Aalto University
School of Electrical Engineering

"Multiple voice disorders in the same individual: Investigating handcrafted features, multi-label classification algorithms, and base-learners"

Feifan Wang, Vikramaditya Malik

ELEC-E5531
Speech and Language Processing Seminar

12.03.2024

Introduction

Acoustic analyses of voice disorders have been at the forefront of current biomedical research. Usual strategies, essentially based on machine learning (ML) algorithms, commonly classify a subject as being either healthy or pathologically-affected. Nevertheless, the latter state is not always a result of a sole laryngeal issue, i.e., multiple disorders might exist, demanding multi-label classification procedures for effective diagnoses. Consequently, the objective of this paper is to investigate the application of five multi-label classification methods based on problem transformation to play the role of base-learners, i.e., Label Powerset, Binary Relevance, Nested Stacking, Classifier Chains, and Dependent Binary Relevance with Random Forest (RF) and Support Vector Machine (SVM), in addition to a Deep Neural Network (DNN) from an algorithm adaptation method, to detect multiple voice disorders, i.e., Dysphonia, Laryngitis, Reinke's Edema, Vox Senilis, and Central Laryngeal Motion Disorder. Receiving as input three handcrafted features, i.e., signal energy (SE), zero-crossing rates (ZCRs), and signal entropy (SH), which allow for interpretable descriptors in terms of speech analysis, production, and perception

Literature review

Authors and references	Main approaches and tools
Al-Naheri et al. (2017)	feature extraction; frequency bands; SVM
Muhammad et al. (2012b)	feature extraction; GMM; MFCC
Muhammad and Melhem (2014)	MPEG-7 features; SVM
Vikram and Umarani (2013)	MFCC; GMM-UBM
Akbari and Arjmandi (2014)	DWPT; energy; entropy
Hemmerling et al. (2016)	cepstrum; PCA; random forest; K-means
Martinez et al. (2012)	GMM; MFCC; glottal-to-noise ratio
Saeedi and Almasganj (2013)	wavelets; GA; SVM
Mekyska et al. (2015)	Mann–Whitney U-test; parametrization
Ali et al. (2016)	psychophysics; GMM
Markaki and Stylianou (2011)	modulation-related features
Pranav and Sabarimalai (2017)	glottal instants; EGG features
Sasou (2017)	HLAC; jitter; shimmer; neural nets
Verde et al. (2018b)	gender; age; fundamental frequency
Lachhab et al. (2014)	GMM; HLDA
Zhong et al. (2016)	HMM; fuzzy MF; STFT
Fonseca and Pereira (2008)	LS-SVM; RBF kernels

Abbreviations to be used

- BR - Binary Relevance.
- CC - Classifier Chains.
- CLMD - Central Laryngeal Motion Disorder.
- DBR - Dependent Binary Relevance.
- DNN - Deep Neural Network.
- DYS - Dysphonia.
- LAR - Laryngitis.
- LP - Label Powerset.
- MLC - Multi-label Classification.
- NS - Nested Stacking.
- RDE - Reinke Edema.
- RF - Random Forest.
- SLC - Single-label Classification.
- SVM - Support Vector Machine.
- VSE - Vox Senilis.

Methodology

Five problem-transformation strategies and one algorithm adaptation method are selected.

The problem-transformation MLC methods, i.e., LP, BR, CC, NS, and DBR were chosen due to their notable performance in previous works and implemented using R language and the.

Our algorithm adaptation implementation was based on artificial neural networks The multi-layer perceptron network (MLP) was constructed using Keras for computational speed boost.

Here an MLP with five hidden layers (n-256-128-64-7), where n is the size of n-dimensional feature vector is proposed.

There are 2 sets made according to the SE where set1 has C = 1% and set2 has C = 10%

Dataset

Dataset distribution of samples and classes without balancing and with several balancing rate. After Barry and Putzer (2007).

