PSI5123 - Lista de Exercícios 1

Marcos Paulo Pereira Moretti

2 de agosto de 2022

NUSP: 9345363
Curso: Mestrado
Domínio: Engenharia de Computação

Exercício 1

A seguir estão os artigos principais que serão utilizados. Ambos tratam da predição do idioma a partir da voz e apresentam diferentes metodologias, nas quais o projeto proposto vai se inspirar:

- Kotsakis R. (2020): Utiliza dados de transmissões via rádio em 4 idiomas diferentes: grego, inglês, francês e alemão. O objetivo é a criação de um modelo de reconhecimento de idioma genérico;
- Srinivas N.S.S. (2019): Utiliza dados de dois datasets diferentes: o Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus (IITKGP-MLILSC) que contém áudios em 15 idiomas falados na Índia; e o Oriental Language Recognition Speech Corpus (AP18-OLR) que contém áudios em 10 idiomas falados na Ásia. O objetivo é propor um novo parâmetro de Fourier (FP) e comparar sua performance em relação às MFCCs no reconhecimento de diferentes idiomas.

Abaixo também são apresentados outros artigos de auxílio. Nesses artigos, são construídas variáveis que poderão ser utilizadas no projeto atual:

- Bhattacharya S. (2022): Utiliza dados de três datasets diferentes: Ryerson-Audio-Visual database (RAVDESS), Berlin Database (EmoDb) and Italian Database (Emo-Vo) os quais estão em inglês, alemão and italiano respectivamente. O objetivo é detectar emoções humanas a partir de áudios em idiomas diferentes;

- Gupta V. (2022): Utiliza dados das salas de *chat* da plataforma *ShareChat* em 10 idiomas falados na Índia. O objetivo é a criação de um dataset contendo palavras que caracterizem conteúdo abusivo, seguida de uma validação da qualidade do dataset via métricas de modelagem.

Exercício 2

O banco de dados escolhido foi obtido a partir do site Open Speech and Language Resources OpenSLR (2019), no tópico Multilingual LibriSpeech (MLS: A Large-Scale Multilingual Dataset for Speech Research) (Pratap V. (2020)). Trata-se de vários datasets em 8 idiomas diferentes derivados dos *audiobooks* do LibriVox.

Foram baixados arquivos de áudio com extensão *.flac* (*Free Lossless Audio Codec*) (diferentemente da proposta inicial), os quais são comprimidos sem perda de informação. Foram escolhidos 4 dos 8 idiomas disponíveis (diferentemente da proposta inicial): português, italiano, polonês e espanhol.

Foi utilizada, para cada idioma, uma amostra de treino de aproximadamente 1 hora, já pré-selecionada na base de origem (pasta *limited_supervision*). Os arquivos de áudio são importados a partir dos IDs fornecidos nessa pasta. São utilizados 947 arquivos de áudio.

Foi utilizada a linguagem Python, além das bibliotecas *librosa*, *pandas* e *numpy* para importar os arquivos e tratar os dados.

Pretende-se dividir o dataset em dois: um de treino e um de teste. A divisão será estratificada por sexo e por idioma, para manter as proporções no treino e no teste.

Também pretende-se utilizar uma janela do tipo Hamming, uma vez que permite reduzir o *leakage* e melhorar a qualidade do espectro após a segmentação do sinal, sendo utilizada amplamente nos *papers* citados. Serão utilizados 4 tamanhos de janela diferentes, com 100 ms, 500 ms, 1000 ms e 2000 ms de duração. O tamanho da janela vai variar do "menor"comprimento, com o qual obtém-se melhor resolução temporal, até o "maior"comprimento, com o qual obtém-se melhor resolução espectral, e serão calculadas as *features* em cada caso.

A partir dos áudios segmentados, para cada tamanho de janela, pretende-se computar as *features* conforme sugeridas pelos *papers*. As tabelas de *features* que cada *paper* sugeriu estão nas figuras 1 e 2.

