CLASSIFICAÇÃO DE CENÁRIO ACÚSTICO UTILIZANDO HANDCRAFTED FEATURES E NON-HANDCRAFTED FEATURES

João Vitor Staub Castanho (PIBIC/CNPq/FA/UEM), Yandre Maldonado e Gomes da Costa (Orientador). E-mai: yandre@din.uem.br

Universidade Estadual de Maringá, Centro de Tecnológica, Maringá, PR.

Ciências Exatas e da Terra / Ciências Ciência da Computação

Palavras-chave: classificação automática, espectrograma, reconhecimento de padrões, cenário acústico

RESUMO

Este trabalho se propôs a analisar o desempenho de duas técnicas de classificação de cenários acústicos, utilizando *handcrafted* e *non-handcrafted features*, e ao fim fundir as duas técnicas para obter melhores resultados na classificação. Foi utilizada a base de dados do desafio DCASE 2016, que abrange amostras de 15 cenários distintos. O sinal de áudio foi convertido para espectrogramas para a classificação, além disso, explorou-se o emprego do uso de *data augmentation* para otimizar as taxas de classificação. Utilizando *handcrafted features*, dos espectrogramas foram extraídas características visuais utilizando um descritor de textura. Já para *non-handcrafted features* o modelo convolucional foi alimentado com as imagens de espectrograma e a extração de características e classificação ficou bom conta das camadas presentes no modelo da rede. Ao final foi obtido uma acurácia de 64,17% utilizando *handcrafted features*, de 91,15% utilizando *non-handcrafted features* e de 92,45% fundindo os dois classificadores. Utilizando *non-handcrafted features* e o método de fusão foi possível superar o *baseline* originalmente previsto pelo próprio desafio, que apresentava acurácia de 77,2%.

INTRODUÇÃO

A paisagem sonora de um ambiente é composta por uma variedade de sons que carregam informações essenciais sobre o local em que foi gravado. A tarefa fundamental de classificar cenários acústicos desempenha um papel crucial ao discernir e categorizar ambientes distintos com base em suas assinaturas sonoras únicas.

Este projeto se concentra em um estudo comparativo da extração e classificação de cenários acústicos, explorando o potencial das características *handcrafted features* e *non-handcrafted features* como parâmetros para a etapa de classificação. Além disso, o estudo incorpora uma abordagem de fusão entre esses dois métodos, buscando ampliar ainda mais a precisão e eficácia da classificação.

Diversos trabalhos indicam métodos que caracterizam tarefas de classificação automática de som, utilizando as duas técnicas que foram comparadas neste trabalho. Em (COSTA, Y. M. G. et al, 2012) foi examinada a eficácia de atributos derivados da textura de uma representação tempo frequência de som, espectrograma, na classificação de gêneros musicais. Já em (HAN, Y. e LEE, K., 2015) foi explorado o uso de redes neurais convolucionais (do inglês *Convolutional Neural Network*, CNN), que, segundo eles, se mostraram superiores ao uso de *handcrafted features* para classificação. O propósito deste trabalho é comparar as duas técnicas de classificação em cenários acústicos, além de realizar uma fusão entre as duas técnicas.

MATERIAIS E MÉTODOS

Base de dados

A base de dados escolhida para realização da classificação foi a *TUTAcousticScenes_2016_DevelopmentSet* (MESAROS, A. et al, 2018), que possui gravações de cenários acústicos divididos em 15 classes, praia do lago, ônibus, café/restaurante, carro (dirigindo ou viajando como passageiro, na cidade), centro da cidade, trilha na floresta, mercearia (tamanho médio), residência, biblioteca, estação de metrô, escritório, parque, área residencial, trem e bonde. Para todas as cenas acústicas, as gravações foram feitas em locais diferentes: ruas diferentes, parques diferentes, casas diferentes, com uma taxa de amostragem de 44,1 kHz e resolução de 24 bits. A base contém um total de 1170 arquivos, cada classe com 78 gravações de 30 segundos cada.

Métodos

A abordagem metodológica adotada neste trabalho foi baseada na utilização dos espectrogramas gerados a partir dos sinais de amostras de áudio, para a extração dos espectrogramas foi utilizada a biblioteca *librosa*, a escolha dessa biblioteca deve-se à sua eficiência e versatilidade em processar dados de áudio e extrair informações relevantes para a análise acústica. Antes de realizar o treinamento dos modelos e a classificação dos cenários, foi realizado o *data augmentation* da base de dados, a base que contém 1170 arquivos de áudio de 30 segundos cada foi transformada em 7020 espectrogramas de 5 segundos cada, ou seja, para cada arquivo foram obtidas 6 janelas de 5 segundos. Utilizando a biblioteca *librosa* o arquivo de áudio foi carregado utilizando uma taxa de amostragem de 22050 o arquivo é dividido e convertido para para decibéis, essa conversão é realizada para que o espectrograma contenha mais informações visuais sobre o cenário acústico, e por fim o espectrograma é salvo em arquivos de imagem,. Essa janela de 5 segundos foi escolhida pois em testes iniciais com os modelos foi utilizada uma janela de 5 segundos que se iniciava no segundo 3 e terminava no segundo 8, foram obtidos bons resultados, porém verificou-se a ocorrência de *overfitting*, e para