Balancing rate	HEA	Pathology							Samples
		DYS	LAR	RDE	VSE	CLMD	DYS-LAR	LAR-RDE	
0% (Original)	686	69	81	33	22	10	4	9	914
20%	686	137	137	132	132	130	136	135	1625
35%	686	207	162	231	220	240	240	234	2220
50%	686	276	324	330	330	340	340	342	2968
65%	686	414	405	429	440	440	444	441	3699
80%	686	483	486	528	528	540	548	540	4339
95%	686	621	648	627	638	650	648	648	5166

Feature Extraction

- Signal energy
- Signal ZCRs
- Signal entropy

Feature Extraction - Signal Energy

It refers to the total amount of energy contained in a signal over a specific period of time or in a particular segment of the signal. Here :

$$SE(s[\cdot]) = \sum_{i=0}^{M-1} (s_i)^2$$

Feature Extraction - Signal ZCRs

ZCR of a signal is the rate at which the signal changes its sign. In other words, it is the number of times the signal crosses the zero axis per unit of time. Here :

$$ZCR(s[\cdot]) = \frac{1}{2} \sum_{j=0}^{M-2} |sign(s_j) - sign(s_{j+1})|$$

Feature Extraction - Signal Entropy

Measure of the randomness or unpredictability of a signal.

Here :

$$SH(s[\cdot]) = -\sum_{i=0}^{K-1} p_i \cdot log_{\beta}(p_i)$$

Base-learners Selected

RF and SVM classifiers (linear, polynomial and radial kernels) were selected.

- Well known algorithms
- A relevant number of speech pathology detection algorithms has employed SVMs for building their classification models

Evaluation

- **Assessed by using a 10-fold cross validation strategy**
- **Two baseline: majority and random**

$$accuracy = 1 - \frac{1}{m}\sum_{i=1}^{m}\frac{|Z_i \Delta Y_i|}{|L|} \quad , \qquad recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Y_i|} \quad ,$$

$$precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Z_i|}{|Z_i|} \quad , \qquad F1\text{-}score = \frac{1}{m}\sum_{i=1}^{m}\frac{2|Y_i \cap Z_i|}{|Y_i| \cup |Z_i|} \quad ,$$

where Y_i represents the ith instance of the true set of labels, Z_i represents ith instance of the predicted set of labels, and Δ represents the symmetric difference.

Results

- MLC predictive assessment for disorder prediction
- Machine learning inductive assessment and balancing improvements
- Related issues

MLC Predictive Assessment

Accuracy

Label	Method				
	LP	BR	DBR	CC	NS
HEA	**71.10%**	70.43%	70.39%	69.95%	69.32%
CLMD	96.35%	**96.88%**	96.78%	96.55%	96.65%
DYS	**90.75%**	90.07%	89.43%	90.12%	89.81%
LAR	**86.76%**	84.86%	85.17%	83.96%	85.05%
RDE	**90.89%**	90.19%	89.72%	87.88%	89.53%
VSE	94.91%	95.48%	**95.63%**	92.53%	95.19%
Average	88.46%	86.31%	87.85%	86.83%	87.59%

F1-score

Label	LP		BR		DBR		CC		NS		DNN	
	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2
HEA	0.828	0.763	0.794	0.730	0.802	0.732	0.798	0.711	0.797	0.713	0.858	0.818
CLMD	0.962	0.801	0.950	0.778	0.943	0.763	0.945	0.705	0.949	0.666	0.982	0.971
DYS	0.810	0.766	0.776	0.735	0.785	0.742	0.754	0.708	0.737	0.700	0.905	0.856
LAR	0.810	0.784	0.778	0.712	0.778	0.732	0.789	0.710	0.793	0.722	0.893	0.861
RDE	0.868	0.760	0.820	0.717	0.825	0.740	0.829	0.706	0.830	0.711	0.936	0.927
VSE	0.857	0.798	0.861	0.786	0.872	0.779	0.853	0.779	0.866	0.751	**0.919**	**0.934**
Avg	0.856	0.779	0.830	0.743	0.834	0.748	0.828	0.720	0.829	0.711	0.916	0.897

- Healthy samples presented the lowest accuracy (71.10%) but were classified with higher performance (F1-score) than the disorders one.
- It is worth mentioning that, even with few data samples in the original dataset, the experiments exposed different patterns from these combinations of multiple disorders. Likewise, the predictive performance increased when using SMOTE to expand the original set of samples.