Algumas características do dataset escolhido estão na tabela 1. Nota-se que há um bom balanceamento de locutores masculinos e femininos, e também aproximadamente 1 hora de áudio para cada idioma.

Time-Domain	Spectral-Domain
rms	Npeaks_spectral
Npeaks_temporal	flux_avr, flux_std
lowenergy	rolloff_0.2, 0.5, 0.8, 0.9, 0.99
Nonsets	bright_500, 1000, 1500, 2000, 3000, 4000, 8000
event_density	sp_roughness, sp_irregularity
rthythm_clarity	Npitch, pitch_avr, pitch_std
zerocross	fundam_freq, inharmonicity
attacktime_avr,std	mode
attackslope_avr,std	Nhcdf, hcdf_avr, hcdf_std
	sp_centroid, sp_spread
Cepstral-Domain	sp_skewness, sp_kurtosis
mfcc1 ... 13	sp_flatness
	entropy

Figura 1: *Features* extraídas dos áudios segmentados (Fonte: Kotsakis R. (2020))

Feature index range	Feature description	Feature index range	Feature description	Feature index range	Feature description
1–120	$\bar{x}\ (H_k^p)$	601–720	$\bar{x}\ (\Delta H_k^p)$	1201–1320	$\bar{x}\ (\Delta\Delta H_k^p)$
121–240	$\tilde{x}\ (H_k^p)$	721–840	$\tilde{x}\ (\Delta H_k^p)$	1321–1440	$\tilde{x}\ (\Delta\Delta H_k^p)$
241–360	$\sigma_x\ (H_k^p)$	841–960	$\sigma_x\ (\Delta H_k^p)$	1441–1560	$\sigma_x\ (\Delta\Delta H_k^p)$
361–480	min (H_k^p)	961–1080	min $(\Delta H_k^p$	1561–1680	min $(\Delta\Delta H_k^p)$
481–600	max (H_k^p)	1081–1200	max (ΔH_k^p)	1681–1800	max $(\Delta\Delta H_k^p)$

Here, the range of k is defined as, $1 \leq k \leq 120$, where k denotes the harmonic coefficients. The range of p is defined as $1 \leq p \leq l$, where l is the number of speech frames. H_k denotes FPs, ΔH_k denotes first-order FPs, and $\Delta\Delta H_k$ denotes second-order FPs. The statistical parameters are computed with respect to k across l. $\bar{x}.$ = mean, $\tilde{x}.$ = median, $\sigma_x.$ = standard deviation, $min.$ = minimum, and max . = maximum

Figura 2: Parâmetros de Fourier (Fonte: Srinivas N.S.S. (2019))

Tabela 1: Características do dataset MLS

Idioma	Qtde. áudios	Feminino	Masculino	Feminino (%)	Masculino (%)	Tempo
Italiano	240	116	124	48.33%	51.67%	59”19’
Polonês	238	118	120	49.58%	50.42%	59”35’
Português	236	119	117	50.42%	49.58%	59”41’
Espanhol	233	115	118	49.36%	50.64%	59”40’

Segue em anexo o *notebook* com detalhes de implementação da obtenção dos dados.

Exercício 3

Na figura 3, é apresentada a sequência de passos do projeto.

