UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MARCOS VINICIUS GOLOM

CLASSIFICAÇÃO AUTOMÁTICA DE CATEGORIA DE CENAS ACÚSTICAS COM REDUÇÃO DE DIMENSIONALIDADE BASEADA EM PROJEÇÕES LINEARES E SELEÇÃO DE INSTÂNCIAS

CAMPO MOURÃO

2021

MARCOS VINICIUS GOLOM

CLASSIFICAÇÃO AUTOMÁTICA DE CATEGORIA DE CENAS ACÚSTICAS COM REDUÇÃO DE DIMENSIONALIDADE BASEADA EM PROJEÇÕES LINEARES E SELEÇÃO DE INSTÂNCIAS

Dimensionality Reduction Based on Linear Projections and Instance Selection for Automatic Acoustic Scene Category Classification

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Juliano Henrique Foleis

Coorientador: Prof. Dr. Diego Bertolini Gonçalves

CAMPO MOURÃO

2021

MARCOS VINICIUS GOLOM

CLASSIFICAÇÃO AUTOMÁTICA DE CATEGORIA DE CENAS ACÚSTICAS COM REDUÇÃO DE DIMENSIONALIDADE BASEADA EM PROJEÇÕES LINEARES E SELEÇÃO DE INSTÂNCIAS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 30/novembro/2021

Juliano Henrique Foleiss
Doutorado
Universidade Tecnológica Federal do Paraná

Rodrigo Hübner
Doutorado
Universidade Tecnológica Federal do Paraná

Frank Helbert Borsato
Doutorado
Universidade Tecnológica Federal do Paraná

CAMPO MOURÃO

2021

AGRADECIMENTOS

A minha mãe e meus avós por todos esses anos de apoio.

E a todos que me deram apoio para superar meus obstáculos.

Ao meu orientador Prof. Dr. Juliano Herique Foleiss e ao meu coorientador Prof. Dr. Diego Bertolini Gonçalves pela oportunidade e pelo apoio.

A Terezinha Martins Davidoff pelo suporte imprescindível.

Ao Paulo Adriano Davidoff pelos conselhos ao decorrer da formação.

RESUMO

A classificação de cenas acústicas consiste na tarefa de reconhecer o ambiente que um áudio foi gravado a partir do seu sinal do áudio. A literatura mostra que métodos baseados em aprendizagem profunda apresentam ótimos resultados. Entretanto, esses métodos possuem custo computacional elevado, inviabilizando seu uso em dispositivos limitados.
Para este trabalho foi utilizado o conjunto de dados disponibilizado pela DCASE para o desafio *Low-Complexity Acoustic Scene Classification* de 2020. Este desafio propõe um limite de 500kb para armazenar os parâmetros do modelo. Um *baseline* também foi publicado pelos organizadores do desafio e consiste em uma rede neural que obteve uma acurácia de 87,30%, com 450kb de parâmetros.
Para lidar com esse limite foram utilizadas duas abordagens que visam reduzir o tamanho dos modelos para a tarefa de classificação de categorias de cenas acústicas.
O objetivo principal deste trabalho foi realizar reduções de dimensionalidade focadas na otimização do custo na fase de predição. Na primeira abordagem avaliada, apenas as técnicas de redução de dimensionalidade PCA, RP, NMF foram usadas com os classificadores KNN (K-Nearest Neighbors) e SVM (Support Vector Machine). O melhor resultado que respeitou o limite de 500KB foi obtido com a técnica NMF com 16 componentes e o classificador SVM. O F1-Score alcançado foi de 86,64%, com 390KB. Na segunda abordagem avaliada, optou-se também por usar a técnica de seleção de instâncias, uma vez que os classificadores usados armazenam o modelo a partir de uma amostragem do conjunto de treinamento. Nesta abordagem, foram combinadas as técnicas de redução de dimensionalidade juntamente com a técnica de seleção de instâncias baseada em K-Means (KMEANSC). O melhor resultado que respeitou o limite de 500KB foi obtido com a técnica PCA com 32 componentes usando 15% do conjunto de treinamento, com o classificador SVM. O F1-score alcançado nesta situação foi de 87,84%, com 410KB.

Palavras-chave: classificação de cenas acústicas; aprendizagem de máquina; seleção de instâncias; redução de dimensionalidade; projeção linear.

ABSTRACT

Acoustic scene classification consists of the task of recognizing the environment that an audio was recorded from its audio signal. The literature shows that deep learning-based methods provide excellent results. However, these methods have a high computational cost, making them unfeasible to use on limited devices.
For this work we used the dataset provided by DCASE for the 2020 challenge *Low-Complexity Acoustic Scene Classification*. This challenge proposes a 500kb limit for storing model parameters. A *baseline* has also been published by the organizers of the challenge and consists of a neural network that achieved an accuracy of 87.30% with 450kb of parameters.
To deal with this limit, two approaches were used that aim to reduce the size of the models for the task of classifying acoustic scene categories.
The main goal of this work was to perform dimensionality reductions focused on cost optimization in the prediction phase. In the first approach evaluated, only the dimensionality reduction techniques PCA, RP, NMF were used with the KNN (K-Nearest Neighbors) and SVM (Support Vector Machine) classifiers. The best result that respected the 500KB limit was obtained with the NMF technique with 16 components and the SVM classifier. The F1-Score achieved was 86.64% with 390KB. In the second approach evaluated, it was also chosen to use the instance selection technique, since the classifiers used store the model from a sampling of the training set. In this approach, the dimensionality reduction techniques were combined together with the K-Means based instance selection technique (KMEANSC). The best result respecting the 500KB threshold was obtained with the PCA technique with 32 components using 15% of the training set, with the SVM classifier. The F1-score achieved in this situation was 87.84% with 410KB.

Keywords: acoustic scene classification; machine learning; instance selection; dimensionality reduction; linear projection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

CCA	Classificação de Cenas Acústicas
CNN	*Convolutional Neural Network*
DCASE	*Detection and Classification of Acoustic Scenes and Events*
DFT	*Discrete Fourier Transform*
KNN	*K-Nearest Neighbors*
LBP	*Local Binary Pattern*
LogMBE	*Long Mel Band Energies*
LPQ	*Local Phase Quantization*
MARSYAS	*Music Analysis, Retrieval and Synthesis for Audio Signals*
MFCC	*Mel-Frequency Cepstral Coefficients*
MFDWC	*Mel-Frequency Discrete Wavelet Coefficient*
MMD	*Maximum Mean Discrepancy*
NMF	*Non-Negative Matrix Factorization*
PCA	*Principal Component Analysis*
RFR	*Receptive-Field Regularized*
RFR CNN	*Receptive-Field Regularized Convolutional Neural Network*
RP	*Random Projection*
STFT	*Short Time Fourier Transform*
SVM	*Support Vector Machine*
SWD	*Sliced Wasserstein Distance*
TF	Transformada de Fourier

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da era da informação, rapidamente aumentamos o volume de dados multimídia como fotos, áudios ou vídeos provenientes da criação de novos conteúdos ou da restauração e digitalização de arquivos antigos. Por essa razão, rotular ou categorizar bases de dados contendo milhões de arquivos de áudio, imagem ou vídeo é uma tarefa difícil. Pode-se empregar pessoas para rotular e/ou categorizar estes arquivos manualmente, pois essa tarefa em muitos cenários é trivial, porém o custo é alto para fazer isso. Contudo, devido à velocidade do crescimento dos dados e também a subjetividade da percepção humana, percebe-se a necessidade de automatizar essa tarefa. (HERSHEY *et al.*, 2017).

A classificação de cenas acústicas consiste na tarefa de, a partir de um determinado áudio, reconhecer em qual ambiente foi gravado. As aplicações da Classificação de Cenas Acústicas (CCA) estão em diversas áreas, como em monitoramento (GOETZE *et al.*, 2012), navegação de robôs autônomos (CHU *et al.*, 2006), *wearables* inteligentes (XU; LI; LEE, 2008), entre outras.

Uma outra tarefa da classificação de cenas acústicas é detectar o tipo de ambiente, ao contrário de detectar exatamente o ambiente da gravação. Por exemplo, na base do DCASE 2020 Task 1B (HEITTOLA; MESAROS; VIRTANEN, 2020), os áudios devem ser categorizados entre *Indoor*, *Outdoor* e *Transportation* (em algum lugar fechado, ao ar livre, ou em algum meio de transporte respectivamente).

Trabalhos anteriores que obtiveram bons resultados utilizaram redes neurais artificiais Koutini *et al.* (2020), Hu *et al.* (2020), McDonnell (2020). Entretanto, na prática, as aplicações podem ser implementadas em dispositivos com restrições de memória ou de processamento. Portanto, nem sempre é viável usar redes neurais nestes dispositivos.

Neste trabalho abordamos o problema de classificação de cenas acústicas usando métodos tradicionais de aprendizado de máquina. Especificamente, avaliamos a viabilidade da combinação de características artesanais (*handcrafted*) e de métodos de redução de dimensionalidade.

1.1 Justificativa

Na última década tem sido observado um grande aumento na utilização de técnicas de Aprendizagem Profunda para obtenção de melhores taxas de classificação. Contudo, muitos trabalhos desconsideram o custo computacional que esses métodos exigem.

Quando as aplicações são restritas a utilizar dispositivos com poder de computação e armazenamento limitados, o uso aprendizagem profunda diretamente no dispositivo é inviável. Neste caso, para viabilizar a aplicação, muitas vezes o dispositivo é responsável apenas por capturar o sinal. Este sinal deve ser transmitido para outro dispositivo local ou via internet para que então o modelo de aprendizagem profundo possa fazer as predições (HEITTOLA; MESAROS;

VIRTANEN, 2020). Em muitos casos, este tipo de solução multi-dispositivo não é adequada, inviabilizando o uso de aprendizagem profunda para resolver o problema.