ML Inductive Assessment and Balancing

F1-score (LP)

Method	Dataset	Classifier					
		RF	L-SVM	P-SVM	R-SVM	Majority	Random
LP	Original	0.7430	0.7503	0.7481	0.7503	0.7503	0.1530
	20% (r)	0.7926	0.5805	0.4799	0.5796	0.4218	0.1854
	35% (r)	0.8388	0.5900	0.4312	0.5869	0.3087	0.1988
	50% (r)	0.8779	0.5755	0.4129	0.5958	0.2309	0.2141
	65% (r)	0.9005	0.6000	0.4608	0.6253	0.1852	0.2169
	80% (r)	0.9162	0.6229	0.5033	0.6607	0.1577	0.2202
	95% (r)	**0.9262**	0.6472	0.5577	0.6852	0.1326	0.2203

ML Inductive Assessment and Balancing

F1-score

Method	Dataset	Classifier					
		RF	L-SVM	P-SVM	R-SVM	Majority	Random
BR	Original	0.7406	0.7503	0.7488	0.7479	0.7503	0.2258
	20% (r)	0.7513	0.5538	0.4774	0.5851	0.0799	0.2571
	35% (r)	0.8123	0.5154	0.4446	0.6010	0.1079	0.2647
	50% (r)	0.8536	0.5043	0.4265	0.6046	0.1143	0.2764
	65% (r)	0.8837	0.5161	0.4489	0.6391	0.1187	0.2781
	80% (r)	0.9008	0.5371	0.4761	0.6746	0.1241	0.2796
	95% (r)	**0.9124**	0.5415	0.4940	0.6968	0.1255	0.2776

Method	Dataset	Classifier					
		RF	L-SVM	P-SVM	R-SVM	Majority	Random
CC	Original	0.7377	0.7501	0.7475	0.7404	0.7503	0.2418
	20% (r)	0.7440	0.5550	0.4357	0.5390	0.0799	0.2640
	35% (r)	0.8028	0.5164	0.4212	0.5644	0.1079	0.2668
	50% (r)	0.8346	0.4968	0.4161	0.5720	0.1143	0.2664
	65% (r)	0.8684	0.5111	0.4412	0.6086	0.1187	0.2738
	80% (r)	0.8867	0.5371	0.4668	0.6440	0.1241	0.2825
	95% (r)	**0.9019**	0.5540	0.4860	0.6672	0.1255	0.2761

Method	Dataset	Classifier					
		RF	L-SVM	P-SVM	R-SVM	Majority	Random
DBR	Original	0.7421	0.7497	0.7485	0.7412	0.7503	0.2286
	20% (r)	0.7585	0.5573	0.4822	0.5855	0.0799	0.2705
	35% (r)	0.8161	0.5341	0.4553	0.6137	0.1079	0.2737
	50% (r)	0.8528	0.5390	0.4680	0.6220	0.1143	0.2719
	65% (r)	0.8850	0.5501	0.4957	0.6506	0.1187	0.2791
	80% (r)	0.9011	0.5710	0.5260	0.6884	0.1241	0.2761
	95% (r)	**0.9133**	0.5808	0.5411	0.7036	0.1255	0.2816