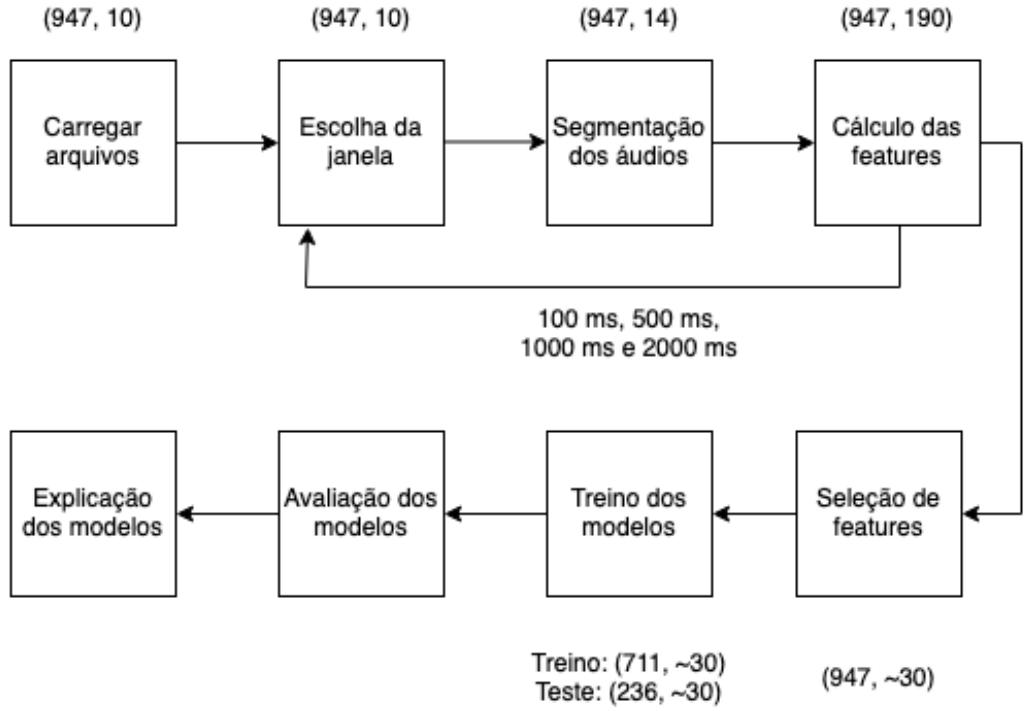

Figura 3: Sequência de passos de execução do projeto, com respectivas volumetrias esperadas entre parênteses

As etapas são descritas a seguir:

1. Carregar arquivos: Serão carregados os arquivos *.flac* e os metadados relacionados aos arquivos (sexo, idioma, id, título do li-

vro, taxa de amostragem, etc.), presentes na pasta que contém os áudios. Os arquivos serão carregados em mono (diferentemente da proposta inicial);

2. Escolha da janela: Serão utilizados 4 tamanhos de janelas diferentes. Trata-se de um *loop* variando o tamanho da janela para a construção das *features* em diferentes segmentações dos sinais de áudio;

3. Segmentação do sinal: O sinal é segmentado e janelado. A dimensão do dataset aumenta em 4 "*features*", que na verdade representam o sinal segmentado para cada tamanho de janela escolhido;

4. Cálculo das *features*: Serão calculadas as *features* descritas anteriormente em cada um dos segmentos. Foram claculadas 14 *features* temporais, 13 *features* cepstrais (MFCCs) e 31 *features* espectrais. Multiplica-se por 4 a soma de *features* cepstrais e espectrais, pois variam com o tamanho da janela. Assim, obtêm-se 176 *features* novas. Aqui não foram contabilizados os parâmetros de Fourier do *paper* Srinivas N.S.S. (2019), que poderão ser utilizados para enriquecer o dataset;

5. Seleção de *features*: Será utilizado um método de seleção de *features* chamado Boruta para capturar as features mais relevantes para modelagem. Estimam-se 30 *features* relevantes, mas é um valor a descobrir;

6. Treino dos modelos: O dataset será dividido em 75% de treino (711 exemplos) e 25% de teste (236 exemplos), estratificado por sexo e idioma. A variável resposta é o idioma. Serão treinados os modelos de Random Forest, SVM, Naive Bayes, Logistic Regression, Convolutional Neural Network e Recurrent Neural Network. Será utilizado um esquema de modelagem *one-vs-all*; ou seja, para cada classe, trata-se a classe como sendo positiva e as outras como sendo negativas;

7. Validação dos modelos: Serão computadas métricas a respeito do desempenho de classificação dos modelos, serão exibidas as matrizes de confusão, e em seguida será decidido o modelo que teve melhor desempenho;

8. Explicação dos modelos: A partir desse "modelo vencedor", será utilizada a biblioteca SHAP (*SHapley Additive exPlanations*) para identificar as variáveis que mais contribuíram para as classificações

realizadas. Com isso, espera-se identificar variáveis que diferenciem os idiomas entre si, e fornecer uma interpretação para as classificações.

