robert O; PERROTT, David. Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, Taylor & Francis, v. 37, n. 2, p. 93–100, 2008. Citado 2 vezes nas páginas 16 e 24.

GUNES, Veyis; MENARD, Michel; LOONIS, Pierre; PETIT-RENAUD, Simon. Combination, cooperation and selection of classifiers: A state of the art. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 17, n. 08, p. 1303–1324, 2003. Citado na página 46.

HANSEN, L.K.; SALAMON, Peter. Neural network ensembles. v. 12, p. 993 – 1001, 11 1990. Citado na página 44.

HEVNER, Kate. Experimental studies of the elements of expression in music. *American journal of Psychology*, v. 48, n. 2, p. 246–268, 1936. Citado 3 vezes nas páginas 7, 30 e 31.

HIRJI, Karim K. Discovering data mining: From concept to implementation. ACM, v. 1, n. 1, p. 44–45, jun 1999. Disponível em: <http://doi.acm.org/10.1145/846170.846181>. Citado na página 53.

HO, Tin Kam; HULL, Jonathan J.; SRIHARI, Sargur N. Decision combination in multiple classifier systems. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 16, n. 1, p. 66–75, 1994. Citado na página 45.

HU, Xiao; DOWNIE, J. Stephen; LAURIER, Cyril; BAY, Mert; EHMANN, Andreas F. The 2007 mirex audio mood classification task: Lessons learned. ISMIR 2008 - 9th International Conference on Music Information Retrieval, p. 462–467, Dez 2008. Disponível em: <http://dblp.uni-trier.de/db/conf/ismir/ismir2008.html#HuDLBE08>. Citado na página 34.

HUANG, Yea S.; SUEN, Ching Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Transactions on Pattern Analysis & Machine Intelligence, n. 1, p. 90–94, 1995. Citado na página 48.

HUQ, Arefin; BELLO, Juan Pablo; ROWE, Robert. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, v. 39, n. 3, p. 227–244, 2010. Citado 2 vezes nas páginas 22 e 23.

Jain, A. K.; Duin, R. P. W.; Jianchang Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 4–37, Jan 2000. Citado na página 59.

JUSLIN, Patrik N; SLOBODA, John A. *Music and emotion: Theory and research.* Oxford University Press, 2001. Disponível em: <https://books.google.com.br/books?id=t8j5pduTkboC>. Citado 2 vezes nas páginas 29 e 30.

JUSLIN, Patrik N; VÄSTFJÄLL, Daniel. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, Cambridge University Press, v. 31, n. 5, p. 559–575, 2008. Citado 2 vezes nas páginas 19 e 30.

KARKANIS, S; GALOUSI, K; MAROULIS, D. Classification of endoscopic images based on texture spectrum. *ACAI99, Workshop on Machine Learning in Medical Applications*, p. 63–69, 1999. Citado na página 37.

KIM, Youngmoo E; SCHMIDT, Erik M; MIGNECO, Raymond; MORTON, Brandon G; RICHARDSON, Patrick; SCOTT, Jeffrey; SPECK, Jacquelin A; TURNBULL, Douglas. Music emotion recognition: A state of the art review. *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010*, p. 255–266, 01 2010. Citado 2 vezes nas páginas 30 e 32.

KITTLER, Josef; HATEF, Mohamad; DUIN, Robert PW; MATAS, Jiri. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 20, n. 3, p. 226–239, 1998. Citado 2 vezes nas páginas 45 e 59.

KO, Albert HR; SABOURIN, Robert; BRITTO JR, Alceu Souza. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, Elsevier, v. 41, n. 5, p. 1718–1731, 2008. Citado 6 vezes nas páginas 44, 45, 47, 49, 50 e 51.

KOERICH, Alessandro Lameiras; POITEVIN, Cleverson. Combination of homogeneous classifiers for musical genre classification. In: . 2005 IEEE International Conference on Systems, Man and Cybernetics, 2005. v. 1, p. 554–559. Disponível em: <http://dblp.uni-trier.de/db/conf/smc/smc2005.html#KoerichP05>. Citado 3 vezes nas páginas 18, 24 e 28.

KONEČNI, Vladimir J. Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts*, Educational Publishing Foundation, v. 2, n. 2, p. 115, 2008. Citado na página 29.