Method	Dataset	Classifier					
		RF	L-SVM	P-SVM	R-SVM	Majority	Random
NS	Original	0.7390	0.7503	0.7497	0.7464	0.7503	0.2911
	20% (r)	0.7431	0.5717	0.4787	0.5749	0.0799	0.2566
	35% (r)	0.7971	0.5238	0.4462	0.5904	0.1079	0.2391
	50% (r)	0.8352	0.5033	0.4433	0.5928	0.1143	0.2360
	65% (r)	0.8667	0.5214	0.4600	0.6233	0.1187	0.2314
	80% (r)	0.8927	0.5373	0.4947	0.6590	0.1241	0.2292
	95% (r)	**0.9016**	0.5451	0.5133	0.6833	0.1255	0.2271

ML Inductive Assessment and Balancing

F1-score (DNN)

Method	Dataset	Feature set		
		Set_1	Set_2	Average
DNN	20% (r)		0.897	0.906
	35% (r)	0.947	0.926	0.936
	50% (r)	0.958	0.938	0.948
	65% (r)	0.962	0.945	0.953
	80% (r)	0.956	0.952	0.954
	95% (r)	**0.972**	**0.955**	**0.963**

- And algorithm adaptation, the proposed DNN model, was capable of overcoming the problem transformation methods but unable to converge towards predictions for all possible class combinations.
- Data balancing was required to support reliable and improved results
- SMOTE with a balancing rate of 20% in Majority can reduce the unbalancing problem and provide classification improvements

Related Issues

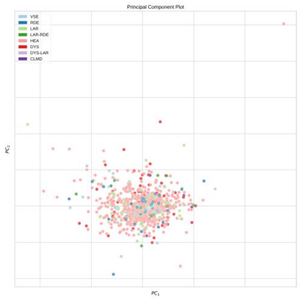
- Historically, laryngitis is known to be a serious research issue for speech technology problems, in particular for speaker recognition.
- A relevant evaluation among the possible strategies is to consider the number of produced models. Some methods could increase the number of models, requiring more computational resources and time to train the solution.
- Results revealed the DNN as the most predictive method demanding a single model to tackle the classification problem. However, additional efforts towards adapting the architecture and hyperparameters are required.

Discussion - Sets

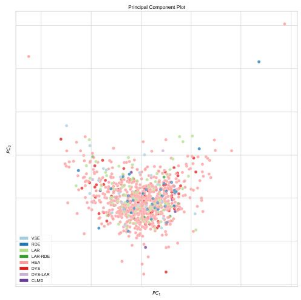
F1-score with RF

Label	LP		BR		DBR		CC		NS		DNN	
	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2	Set_1	Set_2
HEA	0.828	0.763	0.794	0.730	0.802	0.732	0.798	0.711	0.797	0.713	0.858	0.818
CLMD	0.962	0.801	0.950	0.778	0.943	0.763	0.945	0.705	0.949	0.666	0.982	0.971
DYS	0.810	0.766	0.776	0.735	0.785	0.742	0.754	0.708	0.737	0.700	0.905	0.856
LAR	0.810	0.784	0.778	0.712	0.778	0.732	0.789	0.710	0.793	0.722	0.893	0.861
RDE	0.868	0.760	0.820	0.717	0.825	0.740	0.829	0.706	0.830	0.711	0.936	0.927
VSE	0.857	0.798	0.861	0.786	0.872	0.779	0.853	0.779	0.866	0.751	**0.919**	**0.934**
Avg	0.856	0.779	0.830	0.743	0.834	0.748	0.828	0.720	0.829	0.711	0.916	0.897

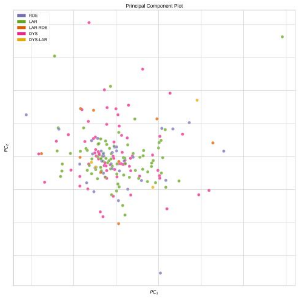
Discussion - PCA



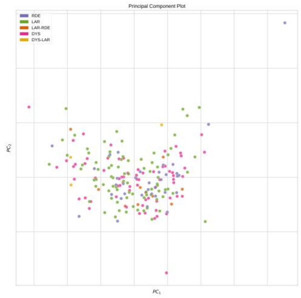
(a) All classes using Set_1 with $PC_1 = 0.69$ and $PC_2 = 0.18$.