Ao reduzir o custo computacional e o tamanho do modelo gerado de um sistema de classificação de cenas acústicas, torna-se possível o seu uso em dispositivos com baixo poder computacional. Assim, tornando possível que um número maior de dispositivos possa ter acesso a soluções para o problema de Classificação de Cenas Acústicas (CCA).

1.2 Estrutura do trabalho

No Capítulo 2 são apresentadas técnicas de processamento de áudio, descritores de áudio, descritores de imagens e técnicas de redução de dimensionalidade. No Capítulo 3 são apresentados trabalhos relacionados. No Capítulo 5 são apresentados os resultados dos experimentos realizados.

2 REFERENCIAL TEÓRICO

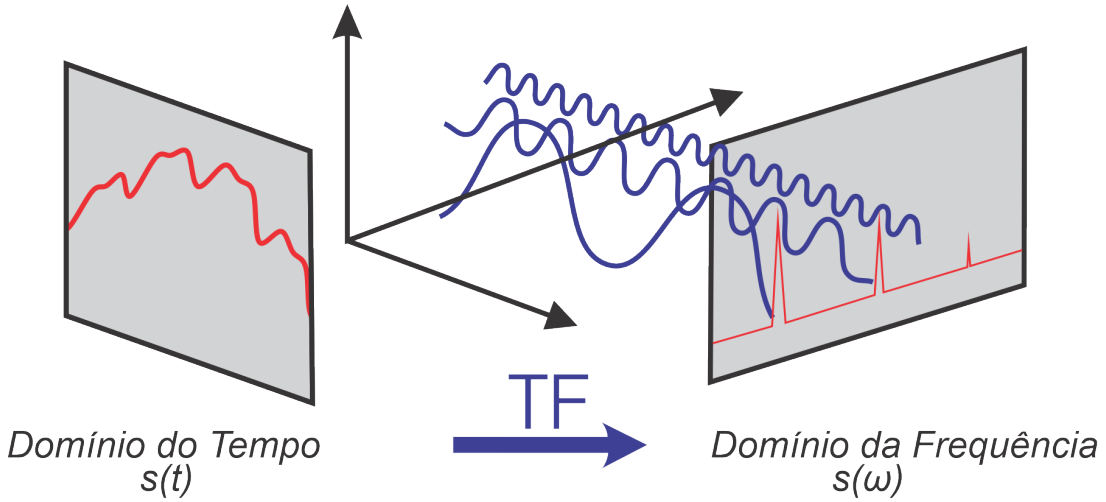
Neste capitulo são apresentados alguns conceitos importantes relacionados a descritores de áudio e redução de dimensionalidade. Primeiramente são apresentados conceitos ligados a descritores de áudio e extração de características e em seguida conceitos a respeito de redução de dimensionalidade.

2.1 Transformada de Fourier

A Transformada de Fourier (TF) é uma transformada integral mais importante em processamento de sinais. Como tal, ela faz parte de um extenso conjunto de ferramentas analíticas usadas em aplicações em várias áreas do conhecimento tais como a Geofísica (SINHA *et al.*, 2005), Medicina (ERNST; ANDERSON, 1966) e Tecnologia de Alimentos (TAY *et al.*, 2002).

Segundo Bailey e Swarztrauber (1994) a TF é usada para decompor um sinal em um determinado intervalo de tempo em uma soma ponderada de senoides e cossenoides. O resultado da TF descreve as frequências fundamentais do sinal, dizemos que a TF transforma um sinal no domínio do tempo em um sinal no domínio da frequência. A Figura 1 mostra a TF.

Figura 1 – Transformada de Fourier



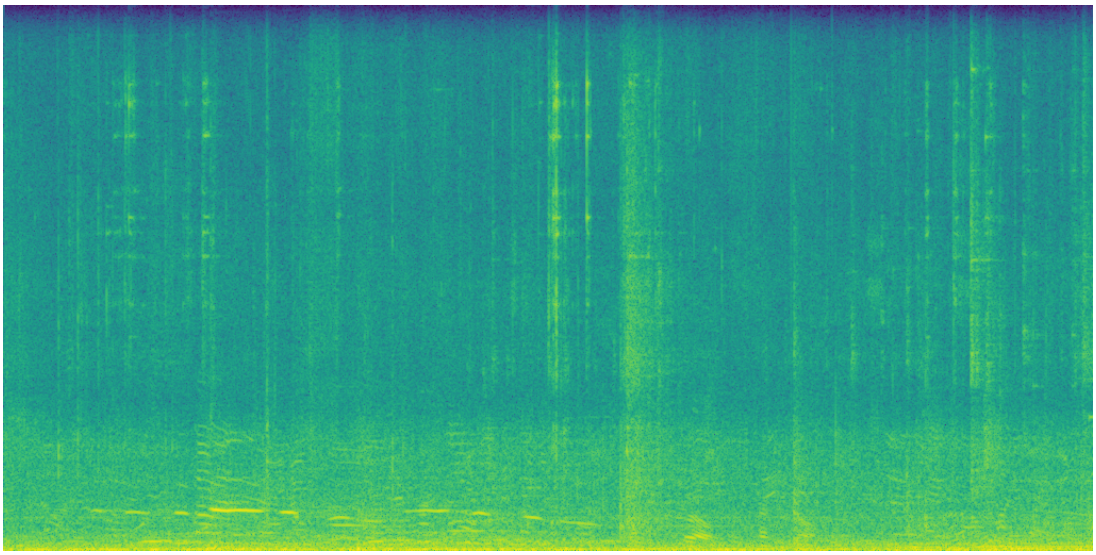
Fonte: Autoria própria (2022).

Helmholtz (2009) mostrou no Século XIX a relação entre as frequências audíveis e sua percepção. Resumidamente, Helmholtz mostrou que quanto mais alta a frequência, mais aguda é a percepção do som. Desta forma, a decomposição do sinal de áudio para o domínio da frequência fornece informações que não são diretamente observáveis no sinal do domínio do tempo.

A *Short Time Fourier Transform* (STFT) (Transformada de Fourier em Tempo Curto) é uma aplicação da TF que consiste em uma sequência de TF aplicadas em pequenas partes justapostas do sinal, denominadas de quadros (*frames*). Assim, a STFT realiza o mapeamento

de um sinal de áudio no domínio do tempo, para uma representação tempo $\times$ frequência. A STFT pode ser usada para gerar espectrogramas como pode ser visto na Figura 2, tal que cada valor representa a magnitude do sinal em dada frequência em cada instante de tempo. Esta representação possibilita acompanhar a evolução do espectro no tempo, descrevendo o áudio ao longo de toda sua duração.

Figura 2 – Espectrograma de um áudio



Fonte: Autoria própria (2022).

2.2 Descritores de Áudio

A STFT é considerada uma descrição de áudio de baixo-nível (TZANETAKIS; COOK, 2002), fornecendo apenas a evolução do espectro do áudio ao longo do tempo. No entanto, para classificar o áudio são normalmente utilizadas características que descrevem diferentes aspectos do sinal e de seu espectro. Estas características são desenvolvidas a partir do conhecimento especialista de cada problema de classificação de áudio.

2.2.1 *Mel-frequency Cepstral Coefficients*(MFCC)

A escala Mel procura codificar a sensibilidade humana sobre as frequências do sinal de áudio de uma forma não-linear, uma vez que o ouvido humano não percebe a variação de frequência linearmente (STEVENS; VOLKMANN; NEWMAN, 1937). Por exemplo, nas frequências entre 0Hz a 1000Hz a percepção é praticamente linear, enquanto acima dos 1000Hz a percepção se torna logarítmica.

MFCC é um descritor de áudio inspirado pela percepção humana aos sinais de áudio. Após o cálculo da STFT sobre um sinal de áudio, as bandas de frequência do espectro são combinadas em bandas de frequência perceptualmente relacionadas, usando a escala Mel (Davis;

Mermelstein, 1980). Por fim, o espectro é compactado usando a transformada discreta do cosseno (TDC).

2.2.2 Características MARSYAS

Em Tzanetakis e Cook (2002), são descritas algumas características computadas a partir da STFT do sinal de áudio. O conjunto dessas características é conhecido como *Music Analysis, Retrieval and Synthesis for Audio Signals* (MARSYAS). Embora essas características tenham sido inicialmente utilizadas no contexto de classificação automática de gêneros musicais, elas são amplamente utilizadas em outros problemas (FOLEIS; TAVARES, 2020).

Seja M o espectro de potência da STFT em dado tempo, então $M[f]$ indica a magnitude da frequência f. N é o número de bins de frequência no espectro. O conjunto MARSYAS é composto das características a seguir.

Spectral Centroid O brilho espectral é calculado pela seguinte equação:

$$C = \frac{\sum_{f=1}^{N} fM[f]}{\sum_{f=1}^{N} M[f]}$$

Spectral Rollof É uma medida da forma do espectro e é dada por R na equação a seguir:

$$\sum_{f=1}^{R} M[f] = 0.85 \sum_{f=1}^{N} M[f]$$

Spectral Flux É uma medida da variação espectral e é calculada por:

$$F = ||M[f] - M[f-1]||_2$$

Energy A medida da potência do sinal é calculada por:

$$E = \sum_{f=1}^{N} M[f]^2$$

Spectral Flatness É uma medida do ruído no sinal de áudio, calculada por:

$$L = \frac{\exp\left(\frac{1}{N}\sum_{f=1}^{N} \ln\left(M[f]\right)\right)}{\frac{1}{N}\sum_{f=1}^{N} M[f]}$$

Zero Crossing Rate É outra medida de ruído no sinal de áudio. Esta característica é calculada sobre o sinal no domínio do tempo $x[i], | \, i = \{1, 2, \ldots, T\}$.

$$Z = \frac{1}{2T} \sum_{t=1}^{T} |\text{sign}(x[t+1]) - \text{sign}(x[t])|$$

tal que sign$(k) = 1$ se $k \geq 0$, ou 0, caso contrário.

