Classification of Developmental Disorders from Speech Signals using Submodular Feature Selection

Katrin Kirchhoff, Yuzong Liu, Jeff Bilmes

Department of Electrical Engineering University of Washington, Seattle

Interspeech 2013 Special Session Monday, Aug 26, 2013

Overview

- Focus of this work:
 - Autism sub-challenge: classification of developmental disorders
 - How to utilize given set of acoustic-prosodic features most effectively?
 - Improve classification / gain better insight into acoustic-prosodic correlates of developmental categories
- Large set of acoustic-prosodic features provided (6,373) but small number of training samples (903)
- Some features may be irrelevant/noisy/redundant
 - $\bullet \, \Rightarrow \, {\rm may}$ affect generalization performance of classifiers trained on this data
- Which features provide the most information for classifying developmental disorders?
- \Rightarrow Novel and general feature selection framework based on <u>submodularity</u>

- Submodular functions: class of set functions traditionally used in economics/operations research/game theory.
- Recent applications in machine learning: viral marketing, sensor placement, document summarization, structured norms
- Set functions defined as follows we are given:
 - a finite ground set of objects $V = \{v_1, ..., v_n\}$, |V| = n,
 - and a function of subsets to values $f: 2^V \to \mathbb{R}^+$.
 - For any $A \subseteq V$, f(A) provides a real number.
- A set function f is submodular if $\forall A \subseteq B$ and $v \notin B$

$$f(A \cup \{v\}) - f(A) \ge f(B \cup \{v\}) - f(B)$$
(1)

 Incremental value of v shrinks as the context in which it is considered grows from A to B (property of *diminishing returns*)

Background - Submodularity

• Example: Let V be a set of possible colored balls, and for any $A \subseteq V$, let f(A) give the number of different colors of the set A.



Initial value: 2 (colors in urn). New value with added blue ball: 3



Initial value: 3 (colors in urn). New value with added blue ball: 3

- On the left, adding a blue ball increases the number of colors. On the right, in the context of a superset, adding a blue ball does not increase the number of colors.
- Having more balls in an urn can never increase the incremental gain of adding a ball.
- Such an *f* is submodular.

- There are $2^{|V|}$ possible values of a set function without further assumptions, optimization is intractable and inapproximable.
- If f is monotone (∀A ⊆ B, f(A) ≤ f(B)) and submodular, however, it can be maximized, subject to a size constraint, using a simple greedy algorithm
- Theoretical performance guarantees: approximates optimal solution to within constant factor $1-1/e \approx 0.63$
- Fast accelerated greedy algorithm, $O(n \log n)$ with same guarantee, scales to large datasets

- Ground set V: original (high-dimensional) feature set
- Goal: find smaller subset A that expresses the same information as V and is non-redundant
- General objective function:

$$f(A) = \mathcal{L}(A) + \lambda \mathcal{R}(A)$$
(2)

- $\mathcal{L}(\mathcal{A})$: measures coverage of V by A
- $\mathcal{R}(\mathcal{A})$: measures diversity of \mathcal{A}
- λ : tradeoff parameter

Submodular Functions for Feature Selection

• Instantiation of $\mathcal{L}(A)$: facility location function

$$\mathcal{L}(A) = \sum_{i \in V} \max_{j \in A} w_{ij} \tag{3}$$

where w is a matrix of pairwise similarity values • Instantiation of $\mathcal{R}(\mathcal{A})$:

$$\mathcal{R}(A) = \sum_{n=1}^{N} \sqrt{\sum_{j \in P_n \cap A} r_j}$$
(4)

where $P_1, ..., P_N$ is partitioning of the ground set into N clusters through k-means clustering

- *N* is tuned on development set
- r_j : relevance score of item j: $r_j = \sum_{i \in V} w_{ij}/|V|$
- *w_{ij}* is mutual information between features *i* and *j*, computed from discretized features

Kirchhoff et al.