KUNCHEVA, Ludmila I; RODRIGUEZ, Juan J. Classifier ensembles with a random linear oracle. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 19, n. 4, p. 500–508, 2007. Citado 2 vezes nas páginas 40 e 50.

KUNCHEVA, Ludmila I.; WHITAKER, Christopher J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, v. 51, n. 2, p. 181–207, May 2003. Disponível em: <https://doi.org/10.1023/A:1022859003006>. Citado na página 59.

LAURIER, Cyril; GRIVOLLA, Jens; HERRERA, Perfecto. Multimodal music mood classification using audio and lyrics. Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008, p. 688–693, 2008. Disponível em: <http://www.grivolla.net/articles/icmla2008.pdf>. Citado 2 vezes nas páginas 21 e 23.

LENC, Ladislav; KRÁL, Pavel. Automatically detected feature positions for lbp based face recognition. In: SPRINGER. *IFIP International Conference on Artificial Intelligence Applications and Innovations*. 2014. v. 436, p. 246–255. Disponível em: <http://dblp.uni-trier.de/db/conf/ifip12/aiai2014.html#LencK14>. Citado na página 38.

LI, Tao; OGIHARA, Mitsunori; LI, Qi. A comparative study on content-based music genre classification. ACM, New York, NY, USA, p. 282–289, 2003. Disponível em: <http://doi.acm.org/10.1145/860435.860487>. Citado 2 vezes nas páginas 24 e 28.

LIDY, Thomas; PÖLZLBAUER, Georg; RAUBER, Andreas. Sound re-synthesis from rhythm pattern features - audible insight into a music feature extraction process. Barcelona, Spain, p. 93–96, September 5–9 2005. Citado na página 58.

LIDY, Thomas; RAUBER, Andreas. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. ISMIR 2005 - 6th International Conference on Music Information Retrieval, p. 34–41, 2005. Disponível em: <http://dblp.uni-trier.de/db/conf/ismir/ismir2005.html#LidyR05>. Citado na página 57.

Lin, Y.; Chung, C.; Chen, H. H. Playlist-based tag propagation for improving music auto-tagging. p. 2270–2274, Sep. 2018. Disponível em: <http://dblp.uni-trier.de/db/conf/eusipco/eusipco2018.html#LinCC18>. Citado 2 vezes nas páginas 16 e 29.

LOPES, M.; GOUYON, F.; KOERICH, A. L.; OLIVEIRA, L. E. S. Selection of training instances for music genre classification. p. 4569–4572, Aug 2010. Citado 2 vezes nas páginas 25 e 28.

LU, Lie; LIU, Dan; ZHANG, Hong-Jiang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, IEEE, v. 14, n. 1, p. 5–18, 2006. Citado 3 vezes nas páginas 21, 23 e 34.

LU, Yi. Knowledge integration in a multiple classifier system. *Applied Intelligence*, Springer, v. 6, n. 2, p. 75–86, 1996. Citado na página 45.

MACIÀ, Núria; ORRIOLS-PUIG, Albert; BERNADÓ-MANSILLA, Ester. In search of targeted-complexity problems. p. 1055–1062, 2010. Citado na página 35.

MAENPAA, Topi; PIETIKAINEN, Matti. Texture analysis with local binary patterns. *Handbook of Pattern Recognition and Computer Vision*, 01 2005. Citado na página 38.

MAYER, John D; SALOVEY, Peter; CARUSO, David R; SITARENIOS, Gill. Emotional intelligence as a standard intelligence. American Psychological Association, 2001. Citado na página 29.

MITCHELL, Tom M. Machine learning and data mining. *Communications of the ACM*, v. 42, n. 11, 1999. Citado na página 41.

NANNI, Loris; COSTA, Yandre M.G.; LUMINI, Alessandra; KIM, Moo Young; BAEK, Seung Ryul. Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, Pergamon Press, Inc., v. 45, n. C, p. 108–117, mar. 2016. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2015.09.018>. Citado 4 vezes nas páginas 26, 28, 36 e 87.

NIAZ, Usman; MERIALDO, Bernard. Fusion methods for multi-modal indexing of web data. p. 1–4, 2013. Citado na página 62.

OJALA, Timo; PIETIKÄINEN, Matti; HARWOOD, David. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, Elsevier, v. 29, n. 1, p. 51–59, 1996. Citado na página 37.