2.3 Descritores de Textura

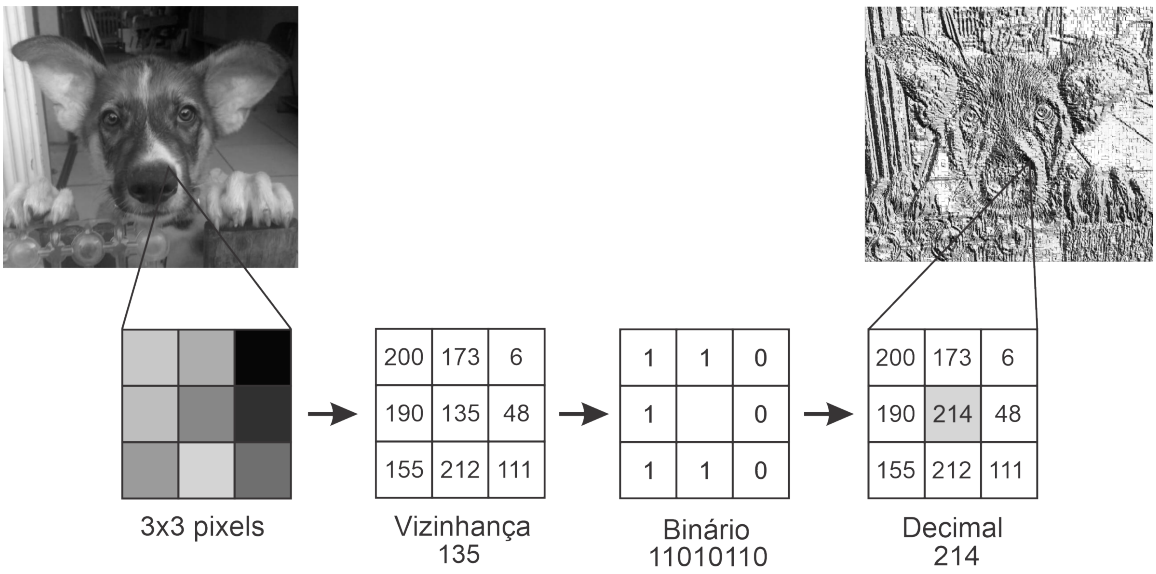
Áudios podem ser descritos por meio de imagens usando a STFT. Para transformar a STFT em imagem, um processo de normalização e quantização é utilizado para determinar o valor do *pixel* correspondente a cada posição da matriz STFT. Uma vez que a STFT é convertida em uma imagem, é possível classificar áudios por meio de descritores de textura (Costa *et al.*, 2011). A seguir são apresentados dois descritores de textura usados na literatura, o *Local Binary Pattern* e o *Local Phase Quantization*.

2.3.1 *Local Binary Pattern*

Local Binary Pattern (LBP) (Ojala; Pietikainen; Maenpaa, 2002) é um operador de textura que primeiramente foi descrito em Ojala, Pietikäinen e Harwood (1996) como medida auxiliar de contraste local de imagens. O LBP um operador que trabalha com imagens em escala de cinza e acentua o contraste local da imagem usando a vizinhança comumente 3×3 de cada *pixel* da imagem. Entretanto, essa vizinhança pode ser parametrizada.

A seguir a Figura 3 ilustra como o valor de LBP é calculado para cada *pixel* da imagem. Com a imagem em escala de cinza, cada *pixel* C é processado e um padrão LBP é obtido por pixel. Considere a vizinhança 3×3 de C. O valor de cada pixel vizinho de C é comparado com C. Se a intensidade do vizinho é maior ou igual a C, então a relação de vizinhança vale 1. Caso contrário, o valor da relação é 0. Como C possui 8 vizinhos, a relação de vizinhança entre C e todos seus vizinhos é representada por um *byte*, com um dígito para cada vizinho. Esse byte resultante é o valor do padrão LBP do pixel C.

Figura 3 – Método LBP



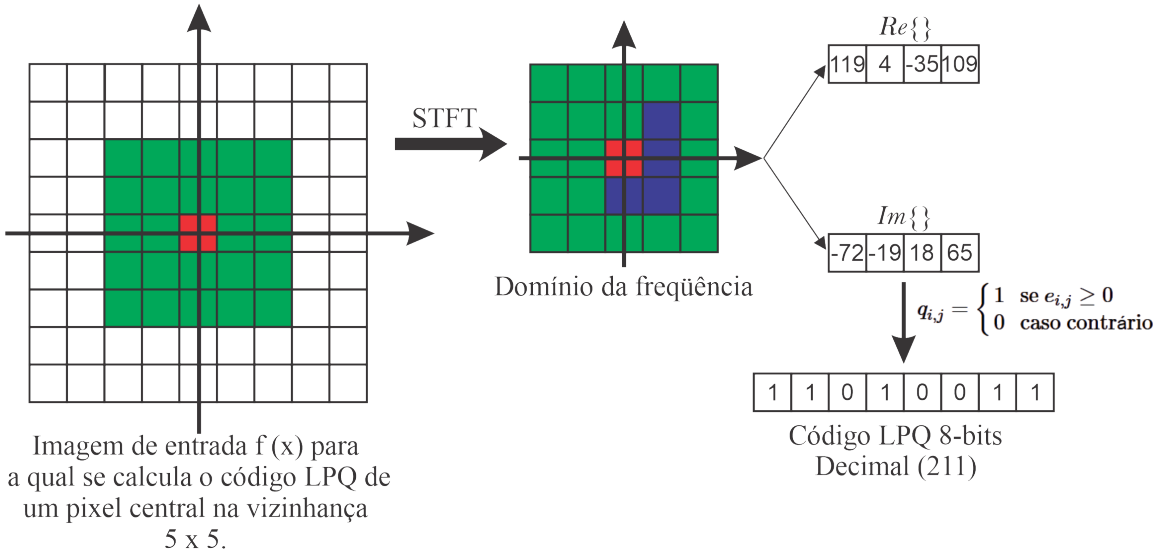
Fonte: Autoria própria (2022).

Após o cálculo do valor LBP para todos os pixeis, um histograma é calculado para representar a variedade de contraste na imagem. O histograma é de tamanho 2^p caso seja calculado de forma integral, sendo p o número de vizinhos. Para reduzir a dimensionalidade do histograma, Ojala, Pietikainen e Maenpaa (2002) propõe fazer o uso do conceito de uniformidade. O conceito de uniformidade é dado pelo número de transições entre uns e zeros contidos no código binário obtido pelo cálculo LBP. Considerando o código binário uma lista circular, caso o número de transições seja menor ou igual a dois, o código é considerado uniforme como o exemplo 00010000, caso contrário não é uniforme como no exemplo 00100100. Assim com o uso da uniformidade é possível fazer o uso apenas dos códigos uniformes, sendo somente possíveis apenas 58 valores, para o cálculo do histograma. Logo, o histograma do LBP é composto dos 58 valores uniformes possíveis e a última coluna do histograma é composta por todos os valores não uniformes que compõe a imagem, assim resultando em um vetor com 59 dimensões.

2.3.2 *Local Phase Quantization*

O descritor de textura *Local Phase Quantization* (LPQ) é um descritor baseado na propriedade de invariância ao borramento do espectro de fase de Fourier (OJANSIVU; HEIKKILä, 2008). Este descritor tem seu comportamento semelhante ao LBP, mas ao contrário de usar operações simples na vizinhança, este opera sobre *Discrete Fourier Transform* (DFT) 2D calculada sobre uma matriz retangular na vizinhança de cada *pixel* da imagem, para extrair a informação de fase local. A Figura 4 demonstra quais os passos necessários para construir o descritor LPQ.

Figura 4 – Método do LPQ



Fonte: Belahcene *et al.* (2016).

Para se obter o código LPQ é necessário calcular os coeficientes locais de Fourier, que são dados por quatro coeficientes complexos das frequências 2D: $u_1 = [a,0]^T$, $u_2 = [0,a]^T$, $u_3 = [a,a]^T$, $u_3 = [a,-a]^T$, no qual $a = 1/m$ e m é o tamanho da janela local.

Após a aplicação dos coeficientes ao resultado da STFT são gerados números complexos, onde a parte real é mantida no vetor $Re\{\}$ e a parte imaginaria em $Im\{\}$. As partes $Re\{\}$ e $Im\{\}$ devem ser quantizadas em forma binaria (ZHOU; YIN; ZHANG, 2013). Assim é possível gerar o código LPQ binário que deve ser transformado para decimal em seguida. Como o descritor tem como base a imagem na escala de cinza, os possíveis valores para cada código LPQ variam de 0 a 255. Logo um vetor de 256 posições corresponde ao histograma LPQ.

2.4 Redução de Dimensionalidade

A dimensionalidade de um vetor de características é definida pela quantidade de características que são usadas para representar um determinado padrão (GABRILOVICH; MARKOVITCH, 2004). A busca pela redução de dimensionalidade é motivada pela redução do conjunto de características original por um conjunto que contenha menos características. A redução de dimensionalidade é considerada eficaz se a perda de informação é minimizada durante o processo (MAATEN; POSTMA; HERIK, 2009). Assim, o classificador tem seu custo de processamento e espaço reduzidos, sem prejudicar de forma significativa a taxa de classificação (GUYON *et al.*, 2002). A redução de dimensionalidade também é uma ferramenta importante usada para evitar a maldição da dimensionalidade (ALTMAN; KRZYWINSKI, 2018), que afeta negativamente alguns classificadores que operam no espaço euclidiano, como o *K-Nearest Neighbors* (KNN).

As técnicas de redução de dimensionalidade podem ser categorizadas em dois conjuntos, a saber: projeção de características ou seleção de características (GUYON; ELISSEEFF, 2003). Na projeção de características, o objetivo é realizar uma transformação do espaço de características original para um espaço de menor dimensionalidade (MAATEN; POSTMA; HERIK, 2009). Estas transformações podem ser realizadas por meio de algoritmos baseados em decomposição de matrizes e outras técnicas numéricas de decomposição.