Exercício 4

Serão utilizadas métricas de classificação usuais. Após um modelo classificar os exemplos, tem-se a seguinte classificação das predições:

- Verdadeiros Positivos (VP): exemplo i pertencia à classe C_k e o classificador o classificou corretamente como pertencente à classe C_k;
- Verdadeiros Negativos (VN): exemplo i não pertencia à classe C_k e o classificador o classificou corretamente em uma classe diferente de C_k;
- Falsos Positivos (FP): exemplo i não pertencia à classe C_k e o classificador o classificou incorretamente como pertencente à classe C_k;
- Falsos Negativos (FN): exemplo i pertencia à classe C_k e o classificador o classificou incorretamente em uma classe diferente de C_k.

A partir dessas classificações, são estabelecidas as métricas de desempenho a seguir:

- Acurácia: Calcula-se a partir da fórmula $(VP+VN)/(VP+VN+FP+FN)$. Significa o total de acertos que o modelo teve nos exemplos classificados, tanto acertos da classe positiva como da negativa. Em geral, é uma métrica enganosa, pois tem-se maior interesse no acerto da classe positiva em detrimento da negativa. Fornece um resultado assertivo somente no caso de classificação balanceada (classes positiva e negativa em igual quantidade);
- Precisão: Calcula-se a partir da fórmula $VP/(VP+FP)$. Significa a taxa de acerto das predições positivas do modelo. Ou seja, dentre todas as classificações preditas pelo modelo como positiva, calcula-se quanto que o modelo acertou. É utilizada quando os falsos negativos são mais prejudiciais que os falsos positivos;
- Revocação (*recall*): Calcula-se a partir da fórmula $VP/(VP+FN)$. Significa a taxa de acerto das predições do modelo em relação aos positivos reais. Ou seja, dentre todos os exemplos

originalmente positivos, calcula-se quanto que o modelo acertou. É utilizada quando os falsos positivos são mais prejudiciais que os falsos negativos;

- F1-*score*: Calcula-se a partir da média harmônica entre a Precisão e a Revocação. É uma ponderação das duas métricas. Nenhuma delas pode ser muito baixa, pois nesse caso o F1-*score* também o será.

Serão calculadas todas as métricas em cada treinamento realizado, mas um foco especial será dado à Precisão e ao F1-*score*.

Repositório de desenvolvimento

Este projeto está sendo desenvolvido em um repositório público no GitHub, e pode ser acompanhado a partir da URL https://github.com/mpereiramoretti/multilingual_classification

Referências

Bhattacharya S., Borah S., M. B. e. a. (2022). Emotion detection from multilingual audio using deep analysis. *Multimed Tools Appl*.

Gupta V., Sharon R., S. R. e. a. (2022). ADIMA: ABUSE DETECTION IN MULTILINGUAL AUDIO. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kotsakis R., Matsiola M., K. G. D. C. (2020). Investigation of Spoken-Language Detection and Classification in Broadcasted Audio Content. *MDPI - Multidisciplinary Digital Publishing Institute*, (4).

OpenSLR (2019). OpenSLR.

Pratap V., Xu Q, S. A. e. a. (2020). Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Srinivas N.S.S., Sugan N., K. N. e. a. (2019). Recognition of Spoken Languages from Acoustic Speech Signals Using Fourier Parameters. *Circuits Syst Signal Process*, (38).