ORAMAS, Sergio; NIETO, Oriol; BARBIERI, Francesco; SERRA, Xavier. Multi-Label Music Genre Classification from Audio, Text, and Images Using Deep Features. *Proceedings of the International Society for Music Information Retrieval {(ISMIR)} Conference*, p. 23–30, 2017. Disponível em: <http://arxiv.org/abs/1707.04916>. Citado 2 vezes nas páginas 27 e 28.

ORIO, Nicola. Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, v. 1, n. 1, p. 1–90, 2006. Disponível em: <http://dx.doi.org/10.1561/1500000002>. Citado na página 16.

PACHECO, André. *K vizinhos mais próximos - KNN*. 2017. <http://computacaointeligente.com.br/algoritmos/k-vizinhos-mais-proximos>. Accessed: 2018-09-02. Citado na página 42.

PACHET, François; CAZALY, Daniel. A Taxonomy of Musical Genres. *Content-Based Multimedia Information Access Conference*, n. April, p. 1238–1245, 2000. Citado na página 23.

PAMPALK, Elias; FLEXER, Arthur; WIDMER, Gerhard et al. Improvements of audio-based music similarity and genre classificaton. v. 5, p. 634–637, 2005. Citado na página 24.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 43.

PEEL, David; MCLACHLAN, Geoffrey J. Robust mixture modelling using the t distribution. *Statistics and computing*, Springer, v. 10, n. 4, p. 339–348, 2000. Citado na página 42.

PEREIRA, Rodolfo Miranda; SILLA, Carlos. Using simplified chords sequences to classify songs genres. IEEE Computer Society, p. 1446–1451, 07 2017. Disponível em: <http://dblp.uni-trier.de/db/conf/icmcs/icme2017.html#PereiraS17>. Citado 2 vezes nas páginas 55 e 96.

PETER, Paula C. Emotional intelligence. *Wiley International Encyclopedia of Marketing*, Wiley Online Library, 2010. Citado na página 29.

PONTI JR; P, Moacir. Combining classifiers: from the creation of ensembles to the decision fusion. p. 1–10, 2011. Citado na página 45.

PORTER, Karen G; FEIG, Yvette S. The use of dapi for identifying and counting aquatic microflora 1. *Limnology and oceanography*, Wiley Online Library, v. 25, n. 5, p. 943–948, 1980. Citado na página 40.

PORTILLA, Javier; SIMONCELLI, Eero P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, Springer, v. 40, n. 1, p. 49–70, 2000. Citado na página 35.

POUYANFAR, Samira; SAMETI, Hossein. Music emotion recognition using two level classification. *Proc. Intelligent Systems*, p. 1–6, 2014. Citado 2 vezes nas páginas 22 e 23.

RANAWANA, Romesh; PALADE, Vasile. Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, IOS Press, v. 3, n. 1, p. 35–61, 2006. Citado na página 46.

REYNOLDS, Douglas A; QUATIERI, Thomas F; DUNN, Robert B. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, Elsevier, v. 10, n. 1-3, p. 19–41, 2000. Citado na página 42.

ROMANO, Luiz Eduardo; ADAMI, André Gustavo. Reconhecimento automático de gêneros musicais utilizando classificadores baseados em múltiplas características. *Revista Brasileira de Computação Aplicada*, v. 7, n. 1, p. 85–99, 2015. Disponível em: <http://www.upf.br/seer/index.php/rbca/article/view/4281>. Citado 2 vezes nas páginas 26 e 28.

RUSSELL, James A. A circumplex model of affect. *Journal of personality and social psychology*, American Psychological Association, v. 39, n. 6, p. 1161, 1980. Citado 3 vezes nas páginas 31, 32 e 33.

SALTON, G. The smart system. *Retrieval Results and Future Plans*, 1971. Citado na página 39.

SALZBERG, Steven L. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, Springer, v. 16, n. 3, p. 235–240, 1994. Citado na página 41.

SANTOS, Carolina L. dos; SILLA, Carlos N. The latin music mood database. *EURASIP Journal on Audio, Speech, and Music Processing*, v. 2015, n. 1, p. 23, Aug 2015. Disponível em: <https://doi.org/10.1186/s13636-015-0065-6>. Citado na página 54.

SCARINGELLA, Nicolas; ZOIA, Giorgio; MLYNEK, Daniel. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, IEEE, v. 23, n. 2, p. 133–141, 2006. Citado na página 35.

SCHöLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. Advances in kernel methods. MIT Press, Cambridge, MA, USA, p. 1–15, 1999. Disponível em: <http://dl.acm.org/citation.cfm?id=299094.299095>. Citado na página 43.

SCHWARTZ, William R; SIQUEIRA, Fernando R de; PEDRINI, Helio. Evaluation of feature descriptors for texture classification. *Journal of Electronic Imaging*, International Society for Optics and Photonics, v. 21, n. 2, p. 023016, 2012. Citado na página 37.

SEYERLEHNER, Klaus; SCHEDL, Markus; POHLE, Tim; KNEES, Peter. Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX*, v. 2010, 2010. Citado 2 vezes nas páginas 25 e 28.

SHAO, Xi; CHENG, Zhiyong; KANKANHALLI, Mohan S. Music auto-tagging based on the unified latent semantic modeling. *Multimedia Tools and Applications*, v. 78, n. 1, p. 161–176, Jan 2019. Disponível em: <https://doi.org/10.1007/s11042-018-5632-2>. Citado na página 18.

SILLA JR., Carlos N.; KAESTNER, Celso A.A; KOERICH, Alessandro L. A machine learning approach to automatic music genre classification. *Journal of the Brazilian Computer Society*, SciELO Brasil, v. 14, n. 3, p. 7–18, 2008. Citado 2 vezes nas páginas 25 e 28.

SILLA JR., Carlos N.; KAESTNER, Celso A. A.; KOERICH, Alessandro L. The Latin Music Database. *Musica*, p. 451–456, 2008. Disponível em: <http://kar.kent.ac.uk/24000/>. Citado 2 vezes nas páginas 25 e 54.

SILLA JR., Carlos N.; KAESTNER, Celso A. A.; KOERICH, Alessandro L. Automatic Music Genre Classification Using Ensemble of Classifiers. 2010. Citado 3 vezes nas páginas 25, 56 e 57.

SRIDHARAN, Anusha; MOH, Melody; MOH, Teng-Sheng. Similarity estimation for classical indian music. IEEE, p. 814–819, 12 2018. Disponível em: <http://dblp.uni-trier.de/db/conf/icmla/icmla2018.html#SridharanMM18>. Citado na página 16.

SUMMERS, Cameron; TRONEL, Greg; CRAMER, Jason; VARTAKAVI, Aneesh; POPP, Phillip. Gnmid14: A collection of 110 million global music identification matches. In: *SIGIR*. New York, NY, USA: ACM, 2016. (SIGIR '16), p. 693–696. Disponível em: <http://doi.acm.org/10.1145/2911451.2914679>. Citado na página 55.

TAVARES, Leonardo G; LOPES, Heitor S; LIMA, Carlos R Erig. Estudo comparativo de métodos de aprendizado de máquina na detecção de regiões promotoras de genes de escherichia coli. *Anais do I Simpósio Brasileiro de Inteligência Computacional*, p. 8–11, 2007. Citado na página 39.

TELLEGEN, Auke; WATSON, David; CLARK, Lee Anna. On the dimensional and hierarchical structure of affect. *Psychological science*, SAGE Publications Sage CA: Los Angeles, CA, v. 10, n. 4, p. 297–303, 1999. Citado 2 vezes nas páginas 30 e 31.

THAYER, Robert E. *The Biopsychology of Mood and Arousal*. Oxford University Press, USA, 1989. 195-222 p. Disponível em: <http://www.loc.gov/catdir/enhancements/fy0637/89002914-t.html>. Citado 4 vezes nas páginas 21, 22, 31 e 32.

TONG, Haoyue; ZHANG, Min; SOLEIMANINEJADIAN, Pouneh; ZHANG, Qianfan; WU, Kailu; LIU, Yiqun; MA, Shaoping. Music mood classification based on lifelog. Springer International Publishing, Cham, p. 55–66, 2018. Citado na página 33.

TSOUMAKAS, Grigorios; PARTALAS, Ioannis; VLAHAVAS, Ioannis. A taxonomy and short review of ensemble selection. *ECAI 2008, Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications*, 2008. Citado na página 45.

TUMER, Kagan; GHOSH, Joydeep. Error correlation and error reduction in ensemble classifiers. *Connection science*, Taylor & Francis, v. 8, n. 3-4, p. 385–404, 1996. Citado na página 45.