Os métodos baseados em seleção de características escolhem um subconjunto das características originais. Estas técnicas normalmente baseiam-se em eliminação de características descorrelacionadas com as classes de interesse, ou de características que sejam redundantes com as demais (GUYON; ELISSEEFF, 2003). Nesta Seção são apresentadas algumas técnicas de redução de dimensionalidade comumente utilizadas em problemas de classificação.

2.4.1 *Principal Component Analysis*

O *Principal Component Analysis* (PCA) (Análise de Componentes Principais) é uma técnica estatística para redução de dimensionalidade. Dado um conjunto de dados que podem estar correlacionados de N registros por p características, são realizadas combinações lineares entre as características originais do conjunto, seguidamente sendo agrupadas em eixos não correlacionados chamados eixos principais. Os eixos principais são ordenados em função da sua quantidade de variância da maior para a menor e a covariância entre cada par de eixos do conjunto é zero (RINGNÉR, 2008).

Para calcular os k melhores eixos o PCA, usa a distância euclidiana como medida de dissimilaridade. A dissimilaridade é calculada para todas as p características e sendo usada como medida de distância entre as N entradas do conjunto (MAATEN; POSTMA; HERIK, 2009).

Basta selecionar os k eixos principais com as maiores variâncias e descartar os eixos que tem uma variância menor assim mantendo no conjunto resultante somente as características com maior significância (HOWLEY *et al.*, 2006). Este é o conceito que permite ao PCA diminuir a dimensionalidade de dados e ainda manter uma baixa perda de informação.

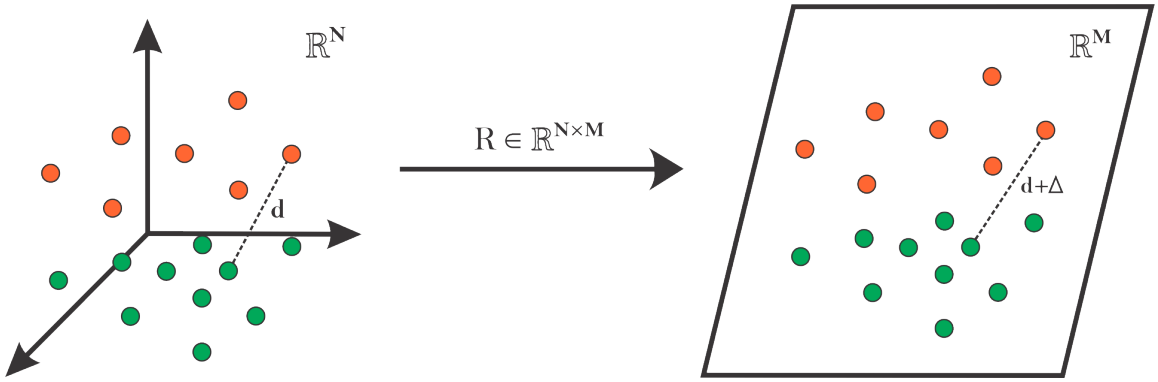
2.4.2 *Random Projection*

É afirmado em Johnson e Lindenstrauss (1984) que os pontos em um espaço vetorial de alta dimensionalidade podem ser projetados em um espaço de baixa dimensionalidade, preservando a topologia de distâncias entre os pontos como pode ser visto na Figura 5. Uma consequência disso é que a classificação pode ser realizada no espaço de menor dimensionalidade.

Random Projection (Projeção Aleatória) é uma técnica para redução da dimensionalidade para pontos no espaço euclidiano. Esta técnica realiza a projeção de uma matriz de

características $D \in \mathbb{R}^{K \times N}$ de um espaço N-dimensional para um espaço M-dimensional, tal que $M < N$, pela transformação linear DR, onde $R \in \mathbb{R}^{N \times M}$ (Candes; Wakin, 2008). R é uma matriz de números aleatórios amostrados de uma distribuição gaussiana padrão $\mathcal{N}(0,1)$. Desde que M não seja muito pequeno, a topologia de distâncias entre os pontos em R^M é preservada em relação a R^N. A escolha de M é normalmente realizada por meio de validação cruzada. Está técnica é usada por sua simplicidade, seu baixo custo computacional e por sua eficiência, quando respeitadas as condições necessárias (BINGHAM; MANNILA, 2001)..

Figura 5 – A técnica de Random Projection

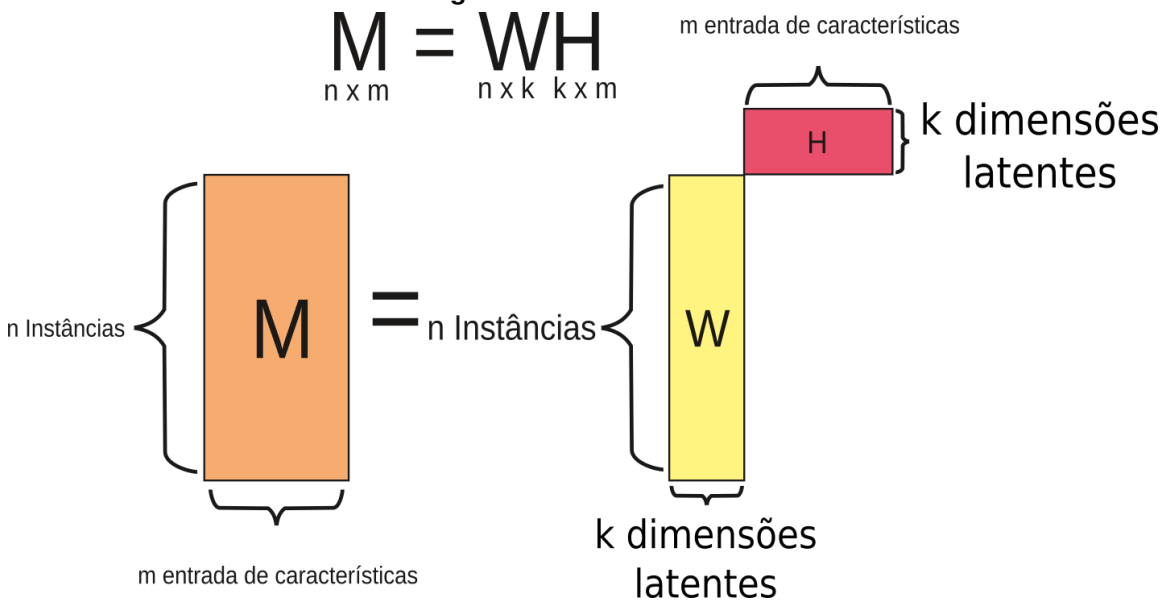


Fonte: Autoria própria (2022).

2.4.3 NMF

O NMF (LEE; SEUNG, 1999) é uma técnica de decomposição matricial que pode ser utilizada para redução de dimensionalidade. O método assume que todos os elementos da matriz de entrada sejam não-negativos e isso dá o nome ao método. Dada uma matriz $M \in \mathbb{R}^{n \times m}$, a fatoração *Non-Negative Matrix Factorization* (NMF) encontra duas matrizes $W \in \mathbb{R}^{n \times k}$ e $H \in \mathbb{R}^{k \times m}$, ambas com todos os elementos não-negativos, que minimiza a distancia entre X e WH. k é a dimensionalidade do espaço latente, que normalmente é escolhida por meio de validação cruzada como pode ser observado na Figura 6.

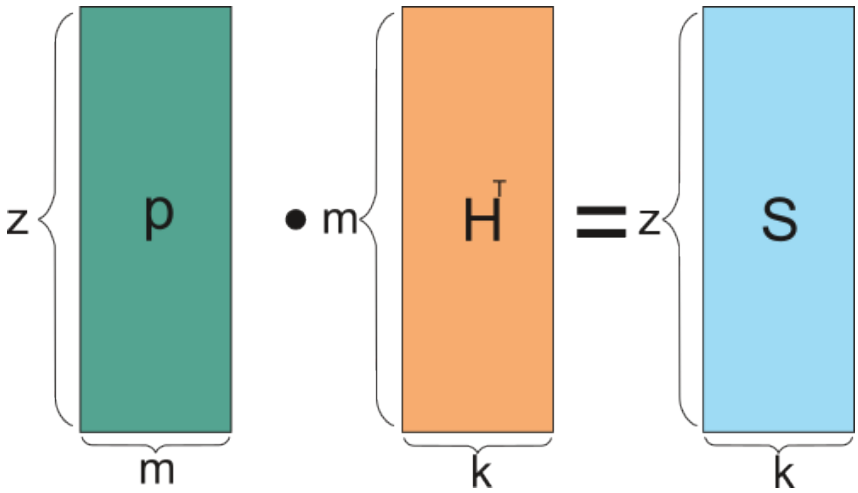
Figura 6 – A técnica NMF



Fonte: Müller (2018).

A redução de dimensionalidade é realizada pela transformação linear descrita por H^T, que projeta o vetor de características de m dimensões para k dimensões como mostrado na Figura 7.

Figura 7 – NMF redução de dimensionalidade



Fonte: Autoria própria (2022).

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados trabalhos sobre classificação de cenas acústicas utilizando modelos de baixa complexidade. Todos utilizaram o mesmo conjunto de dados fornecido pela *Detection and Classification of Acoustic Scenes and Events* (DCASE) (MESAROS; HEITTOLA; VIRTANEN, 2018).

3.1 CP-JKU Submissions To DCASE'20: Low-Complexity Cross-Device Acoustic scene Classification With RF-Regularized CNNs

Neste artigo foram apresentadas algumas técnicas, para o problema de classificação de cenas acústicas de baixa complexidade. Esse trabalho foi centrado em técnicas de redes neurais convolucionais.

Foram empregadas na classificação *Receptive-Field Regularized Convolutional Neural Network* (RFR CNN) que com desafios anteriores da DCASE, se provaram efetivas para classificação de cenas acústicas. Foram feitas melhorias na *Convolutional Neural Network* (CNN) usando *Sliced Wasserstein Distance* (SWD) e *Maximum Mean Discrepancy* (MMD) para controlar melhor a efetividade do *Receptive-Field Regularized* (RFR), assim a CNN se adaptaria melhor para diferentes tarefas.