resolver esse problema utilizamos a estratégia de ampliação da quantidade de amostras de espectrogramas como uma estratégia de *data augmentation*,.
Para a extração de características *handcrafted* foi utilizado o *Local Binary Pattern* (LBP) que é um método de análise de textura que codifica a relação entre um pixel central e seus vizinhos em uma imagem. A classificação foi realizada através de uma biblioteca que implementa o algoritmo *Support Vector Machine* (SVM). Já para as características automáticas foi utilizada uma rede CNN, essa rede é composta por 10 camadas de convolução intercaladas com camadas de *max-pooling* para extrair características e reduzir a dimensionalidade, camadas de *dropout* são adicionadas para regularização e a rede culmina em camadas totalmente conectadas para classificação, com uma camada de saída *softmax* para produzir probabilidades de categorias de destino.
Para o treinamento de ambos os modelos foi utilizada uma técnica de validação cruzada onde a base de dados, composta por 7020 espectrogramas, foi dividida em 5 *folds*. Dentro de cada *fold* de treinamento, foram utilizados 1404 espectrogramas para validação e 5616 para o treinamento de ambos os modelos, CNN e SVM. Além de que todas as 6 representações retiradas de um arquivo de áudio foram colocadas no mesmo *fold* para evitar viés na classificação. Ao término do treinamento, as classes preditas foram coletadas e uma fusão foi realizada. Foram adotados diversos métodos de fusão, tais como borda count, soma simples, produto, média simples, média ponderada, maiores valores, maioria ponderada e fusão com rejeição.

RESULTADOS E DISCUSSÃO

Inicialmente foram realizados testes com a base antes do *data augmentation* para uma análise de quais seriam os melhores parâmetros para o modelo SVM. Além de que com esses testes foi possível detectar o *overfitting* no modelo CNN. Após o *data augmentation*, utilizando o modelo SVM foi obtido uma acurácia de 64,17%. Já para o modelo CNN foi obtido uma acurácia de 91,56%. Porém com a fusão dos dois modelos a melhor acurácia obtida foi pelo método do produto foi de 92,45%.

Tabela 1 Resultados dos treinamentos dos modelos antes e depois do *data augmentation*

	antes do *data augmentation*		depois do *data augmentation*	
	acurácia	desvio padrão	acurácia	desvio padrão
SVM	51,47%	0,05132	64,17%	0,20068
CNN	87,86%	0,87868	91,15%	0,23163
f_produto	**-**	**-**	**92,45%**	**0,23363**
f_media	-	-	91,85%	0,23198

f_media_pond.	-	-	91,73%	0,23223
f_maiores_val.	-	-	91,46%	0,23161

Na tabela a letra “f” representa fusão e “pond.” e “val.” representam ponderada e valores respectivamente.

A melhor acurácia dos métodos de fusão foi alcançada pela fusão do produto, que consiste em multiplicar a probabilidade de cada classe de um modelo pela probabilidade do outro modelo.

CONCLUSÕES

Neste estudo, foram exploradas e comparadas duas técnicas distintas de classificação de cenários acústicos, além de investigar a fusão entre as duas. A aplicação do *data augmentation* destacou-se por seus importantes resultados na classificação de ambas as abordagens. O modelo SVM, alimentado com histogramas LBP, alcançou uma acurácia de 64,17%, enquanto o modelo CNN, apresentou uma acurácia de 91,56%, o que evidencia uma melhora nos resultados de treinamento dos modelos antes do *data augmentation* que foram de 51,47% para o SVM e 87,86% para o CNN. Já a fusão desses modelos culminou em uma acurácia de 92,45%. Essas descobertas reforçam a compreensão de que a classificação utilizando *non-handcrafted features* se mostra superior e como diferentes abordagens podem se complementar, contribuindo para uma classificação mais precisa e robusta de cenários acústicos.

AGRADECIMENTOS

Ao professor Yandre Maldonado e Gomes da Costa pela orientação e o financiamento da pesquisa pelo CNPq.

REFERÊNCIAS

MESAROS, A.; HEITTOLA, T.; BENETTOS, E.; FOSTER, P.; LAGRANGE, M.; VIRTANEN, T.; PLUMBLEY, M. D. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 26, n. 2, p. 379-393. 2018. Disponível em: https://ieeexplore.ieee.org/document/8123864. Acesso em: 12 de ago. de 2023.

COSTA, Y. M. G.; OLIVEIRA, L. S.; KOERICH, A. L. Music genre classification using LBP textural features. **Signal Processing**, v. 92, n. 11, p 2723-2737. 2012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0165168412001478. Acesso em: 12 de ago. de 2023.

HAN, Y.; LEE, K. Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation. **arXiv.**, v. 14, n. 8, jul. 2016. Disponível em: https://arxiv.org/abs/1607.02383. Acesso em: 12 de ago. de 2023