(b) All classes using Set_2 with $PC_1 = 0.67$ and $PC_2 = 0.20$.



(c) Selected classes from Set_1 with $PC_1 = 0.84$ and $PC_2 = 0.11$.



(d) Selected classes from Set_2 with $PC_1 = 0.85$ and $PC_2 = 0.11$.

- Four PCAs were calculated to support a general overview of the features sets
- Two scenarios were built over all classes
- Other two PCAs were computed to expose the behaviour of single and multiple diseases focusing on DYS, LAR, RDE, DYS-LAR, and LAR-RDE patterns

Discussion - Summary

- Promising results depend on the discriminative capacity of the selected features. Thus, many features have been proposed and intensively experimented to describe temporal, spectral or time–frequency characteristics from voice data.
- Feature vectors were designed to achieve suitable results, where some detection scenarios, such as IoT, m-Health, and big data environments demand either a reduced usage of resources or the processing of a massive amount of voice data.
- DNN superiority was obtained considering the usage of synthetic samples to balance the training set and handcrafted features.
- DNN capacity to process raw data directly was not employed in this work to match our dataset size and to provide a fair comparison among all MLC methods and also, besides to study the handcrafted features.

Conclusion

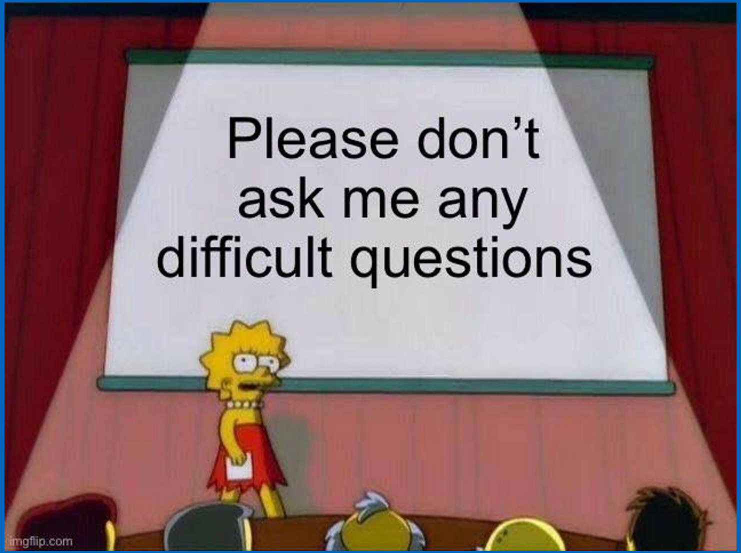
- Multi-label classification methods were successfully employed to identify subjects with healthy or pathologically-affected voices.
- The results have showed that all MLC methods were statistically superior to Random and Majority. The most complex prediction was related to the disorders that occur at the same time, however, all the disorders have superior predictive performance when compared to healthy subjects.
- Particularly, the DNN-based approach presented the best values of F1-score among the tested methods. C = 1% used to compute feature vectors composed by SE, ZCR, and SH is the best option.

Limitation and Future Work

- 13 multi-labelled samples were hundred times oversampled, causing a low variance in the dataset and, thus, degrading the statistical significance of the accuracies.
- As a future work, they suggested applying MLC to a database that presents the co-occurrence of additional voice pathologies, especially the complex ones.

Thank you for listening!

Feifan Wang, Vikramaditya Malik



Aalto University
School of Electrical Engineering

Assignments

1. What are the features used in the paper ? Define them.
2. How do they overcome the restriction of dataset? Explain the methods.