Para a tarefa B também era necessário tratar da condição de limite do tamanho total do modelo sendo esse 500 kB. Para isso foram usadas técnicas e redução de dimensionalidade. Sendo elas a *Weight pruning*, ou poda de peso, *Layer decomposition*, ou decomposição de camadas e a redução de largura e profundidade da base da rede neural.

Esse artigo teve o resultado com taxa de acerto de 97,6% com o tamanho de 483,5 KB tendo um total de 247.562 parâmetros contra 3 milhões de parâmetros caso não fosse aplicado nenhuma das técnicas de redução de tamanho.

3.2 Exploring Compact Alternatives To Deep Learning In Task 1B

A abordagem apresentada por Patki (2020) foi usar técnicas de baixo custo computacional para a classificação de áudio. Foi usada uma camada de rede neural em paralelo com o *Support Vector Machine* (SVM) com o *Kernel* linear. Para a escolha das características foi usado um sistema proprietário que não foi mencionado qual o utilizado, para buscar e otimizar a escolha de quais conjuntos de características usar, com isso foram escolhidas as *Mel Spectral* como características e ao todo foram extraídas 500 delas e sendo codificadas em 16 bits.

Ao final dos experimentos a melhor taxa de acerto obtida foi de 88,5% com o tamanho de 26,3 KB, mas alguns experimentos apresentaram taxas abaixo de 88%, mas o tamanho que apresentavam chegavam a somente 8 KB. Foram alcançadas as taxas que não ultrapassaram

a *baseline* proposta pela DCASE, porém os experimentos conseguiram modelos realmente reduzidos.

3.3 Mel-Scaled Wavelet-Based Features For Sub-Task A And Texture Features for Sub-Task B Of DCASE 2020 Task1

A proposta de Waldekar, A e Saha (2020) é de usar o SVM para o problema de Classificação de Cenas Acústicas apresentado pela DCASE com as características *Long Mel Band Energies* (LogMBE), *Mel-Frequency Discrete Wavelet Coefficient* (MFDWC) e LBP.

Os áudios foram convertidos para o domínio da frequência para serem aplicadas as extrações. Foram extraídas as LogMBE que são parecidas com *Mel-Frequency Cepstral Coefficients* (MFCC). Após esse procedimento também ocorreu a extração das MFDWC aplicado nas LogMBE e usando também a LogMBE, foram extraídas as características de LBP e todas essas características foram usadas para treinar o classificador SVM que usava como configuração um *kernel* de intersecção.

A *baseline* usada pela DCASE (HEITTOLA; MESAROS; VIRTANEN, 2020), que é baseada em uma rede neural convolucional, teve uma acurácia de 87,3% com o tamanho de 450 KB e usa somente LogMBE para treinar e classificar o áudio. Já a abordagem proposta pelos autores usa a fusão de LogMBE+ MFDWC+ LBP para treinar um SVM e como resultado obtiveram uma taxa de acerto de 90,0% e um tamanho de 40KB.

O resultado obtido por esse trabalho por uma boa combinação de características e em conjunto com o classificador o SVM, obteve uma taxa de acerto que superou a apresentada pela *baseline* e um modelo com o tamanho bem reduzido chegando em 40KB.

3.4 Considerações Finais

Todos os trabalhos apresentados neste capítulo tiveram modelos bem reduzidos, porém o trabalho com a maior acurácia usa uma categoria de CNN como classificador. Contudo, os modelos gerados por classificadores CNN tendem a ser custosos, necessitando de poder computacional mais elevado, para ser usados na predição. Deste modo, sendo restrito o seu uso a ambientes computacionais com baixo poder de processamento ou sem acesso a *hardwares* aceleradores. Os trabalhos que usaram técnicas de baixo custo computacional, se comparadas às CNN, são os que fizeram o uso do SVM e tiveram uma boa taxa de acerto e ainda os modelos com tamanho reduzido.

4 METODOLOGIA

Neste capítulo são apresentadas algumas definições importantes relacionadas a metodologia utilizada nos experimentos. Primeiramente são apresentadas definições ligadas ao conjunto de dados, a *baseline*, descrição do sistema e deslocamento dos dados após a normalização.

4.1 Conjunto de Dados

O conjunto de dados escolhido para este trabalho foi o *TAU Urban Acoustic Scenes 2020 3Class* (Heittola, Toni; Mesaros, Annamaria; Virtanen, Tuomas, 2020) que é derivado do *TAU Urban Acoustic Scenes 2019* (HEITTOLA; MESAROS; VIRTANEN, 2019). Contudo, neste conjunto de dados suas dez classes foram agrupadas em três classes de cenas acústicas, sendo elas:

- Indoor – Aeroporto, *shopping center*, estação de metrô;
- Outdoor – Rua para pedestres, praça pública, rua com tráfego médio e parque urbano;
- Transportation – Transporte em ônibus, transporte em bonde, transporte em metrô.

Este conjunto de dados foi utilizado no desafio DCASE 2020, Tarefa 1B. O objetivo dessa tarefa foi desenvolver métodos com baixo custo computacional, considerando o tamanho máximo do modelo em 500KB. Para ter uma ideia de quão próximos nossos resultados estão dos resultados obtidos no desafio, nosso objetivo foi manter o modelo com no máximo 500KB.

A *Tampere University* foi responsável pela gravação do som ambiente e sua rotulação. A gravação foi realizada em doze grandes cidades europeias, sendo elas: Amsterdam, Barcelona, Lisboa, Lyon, Madri, Praga, Milão, Helsinque, Londres, Paris, Estocolmo e Viena. O conjunto de dados contém 40h de áudio (14400 segmentos de 10 segundos). O conjunto de dados já vem com um *split* treino e teste com 70% e 30% respectivamente. A taxa de amostragem de todos os áudios é de 48KHz com resolução de 24 bits.

4.2 *Baseline*

Com o conjunto de dados, a DCASE também disponibiliza para cada um dos seus desafios um *baseline* que consiste em códigos que oferecem uma solução inicial. A ideia é que os resultados obtidos com esse experimento sirvam de referência para os participantes.

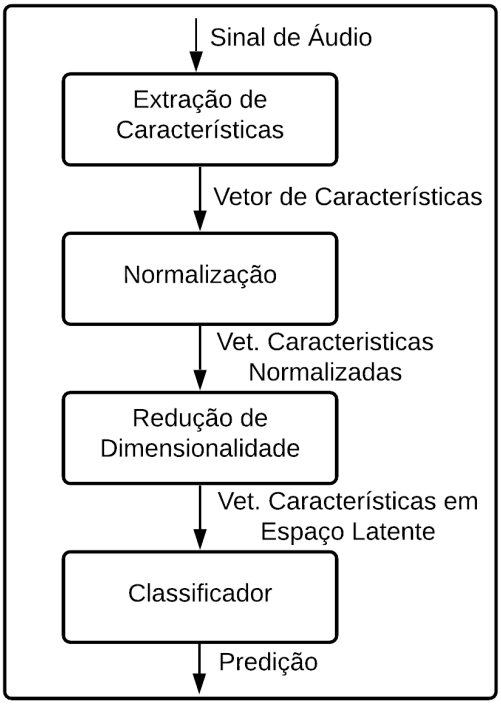
No *baseline* Heittola, Mesaros e Virtanen (2020), foram extraídas as *Log Mel-band Energies* em 40 *bands* para cada 10 segundos de sinal de áudio. O sistema implementa uma abordagem baseada em CNN. A CNN consiste em dois blocos convolucionais 2D (camada de convolução 2D seguida por uma camada de *max pooling*). Onde a camada de *max pooling* é a

responsável por compactar as camadas convolucionais anteriores em um mapa de características compactas, seguidos por uma camada totalmente conectada. A entrada tem o tamanho de 40×500, referente a quantidade de características *Log Mel-bands* extraídas. O modelo gerado tem 450 KB e obteve 87,30% de acurácia, sendo esse resultado obtido pela *baseline* não sendo trivial de ser alcançado.

4.3 Descrição do Sistema

A arquitetura usada neste trabalho é representada pela figura 8 que apresenta as principais etapas para a tarefa de classificação automática de cenas acústicas.

Figura 8 – Representação do Sistema Proposto



Fonte: Autoria própria (2022).

Primeiramente os atributos apresentados no Capitulo 2 são extraídos do sinal de áudio. Em seguida, o vetor de características é normalizado utilizando a estratégia *z-score*. Em seguida, a redução de dimensionalidade é realizada por meio da projeção para um espaço de menor dimensionalidade, usando uma das técnicas descritas na Seção 2.4 sendo elas PCA, RP, NMF . Por fim, o vetor de características pode servir de entrada para um classificador.

4.4 Deslocamento Positivo

Os atributos utilizados neste trabalho podem assumir valores negativos. Portanto, para condicionar os dados para aplicar NMF, o seguinte procedimento foi realizado. Após a normalização usando *z-score* (segunda etapa mostrada na Figura 8), o menor valor do conjunto de treino T_{min} é encontrado. O conjunto de treino é então deslocado por $|T_{min}|$, resultando em um vetor com apenas valores não-negativos. Para condicionar o conjunto de teste um procedimento semelhante é aplicado. Primeiramente o conjunto de teste é normalizado com os parâmetros do *z-score* encontrados no conjunto de treino. Em seguida, os elementos do conjunto de teste que forem menores que T_{min} são zerados arbitrariamente. Pelas propriedades do $z - score$, poucos elementos serão menores que T_{min}. Por fim, o deslocamento também é realizado adicionando $|T_{min}|$ a todos os elementos do conjunto de teste.