TZANETAKIS, George; COOK, Perry. Automatic Musical Genre Classification Of Audio Signals. *IEEE transactions on Speech and Audio Processing*, v. 10, n. 5, p. 292–302, 2002. Citado 5 vezes nas páginas 23, 24, 28, 35 e 57.

VAPNIK, V. N. An overview of statistical learning theory. *Trans. Neur. Netw.*, IEEE Press, Piscataway, NJ, USA, v. 10, n. 5, p. 988–999, set. 1999. ISSN 1045-9227. Disponível em: <https://doi.org/10.1109/72.788640>. Citado na página 43.

VRIESMANN, Leila M.; BRITTO, Alceu S.; OLIVEIRA, Luiz S.; KOERICH, Alessandro L.; SABOURIN, Robert. Combining overall and local class accuracies in an oracle-based method for dynamic ensemble selection. *Proceedings of the International Joint Conference on Neural Networks*, v. 2015-Septe, 2015. Citado 2 vezes nas páginas 26 e 28.

VRIESMANN, Leila Maria; JR, Alceu de Souza Britto; OLIVEIRA, Luiz Eduardo Soares De; SABOURIN, Robert; KO, Albert Houng-Ren. Improving a dynamic ensemble selection method based on oracle information. *International Journal of Innovative Computing and Applications 17*, Inderscience Publishers Ltd, v. 4, n. 3-4, p. 184–200, 2012. Citado na página 17.

WANG, Li; HE, Dong-Chen. Texture classification using texture spectrum. *Pattern Recognition*, Elsevier, v. 23, n. 8, p. 905–910, 1990. Citado na página 37.

WATSON, David; TELLEGEN, Auke. Toward a consensual structure of mood. *Psychological bulletin*, American Psychological Association, v. 98, n. 2, p. 219, 1985. Citado 3 vezes nas páginas 7, 30 e 32.

WATSON, David; WIESE, David; VAIDYA, Jatin; TELLEGEN, Auke. The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of personality and social psychology*, American Psychological Association, v. 76, n. 5, p. 820, 1999. Citado 2 vezes nas páginas 31 e 32.

WITTEN, I.H.; FRANK, E.; HALL, M.A. Practical machine learning tools and techniques. Morgan Kaufmann, p. 629, 2011. Citado na página 53.

WITTEN, Ian H; PAYNTER, Gordon W; FRANK, Eibe; GUTWIN, Carl; NEVILL-MANNING, Craig G. Kea: Practical automated keyphrase extraction. IGI Global, p. 129–152, 2005. Citado na página 52.

WOLOSZYNSKI, Tomasz; KURZYNSKI, Marek; PODSIADLO, Pawel; STACHOWIAK, Gwidon W. A measure of competence based on random classification for dynamic ensemble selection. *Information Fusion*, Elsevier, v. 13, n. 3, p. 207–213, 2012. Citado na página 51.

WOODS, K.; KEGELMEYER, W. P.; BOWYER, K. Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, v. 19, n. 4, p. 405–410, abr. 1997. Disponível em: <http://dx.doi.org/10.1109/34.588027>. Citado 3 vezes nas páginas 47, 48 e 53.

WU, Ming-Ju; CHEN, Zhi-Sheng; JANG, Jyh-Shing Roger; REN, Jia-Min; LI, Yi-Hsung; LU, Chun-Hung. Combining visual and acoustic features for music genre classification. v. 2, p. 124–129, 2011. Citado 2 vezes nas páginas 26 e 28.

WU, Xiaosheng; SUN, Junding. An effective texture spectrum descriptor. v. 2, p. 361–364, 2009. Citado 2 vezes nas páginas 37 e 38.

ZHANG, Tong; KUO, C-CJ. Hierarchical classification of audio data for archiving and retrieving. v. 6, p. 3001–3004, 1999. Citado na página 35.

ZHANG, W1; NUKI, G; MOSKOWITZ, RW; ABRAMSON, S; ALTMAN, Roy D; ARDEN, NK; BIERMA-ZEINSTRA, S; BRANDT, KD; CROFT, P; DOHERTY, M et al. Oarsi recommendations for the management of hip and knee osteoarthritis: part iii: Changes in evidence following systematic cumulative update of research published through january 2009. *Osteoarthritis and cartilage*, Elsevier, v. 18, n. 4, p. 476–499, 2010. Citado na página 29.