- Feature set provided by Challenge (6,373 acoustic-prosodic features)
- Multi-layer perceptron (MLP) classifier
 - Softmax output function
 - Trained on either 2 (Typicality) or 4 (Diagnostic) classes
 - Trained using backpropagation to minimize

$$F(x,\theta) = KL(p(c|x)||\hat{p}_{\theta}(c|x)) + \lambda||\theta||_{2}$$
(5)

x: input; θ : parameters (weights); c: class

• Use performance on development set to determine early stopping

- 6 different feature set sizes: 500, 1000, 2000, 3000, 4000, 5000
- For each feature set size, tested different number of hidden units in MLP:

100, 200, 300, 400, 500, 800, 1000, 2000, 3000, 4000

- Tested different values for N (number of clusters in diversity term), different values of λ
- Values were optimized on development set
 - Typicality: $\lambda = 5, N = 8$, features: 3000, HUs: 400
 - Diagnostic: $\lambda = 20, N = 32$, features: 3000, HUs: 800
- Comparison: modular feature selection method
 - rank all features by their mutual information with class label
 - select the top N features

Typicality task	System	Acc (%)	UAR (%)	# features
	Official baseline	92.6	92.8	6373
	MLP baseline	93.5	93.7	6373
	Modular	92.7	92.7	2000
	Submodular	93.7	94.1	3000

Diagnostic task	System	Acc (%)	UAR (%)	# features
	Official baseline	69.8	51.4	6373
	MLP baseline	76.9	51.6	6373
	Modular	76.8	54.2	2000
	Submodular	78.6	56.5	3000

System	Acc (%)	UAR (%)		
Typicality				
Official baseline*	_	90.7		
Submodular system	92.7	92.5		
Submodular system*	93.8	92.6		
Diagnostic				
Official baseline*	_	67.1		
Submodular system	79.5	57.4		
Submodular system*	83.9	64.4		

*: system was retrained on combined training and dev set 10% of data for submodular system was held out

Top-ranking features selected by submodular criterion (most representative, non-redundant features)

	Typicality	Diagnostic
Rank	Feature	Feature
1	pcm_Mag_spectralCentroid_sma_minPos	pcm_Mag_spectralCentroid_sma_minPos
2	pcm_Mag_psySharpness_sma_percentile99.0	pcm_Mag_psySharpness_sma_percentile99.0
3	audSpec_Rfilt_sma[12]_lpc0	audSpec_Rfilt_sma[12]_lpc0
4	pcm_Mag_spectralRollOff75.0_sma_maxPos	pcm_Mag_spectralRollOff75.0_sma_maxPos
5	pcm_Mag_spectralRollOff75.0_sma_de_pctIrange0-1	pcm_Mag_spectralRollOff75.0_sma_de_pctIrange0-1
6	audSpec_Rfilt_sma[24]_lpc0	audSpec_Rfilt_sma_de[2]_minPos
7	audSpec_Rfilt_sma[19]_lpc0	audSpec_Rfilt_sma[24]_lpc0
8	pcm_Mag_spectralSkewness_sma_maxPos	audSpec_Rfilt_sma[19]_lpc0
9	audSpec_Rfilt_sma[5]_lpc0	audSpec_Rfilt_sma[5]_lpc0
10	audSpec_Rfilt_sma[10]_flatness	audSpec_Rfilt_sma[10]_flatness
11	pcm_Mag_psySharpness_sma_segLenStddev	audSpec_Rfilt_sma[1]_pctIrange0-1
12	pcm_Mag_spectralKurtosis_sma_pctIrange0-1	logHNR_sma_amean
13	audSpec_Rfilt_sma[15]_lpc0	audSpec_Rfilt_sma[15]_lpc0
14	audSpec_Rfilt_sma[8]_lpc0	pcm_Mag_spectralKurtosis_sma_pctlrange0-1
15	audSpec_Rfilt_sma[1]_pctIrange0-1	pcm_Mag_fband250-650_sma_pctIrange0-1
16	pcm_Mag_fband1000-4000_sma_rqmean	logHNR_sma_de_percentile99.0
17	pcm_Mag_psySharpness_sma_peakRangeAbs	audSpec_Rfilt_sma[2]_peakRangeAbs
18	logHNR_sma_amean	pcm_Mag_fband1000-4000_sma_rqmean
19	pcm_Mag_fband250-650_sma_pctIrange0-1	pcm_RMSenergy_sma_quartile2
20	audspecRasta_lengthL1norm_sma_de_maxPos	pcm_Mag_psySharpness_sma_segLenStddev

Thank you! Questions?