5 RESULTADOS EXPERIMENTAIS

Visando a construção de um sistema de baixo custo computacional, realizamos dois experimentos baseados em redução de dimensionalidade e seleção de instâncias, respectivamente. No primeiro experimento avaliamos o impacto de três métodos de redução de dimensionalidade baseado projeções lineares (PCA, Random Projection e NMF) no tamanho dos modelos e nas taxas de acerto. No segundo experimento avaliamos o impacto seleção de instâncias representativas para as classes usando um método baseado em agrupamento por *KMeans*.

5.1 Experimento com Projeções Lineares

Neste experimento, nosso principal objetivo foi aplicar técnicas de projeção linear e observar o impacto sobre as taxas de acerto e os tamanhos dos modelos gerados. Para isso escolhemos a base de dados já citada, três diferentes conjuntos de características LBP, LPQ e MARSYAS e três abordagens de redução de dimensionalidade PCA, *Random Projection* (RP) e NMF. As abordagens avaliadas são largamente empregadas em diversas tarefas descrevendo ótimos resultados. A etapa de extração de características utiliza uma abordagem global de forma que é gerado um vetor de característica por amostra.

O SVM foi escolhido como classificador por já ter sido utilizado com sucesso nesta tarefa em diversos trabalhos descritos na literatura (Costa *et al.*, 2011), (NANNI *et al.*, 2018). A fim de comparação, usamos o classificador KNN para avaliarmos as diferenças de tamanhos e desempenho entre os modelos. Para o KNN foi realizada uma busca por parâmetros de distância e número de vizinhos (k). Os parâmetros de ambos classificadores foram determinados por busca exaustiva via validação cruzada. Na validação cruzada, o conjunto de treino foi particionado em um conjunto de treino, com 80% dos exemplos e outro, de validação, com 20% dos exemplos. O particionamento foi feito de forma que uma mesma combinação (lugar $\times$ cidade) não estivesse presente em ambos conjuntos de treino e validação. Foi usada a biblioteca *Scikit Learn* (PEDREGOSA *et al.*, 2011) em todos nossos experimentos. A Tabela 1 apresenta os valores avaliados em cada parâmetro.

Tabela 1 – Hiperparâmetros avaliados durante a busca exaustiva

SVM	C	$\{1, 10, 100, 1000\}$
	γ	$\{0.125, 0.25, 0.5, 1/\text{n\_features}\}$
	kernel	{rbf, polinomial}
KNN	k	$\{1, 3, 5, 7, 9\}$
	distância	{euclidean, manhattan, chebyshev, minkowski}

Adotamos o seguinte critério para avaliar o tamanho dos modelos. No caso da SVM o tamanho dos modelos é relacionado ao número de vetores de suporte selecionados durante o treinamento e a dimensionalidade dos vetores. No caso do KNN, o tamanho dos modelos é proporcional ao tamanho do conjunto de treinamento e a dimensionalidade dos vetores. Os

métodos PCA e NMF codificam a transformação linear em uma matriz, cujo tamanho também é considerado em todos os resultados apresentados. Embora o método RP também codifique a transformação linear em uma matriz, esta matriz não precisa ser armazenada, uma vez que pode ser gerada a partir de uma semente pré-determinada, portanto somente a semente precisa ser armazenada. Portanto, os resultados do método RP não contemplam o tamanho da matriz.

Desta forma, com o intuito de avaliar o desempenho dos conjuntos de características e o tamanho em megabytes dos modelos gerados, avaliamos individualmente os descritores de textura LBP e LPQ e as características MARSYAS. A Tabela 2 apresenta o *F1-Score(%)* destes conjuntos de características usando os classificadores KNN e SVM.

Tabela 2 – Desempenho e tamanho dos modelos individuais

Descritor	F1-Score(%)		Tamanho (MB)	
	KNN	SVM	KNN	SVM
LBP	75,13	81,89	3,9	1,6
LPQ	76,96	84,27	16,3	6,6
MARSYAS	77,86	84,03	4,5	1,9

Em (NANNI *et al.*, 2017) é demonstrado que a combinação de características podem ser benéficas para classificação de áudio. Desta forma, realizamos a concatenação a nível de características dos três conjuntos, LBP, LPQ e MARSYAS, os quais possuem dimensões de 59, 256 e 72, respectivamente, totalizando um vetor de características com 387 atributos. A Tabela 3 descreve os resultados alcançados através da concatenação (*early fusion*) dos conjuntos de características.

Tabela 3 – Desempenho e tamanho do modelo com early-fusion

Classificador	F1-Score(%)	Tamanho (MB)
KNN	82,16	24,6
SVM	88,76	9,2

É possível notar que através da abordagem de fusão de características conseguimos um aumento no desempenho de ambos os classificadores. Entretanto, o tamanho do modelo também aumentou consideravelmente.

A partir da concatenação das características empregamos três algoritmos de redução de dimensionalidade a fim de conseguirmos modelos mais robustos e compactos. Desta forma, nosso objetivo é utilizar todos os 387 atributos gerados a partir da concatenação do LBP, LPQ e MARSYAS, além de avaliar o impacto da redução do número de atributos para, 256, 128, 64, 32, 16 e 8.

A Tabela 4 apresenta os resultados alcançados utilizando três diferentes abordagens de redução de dimensionalidade através dos classificadores KNN e SVM. Os resultados são apresentados usando seis diferentes níveis de compactação, desta forma é possível notar o impacto em relação ao desempenho do modelo ao reduzirmos o número de atributos.

Tabela 4 – Desempenho usando diferentes abordagens de redução de dimensionalidade, PCA, RP e NMF

Número de Atributos	F1-Score(%)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	83,14	**88,76**	81,25	85.90	81,42	88.28
128	83,76	**88,79**	80,52	85,66	82,03	88,40
64	84,29	**88,50**	79,94	80,32	81,86	88,09
32	83,50	**88,06**	76,30	75,31	82,91	86,24
16	80,18	85,04	67,62	70,69	81,43	**86,64**
8	78,02	**81,89**	67,62	66,09	76,42	80,21

Através da Tabela 4 podemos observar que o classificador SVM continua apresentando os melhores resultados. O RP apresenta os piores resultados para dimensões mais altas, ou seja, quando há pouca redução de atributos. O PCA e NMF apresentaram taxas similares, porém podemos destacar o PCA como apresentando o melhor desempenho através do menor número de componentes principais. Desta forma, foi possível reduzir de 387 para 8 dimensões, obtendo 81,89% de *F1-Score* e um tamanho de 293,49 kB.

A Tabela 5 mostra a redução de tamanhos em megabytes ao usar os 387 atributos e ao aplicar o PCA. É possível notar que evoluímos de um modelo de 9,2 MB para aproximadamente 400 KB com o mesmo desempenho usando a SVM. A Tabela também mostra que os modelos SVM são menores que os modelos KNN em todos os casos analisados. Este é um padrão esperado, uma vez que o KNN armazena todos os vetores do conjunto de treino, enquanto SVM armazena apenas os vetores de suporte.

A partir destes experimentos observamos que mesmo empregando técnicas simples é possível gerar modelos robustos e com baixa dimensionalidade.

Tabela 5 – Tamanho em Megabytes usando diferentes abordagens de redução de dimensionalidade, PCA, RP e NMF

Número de Atributos	Tamanhos(MB)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	15,80	6,00	16,30	**5,30**	15,80	**5,30**
128	8,40	3,40	8,60	3,10	8,40	**2,50**
64	4,20	1,70	4,40	1,60	4,20	**1,20**
32	2,20	0,85	2,20	0,83	2,20	**0,65**
16	1,10	0,40	1,10	1,04	1,10	**0,39**
8	0,65	0,29	1,10	0,48	0,64	**0,28**

Os resultados alcançados demonstram que utilizar métodos de redução de dimensionalidade pode ser uma estratégia simples e que possibilita gerarmos modelos computacionalmente baratos e com praticamente o mesmo desempenho. Avaliamos três diferentes abordagens de redução de dimensionalidade e foi possível notar que em todos os casos conseguimos gerar modelos com baixa dimensionalidade e com uma melhora no desempenho. Usando o NMF foi possível gerar um modelo de 16 componentes com 393,58 Kilobytes e com F1-Score de

86,64%, ou seja, uma redução de apenas 2,1 pontos percentuais e uma redução de 9,2 Mb para 393,58 kB. Modelos gerados usando o SVM conseguiram ter desempenho melhores e em alguns casos modelos com menos de 500Kb. Usando o classificador KNN, os modelos gerados superam os 500kb e o desempenho foi pior, se comparado ao SVM.

A partir dos experimentos é possível notar que conseguimos alcançar desempenho similar aos apresentados na literatura empregando técnicas simples de extração de características, redução de dimensionalidade e classificação.

5.2 Experimento com Seleção de Instâncias

O objetivo deste experimento é reduzir o número de instâncias do conjunto de treino e, ao mesmo tempo, evitar a deterioração das taxas de acerto. Para este segundo experimento foram mantidas as configurações listadas na Tabela 1 para a busca exaustiva com os classificadores KNN e SVM. Também foi mantido o conjunto com as 387 características oriundas da concatenação do LBP, LPQ e MARSYAS como também as técnicas de redução de dimensionalidade PCA, RP e NMF e os números de dimensões-alvo das reduções em 256, 128, 64, 32, 16, 8.

Foi adicionada uma nova camada ao sistema. Essa camada é responsável pela redução do número de instâncias do conjunto de treino. Essa camada faz o uso do *K-meansc* (FOLEIS; TAVARES, 2020). Esta técnica é baseada no algoritmo de clusterização e aprendizado não supervisionado *K-means* (MACQUEEN, 1967). O *K-means* realiza a estimação de pontos chamados centroides que caracterizam as várias tendências de um determinado conjunto de dados que após calculados podem ser usados para inferir a tendência de outros pontos.

A técnica *K-meansc* consiste em usar o *K-means* para encontrar as tendencias de um conjunto de dados e representá-lo pelos centróides. Foram avaliados diferentes números de centróides para representar as classes. O número de centróides corresponde a 5%, 10%, 15% e 20% do número total de instâncias no conjunto de treinamento.

Após a etapa de normalização e redução de dimensionalidade por projeções lineares, o *K-means* é utilizado para calcular as tendências de todas as instâncias de cada classe separadamente. Após o agrupamento, os centróides que representam cada classe são usados como conjunto de treinamento. Por fim, este conjunto de treinamento é usado para o treinamento dos classificadores.

Conforme pode ser visto na Tabela 6 somente utilizando o *K-meansc* a redução no tamanho dos modelos é significativa quando comparados aos resultados sem a redução do conjunto de treinamento, conforme mostrado na Tabela 3. Isto é esperado, uma vez que os modelos treinados armazenam parte dos conjuntos de treinamento. Desta forma, o tamanho dos modelos diminui proporcionalmente. Entretanto, o K-meansc sozinho não é suficiente para reduzir o modelo para menos que 500KB.

Tabela 6 – F1-Score usando *K-meansc* sem redução de dimensionalidade

Centróides	F1-Score(%)		Tamanho(MB)	
	KNN	SVM	KNN	SVM
20%	80,08	88,4	15,4	6,5
15%	80,38	88,02	11,1	5,1
10%	82,61	87,51	7,7	3,8
5%	80,46	86,52	3,8	2,2

A próxima etapa foi adicionar as técnicas de redução de dimensionalidade antes da redução do conjunto de treinamento com *K-meansc*. As taxas de acerto estão nas Tabelas 7, 9, 11 e 13.

Como pode ser visto na Tabela 7 e Tabela 8 mesmo usando somente 5% de cada classe para o treino, não houve uma grande deterioração na taxa de acerto quando comparados aos resultados alcançados no experimento anterior 5.1. Quando comparamos com o tamanho do modelo é vista uma grande redução sendo esses os resultados com os menores modelos encontrados.

Tabela 7 – F1-Score usando *K-meansc* 5% e diferentes abordagens de redução de dimensionalidade

Número de Atributos	F1-Score(%)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	83,76	**86,52**	79,76	83,16	83,18	84,46
128	85,01	**86,39**	81,76	82,53	81,93	85,43
64	84,09	**87,07**	79,66	76,67	81,88	86,26
32	82,43	**86,69**	75,69	74,37	82,69	84,96
16	80,96	83,56	71,44	61,98	80,72	**84,02**
8	76,22	**80,36**	67,53	69,01	74,38	78,19

Tabela 8 – Tamanhos dos modelos (em MB) usando *K-meansc* 5% e diferentes abordagens de redução de dimensionalidade

Número de Atributos	Tamanho(MB)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	3,3	2,3	16,6	**1,4**	3,3	1,9
128	1,7	1	8,6	**0,73**	1,6	0,84
64	0,82	0,53	4,2	**0,38**	0,79	0,44
32	0,40	0,27	2,2	**0,19**	0,40	0,27
16	0,22	0,14	1,1	0,20	0,21	**0,10**
8	0,12	**0,07**	0,62	0,11	0,11	**0,07**

Os resultados encontrados na Tabela 9 e Tabela 10 usando 10% de cada classe do treino não exibem grandes mudanças se comparado aos de 5% encontrados nas tabelas 7 e 8 somente havendo um aumento nos tamanhos dos modelos.

Tabela 9 – F1-Score usando *K-meansc* 10% e diferentes abordagens de redução de dimensionalidade

Número de Atributos	F1-Score(%)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	82,05	**88,02**	80,34	83,64	82,88	86,53
128	82,23	**87,66**	80,79	81,97	83,12	85,35
64	83,48	**87,23**	78,12	78,27	83,56	87,15
32	83,70	**86,79**	78,90	75,35	83,83	85,86
16	80,48	84,46	71,59	64,17	82,20	**85,04**
8	76,82	**81,70**	63,79	59,58	75,96	79,45

Tabela 10 – Tamanhos dos modelos (em MB) usando *K-meansc* 10% e diferentes abordagens de redução de dimensionalidade

Número de Atributos	Tamanho(MB)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	5,9	3,3	16,6	**2,3**	5,9	2,7
128	3	1,6	8,6	**1,1**	3	1,2
64	1,4	0,77	4,4	0,58	1,5	**0,66**
32	0,71	0,39	2,2	**0,30**	0,75	0,33
16	0,37	0,17	1,1	0,36	0,39	**0,16**
8	0,20	**0,10**	0,64	0,19	0,21	0,11

Na Tabela 11 e Tabela 12 estão os resultados com a maior quantidades de modelos com o tamanho abaixo do limite de 500kb definido pela *baseline* da DCASE em 4.4 e o resultado do PCA com 32 componentes com o classificador SVM foi o resultado que superou tanto em taxa quanto em tamanho os resultados apresentados pela *baseline* do DCASE.

Tabela 11 – F1-Score usando *K-meansc* 15% e diferentes abordagens de redução de dimensionalidade

Número de Atributos	F1-Score(%)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	81,64	**87,91**	81,88	83,97	79,44	86,36
128	83,17	**87,93**	79,86	82,66	81,36	87,27
64	83,69	87,28	78,31	79,47	82,41	**87,34**
32	83,83	**87,84**	76,06	74,84	83,39	86,68
16	80,04	**85,53**	69,86	68,05	81,68	85,44
8	77,62	**81,36**	63,33	68,25	76,03	78,91

Tabela 12 – Tamanhos dos modelos (em MB) usando *K-meansc* 15% e diferentes abordagens de redução de dimensionalidade

Número de Atributos	Tamanho(MB)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	7,9	4,2	17,1	**3**	7,9	3,4
128	4	2,1	8,4	**1,6**	4	1,7
64	2	1	4,2	**0,81**	2	0,84
32	1	**0,41**	2,2	0,42	1	0,43
16	0,53	0,23	1,1	0,52	0,55	**0,22**
8	0,29	**0,14**	0,62	0,26	0,30	0,15

Em Tabela 13 e Tabela 14 estão resultados que não tiveram grandes mudanças em comparação com os demais, porém são os que tem os maiores modelos e o menor número de resultados dentro do limite proposto no Capítulo 4.4.

Tabela 13 – Taxas usando a abordagem do *K-meansc* 20% e usando diferentes abordagens de redução de dimensionalidade, PCA, RP e NMF

Número de Atributos	F1-Score(%)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	82,05	**88,02**	80,34	83,64	82,88	86,53
128	82,23	**87,66**	80,79	81,97	83,12	85,35
64	83,48	**87,23**	78,12	78,27	83,56	87,15
32	83,7	**86,79**	78,9	75,35	83,83	85,86
16	80,48	84,46	71,59	64,17	82,2	**85,04**
8	76,82	**81,7**	63,79	59,58	75,96	79,45

Tabela 14 – Tamanhos usando a abordagem do *K-meansc* 20% e usando diferentes abordagens de redução de dimensionalidade, PCA, RP e NMF

Número de Atributos	Tamanho(MB)					
	PCA		RP		NMF	
	KNN	SVM	KNN	SVM	KNN	SVM
256	10,5	5	5,9	**3,9**	11	4,3
128	5,3	2,5	2,8	**2**	5,5	2,1
64	2,7	**0,96**	1,4	0,97	2,7	1
32	1,4	0,61	0,71	0,55	1,4	**0,52**
16	0,74	0,28	0,37	0,66	0,71	**0,26**
8	0,41	**0,17**	0,21	0,32	0,39	0,18

Os experimentos utilizando a junção das técnicas de redução de dimensionalidade e *K-meansc* atingiram uma maior quantidade de resultados que ficam abaixo do limite de 500KB definido, incluindo pela primeira vez o KNN dentro desse limite. Os piores resultados foram alcançados pelo Random Projection que conseguiu somente em um único experimento ficar abaixo do limite. O experimento utilizando o PCA e o *K-meansc* 15% atingiu a melhor marca entre todos os experimento.

6 CONCLUSÕES

Esse trabalho teve como objetivo principal investigar a classificação automática de categorias de cenas acústicas usando algoritmos clássicos. Especificamente, apresentamos um estudo de como técnicas de redução de dimensionalidade baseadas em projeções lineares e a subamostragem do conjunto de treinamento podem impactar o tamanho dos modelos e as taxas de acerto na classificação.

Foram apresentados dois experimentos que avaliaram a combinação das técnicas de redução de dimensionalidade por projeções lineares e subamostragem do conjunto de treinamento, separadamente e conjuntamente. O primeiro experimento nos permitiu avaliar o impacto das técnicas de redução de dimensionalidade nas taxas de classificação e no tamanho dos modelos. Os resultados mostraram reduções nos tamanhos dos modelos e baixa perda nas taxas de acerto. Entretanto, somente estas abordagens não foram suficientes para atingir os requisitos propostos no desafio DCASE.

Com base nos resultados do primeiro experimento notamos que seria necessário aplicar outra técnica que fosse complementar aos algoritmos já utilizados para buscar uma melhora na redução do tamanho do modelo. Como ambos classificadores guardam dados para realizar a classificação, a técnica *K-meansc* foi usada nesse segundo experimento, trazendo para os classificadores uma redução no conjunto de treinamento. Esta redução no conjunto de treinamento resultou na necessidade de armazenar menos elementos no modelo, o que diminuiu satisfatoriamente o tamanho dos modelos, sem perda considerável nas taxas de acerto.

Esta redução no tamanho dos modelos colocou resultados de ambos classificadores dentro do limite de tamanho estipulado na competição. Além disso, no segundo experimento, a taxa de acerto obtida pelo *baseline* foi ultrapassada deixando vários resultados competitivos com as técnicas que usaram aprendizagem profunda, somente usando classificadores clássicos que não precisam de *hardwares* aceleradores.

Os resultados obtidos ainda ficaram abaixo dos encontrados nos trabalhos relacionados. Entretanto, eles nos indicaram que é possível obter modelos com acurácia e tamanho competitivos utilizando somente classificadores clássicos e técnicas de projeção linear e seleção de instâncias. Para trabalhos futuros é possível avaliar outras técnicas de redução de dimensionalidade baseadas em métodos de seleção de características e outras técnicas de subamostragem do conjunto de treinamento.

REFERÊNCIAS

ALTMAN, N.; KRZYWINSKI, M. The curse(s) of dimensionality. **Nature Methods**, v. 15, n. 6, p. 399–400, jun. 2018. ISSN 1548-7091, 1548-7105. Disponível em: http://www.nature.com/articles/s41592-018-0019-x.

BAILEY, D. H.; SWARZTRAUBER, P. N. A Fast Method for the Numerical Evaluation of Continuous Fourier and Laplace Transforms. **SIAM Journal on Scientific Computing**, v. 15, n. 5, p. 1105–1110, set. 1994. ISSN 1064-8275, 1095-7197. Disponível em: http://epubs.siam.org/doi/10.1137/0915067.

BELAHCENE, M. *et al.* Local descriptors and tensor local preserving projection in face recognition. *In*: **2016 6th European Workshop on Visual Information Processing (EUVIP)**. Marseille, France: IEEE, 2016. p. 1–6. ISBN 9781509027811. Disponível em: http://ieeexplore.ieee.org/document/7764608/.

BINGHAM, E.; MANNILA, H. Random projection in dimensionality reduction: Applications to image and text data. *In*: **Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01**. San Francisco, California: ACM Press, 2001. p. 245–250. ISBN 9781581133912. Disponível em: http://portal.acm.org/citation.cfm?doid=502512.502546.

Candes, E. J.; Wakin, M. B. An introduction to compressive sampling. **IEEE Signal Processing Magazine**, v. 25, n. 2, p. 21–30, 2008.

CHU, S. *et al.* Where am i? scene recognition for mobile robots using audio features. *In*: IEEE. **2006 IEEE International conference on multimedia and expo**. Toronto, Ont., Canada, 2006. p. 885–888.

Costa, Y. M. G. *et al.* Music genre recognition using spectrograms. *In*: **2011 18th International Conference on Systems, Signals and Image Processing**. Sarajevo, Bosnia-Herzegovina: IEEE, 2011. p. 1–4.

Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 28, n. 4, p. 357–366, 1980.

ERNST, R. R.; ANDERSON, W. A. Application of Fourier Transform Spectroscopy to Magnetic Resonance. **Review of Scientific Instruments**, v. 37, n. 1, p. 93–102, jan. 1966. ISSN 0034-6748, 1089-7623. Disponível em: http://aip.scitation.org/doi/10.1063/1.1719961.

FOLEIS, J. H.; TAVARES, T. F. Texture selection for automatic music genre classification. **Applied Soft Computing**, v. 89, p. 106127, abr. 2020. ISSN 15684946. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1568494620300673.

GABRILOVICH, E.; MARKOVITCH, S. Text categorization with many redundant features: Using aggressive feature selection to make svms competitive with c4.5. *In*: **Twenty-first international conference on Machine learning - ICML '04**. Banff, Alberta, Canada: ACM Press, 2004. p. 41. Disponível em: http://portal.acm.org/citation.cfm?doid=1015330.1015388.

GOETZE, S. *et al.* Acoustic monitoring and localization for social care. **Journal of Computing Science and Engineering**, v. 6, n. 1, p. 40–50, 2012.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 1157–1182, mar. 2003. ISSN 1532-4435.

GUYON, I. *et al.* Gene selection for cancer classification using support vector machines. **Machine Learning**, v. 46, n. 1/3, p. 389–422, 2002. ISSN 08856125. Disponível em: http://link.springer.com/10.1023/A:1012487302797.

HEITTOLA, T.; MESAROS, A.; VIRTANEN, T. **TAU Urban Acoustic Scenes 2019, Development dataset**. Zenodo, 2019. Type: dataset. Disponível em: https://zenodo.org/record/2589280.

HEITTOLA, T.; MESAROS, A.; VIRTANEN, T. Acoustic scene classification in DCASE 2020 Challenge: Generalization across devices and low complexity solutions. **arXiv:2005.14623 [eess]**, maio 2020. ArXiv: 2005.14623. Disponível em: http://arxiv.org/abs/2005.14623.

Heittola, Toni; Mesaros, Annamaria; Virtanen, Tuomas. **TAU Urban Acoustic Scenes 2020 3Class, Evaluation dataset**. Zenodo, 2020. Type: dataset. Disponível em: https://zenodo.org/record/3685835.

HELMHOLTZ, H. L. F. **On the Sensations of Tone as a Physiological Basis for the Theory of Music**. 3. ed. Cambridge: Cambridge University Press, 2009. (Cambridge Library Collection - Music).

HERSHEY, S. *et al.* cnn architectures for large-scale audio classification. *In*: **2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. New Orleans, LA: IEEE, 2017. p. 131–135. ISBN 9781509041176. Disponível em: http://ieeexplore.ieee.org/document/7952132/.

HOWLEY, T. *et al.* The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. *In*: MACINTOSH, A.; ELLIS, R.; ALLEN, T. (Ed.). **Applications and Innovations in Intelligent Systems XIII**. London: Springer London, 2006. p. 209–222. ISBN 9781846282232. Disponível em: http://link.springer.com/10.1007/1-84628-224-1_16.

HU, H. *et al.* **Device-Robust Acoustic Scene Classification Based on Two-Stage Categorization and Data Augmentation**. Atlanta, Geórgia, 2020.

JOHNSON, W. B.; LINDENSTRAUSS, J. Extensions of lipschitz mappings into a hilbert space. **Contemporary mathematics**, v. 26, n. 189-206, p. 1, 1984. ISSN 978-0-8218-5030-5.

KOUTINI, K. *et al.* **CP-JKU Submissions to DCASE'20: Low-Complexity Cross-Device Acoustic Scene Classification with RF-Regularized CNNs**. Linz, Austria, 2020.

LEE, D.; SEUNG, H. Learning the parts of objects by non-negative matrix factorization. **Nature**, v. 401, p. 788–91, 11 1999.

MAATEN, L. V. D.; POSTMA, E.; HERIK, J. Van den. Dimensionality reduction: A comparative review. **J Mach Learn Res**, v. 10, n. 66-71, p. 13, 2009.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In*: CAM, L. M. L.; NEYMAN, J. (Ed.). **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics**. Berkeley, Calif.: University of California Press, 1967. (Statistics, v. 1), p. 281–297. Disponível em: https://projecteuclid.org/euclid.bsmsp/1200512992.

MCDONNELL, M. **Low-Complexity Acoustic Scene Classification Using One-Bit-Per-Weight Deep Convolutional Neural Networks**. Adelaide, Austrália, 2020.

MESAROS, A.; HEITTOLA, T.; VIRTANEN, T. A multi-device dataset for urban acoustic scene classification. *In*: DCASE. **Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)**. 2018. p. 9–13. Disponível em: https://arxiv.org/abs/1807.09840.

MüLLER, A. C. **NMF & Outlier Detection**. 2018. Disponível em: https://amueller.github.io/COMS4995-s19/slides/aml-16-nmf-outlier-detection/#4.

NANNI, L. *et al.* Bird and whale species identification using sound images. **IET Computer Vision**, v. 12, n. 2, p. 178–184, 2018. Disponível em: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cvi.2017.0075.

NANNI, L. *et al.* Combining visual and acoustic features for audio classification tasks. **Pattern Recognition Letters**, v. 88, p. 49–56, 2017. ISSN 0167-8655. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167865517300132.

Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 971–987, 2002.

OJALA, T.; PIETIKäINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. **Pattern Recognition**, v. 29, n. 1, p. 51–59, 1996. ISSN 0031-3203. Disponível em: http://www.sciencedirect.com/science/article/pii/0031320395000674.

OJANSIVU, V.; HEIKKILä, J. Blur Insensitive Texture Classification Using Local Phase Quantization. *In*: HUTCHISON, D. *et al.* (Ed.). **Image and Signal Processing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. v. 5099, p. 236–243. ISBN 9783540699040 9783540699057. Disponível em: http://link.springer.com/10.1007/978-3-540-69905-7_27.

PATKI, P. **Exploring Compact Alternatives to Deep Learning in Task 1B**. Maryland, United States, 2020.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RINGNÉR, M. What is principal component analysis? **Nature biotechnology**, Nature Publishing Group, v. 26, n. 3, p. 303–304, 2008.

SINHA, S. *et al.* Spectral decomposition of seismic data with continuous-wavelet transform. **GEOPHYSICS**, v. 70, n. 6, p. P19–P25, nov. 2005. ISSN 0016-8033, 1942-2156. Disponível em: https://library.seg.org/doi/10.1190/1.2127113.

STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. **The Journal of the Acoustical Society of America**, v. 8, n. 3, p. 185–190, 1937.

TAY, A. *et al.* Authentication of Olive Oil Adulterated with Vegetable Oils Using Fourier Transform Infrared Spectroscopy. **LWT - Food Science and Technology**, v. 35, n. 1, p. 99–103, fev. 2002. ISSN 00236438. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0023643801908643.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. **IEEE Transactions on Speech and Audio Processing**, v. 10, n. 5, p. 293–302, jul. 2002. ISSN 1063-6676, 1558-2353. Disponível em: https://ieeexplore.ieee.org/document/1021072/.

WALDEKAR, S.; A, K. K.; SAHA, G. **Mel-Scaled Wavelet-Based Features for Sub-Task A and Texture Features for Sub-Task B of DCASE 2020 Task 1**. Tampere, Finland, 2020.

XU, Y.; LI, W. J.; LEE, K. K. C. **Intelligent Wearable Interfaces**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008. ISBN 9780470222867 9780470179277. Disponível em: http://doi.wiley.com/10.1002/9780470222867.

ZHOU, S.-R.; YIN, J.-P.; ZHANG, J.-M. Local binary pattern (LBP) and local phase quantization (LBQ) based on gabor filter for face representation. **Neurocomputing**, v. 116, p. 260–264, set. 2013. ISSN 09252312. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0925231212008181.