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, WA, 98195, USA
{kk2,yzliu,bilmes}@uw.edu

Abstract

We present our system for the Interspeech 2013 Computational Paralinguistics Autism Sub-challenge. Our contribution focuses on improving classification accuracy of developmental disorders by applying a novel feature selection technique to the rich set of acoustic-prosodic features provided for this purpose. Our feature selection approach is based on submodular function optimization. We demonstrate significant improvements over systems using the full feature set and over a standard feature selection approach. Our final system outperforms the official Challenge baseline system significantly on the development set for both classification tasks, and on the test set for the Typicality task. Finally, we analyze the subselected features and identify the most important ones.

Index Terms: classification, feature selection, neural networks, submodular functions

1. Introduction

This paper describes the University of Washington's contribution to the 2013 Interspeech Computational Paralinguistics Challenge. Our study focuses on the Autism Sub-Challenge only, with the express purpose of studying novel machine learning methods to enhance classification performance on the provided feature set. In particular, we utilize a novel feature selection technique based on submodular function optimization. This method is designed to select a feature subset that expresses the same information as the original feature set while taking into account the dependencies between the selected features. As a result the subsequent classifier can devote all of its modeling power to only the relevant features rather than modeling redundant features. We test submodular feature selection in combination with a neural-network classifier and demonstrate that is outperforms the baseline system on the development sets and on the test set for the Typicality task. Finally, we examine the selected lists of features under various parameterizations of the selection algorithm and identify the most relevant features, i.e. those that collectively 'summarize' most of the information in the full feature set.

2. Feature selection

The task of the Autism Sub-challenge in the Interspeech 2013 Computational Paralinguistics Challenge is to classify children's speech samples into either *typical* vs. *atypical* (for the Typicality task) or into one of the four categories *typical*, *pervasive developmental disorder*, *pervasive developmental disorder - not otherwise specified*, and *dysphasia* (for the Diagnostic task). A precomputed set of 6,373 acoustic-prosodic features was provided for this purpose. At the same time, the training and development sets contain only 903 and 819 samples, respectively. Due to the high-dimensional feature space and small training set, it is possible that statistical classifiers trained on this data will overfit to the training set. It is also likely that a feature set of this size contains redundant or irrelevant features. Most importantly, in order gain deeper insight into the problem of classifying neurodevelopmental disorders from speech, it would be desirable to have an explicit assessment of the importance of different features. Several feature selection techniques have been proposed in the past. Feature transformations such as linear or non-linear principal component analysis, or linear discriminant analysis, provide one way of reduce the feature set size. However, they project the original dimensions into a new space where they lose their original interpretation, which would defeat our purpose. Explicit feature selection (as opposed to transformation) methods typically apply a search procedure for a feature subset in combination with an optimization criterion. The search procedure may involve forward selection, i.e. adding features one by one, or may take the form of backward selection or filtering, where features are iteratively removed from the original set. Both methods usually involve a "rank-and-select" approach, which computes the quality of each feature in isolation, ranks all features by that measure, and selects the top-scoring (or removes the bottom-scoring) feature. Such approaches do not consider the quality or informativeness of a feature when combined with the already selected set; therefore, the resulting set may still be redundant. Methods that do take dependencies into account, such as correlation-based feature selection [1], or maximum-relevancy-minimum redundancy [2], are computationally expensive and often do not scale well to high-dimensional feature spaces.

In this paper we consider a feature selection technique based on submodular functions that, in terms of the objective being optimized, provides near-optimal performance guarantees and that can be carried out by a fast accelerated greedy algorithm and hence is scalable to large data sets.

2.1. Submodular functions

While submodular functions have been popular in mathematics, economics, and operations research, they have recently been used in various machine learning problems, such as sensor placement [3], document summarization [4], dictionary selection [5], training data subset selection [6, 7], and random variable subset selection [8]. Given a finite ground set of objects $V = \{v_1, ..., v_n\}$ and a function $f : 2^V \rightarrow \mathbb{R}_+$ that returns a real value for any subset $S \subseteq V$, f is submodular if $\forall A \subseteq B$, and $v \notin B$,

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B). \quad (1)$$

Thus, the incremental "value" of v decreases as the context in which v is considered grows from A to B. If a function is monotone ($\forall A \subseteq B, f(A) \leq f(B)$) and submodular, it can be maximized under a cardinality constraint by a greedy algorithm [9] that guarantees a solution to within a constant factor $1 - 1/e$ of optimal. This algorithm, moreover, scales to large data sets.

2.2. Submodular functions for feature selection

Initial work on applying submodular functions to feature selection in acoustic feature spaces was presented in [10]. Building on this work we utilize a general surrogate objective function as in [4] which has the form:

$$f(S) = \mathcal{L}(S) + \lambda \mathcal{R}(S), \tag{2}$$

where $\mathcal{L}(S)$ measures how well the selected feature subset S *covers*, or *represents*, the original full feature set and the second term $\mathcal{R}(S)$ measures how *diverse* or non-redundant the selected features are, and λ is a tradeoff parameter. We use a surrogate objective since the exact objective (accuracy) is intractable to repeatedly evaluate, and even intractable to estimate measures of quality as in [8]. We define two instantiations of the $\mathcal{L}(S)$ term: The **facility location function** is defined as

$$\mathcal{L}_1(S) = \sum_{i \in V} \max_{j \in S} w_{ij}, \tag{3}$$

where $w_{ij} \geq 0$ measures the similarity between features i and j. This functions indicates how well each feature $i \in V$ is represented by the selected subset S.
The **saturated coverage function** is defined as:

$$\mathcal{L}_2(S) = \sum_{i \in V} \min\{C_i(S), \beta C_i(V)\}, \tag{4}$$

where $C_i(S) = \sum_{j \in S} w_{i,j}$ measures the degree to which i is "covered" by S, and $\beta \in [0, 1]$ is a hyperparameter that determines a global saturation threshold, the minimum within each term keeps features from being over-represented by subset S.
A **diversity function** $\mathcal{R}(S)$ can be added to these to reduce the redundancy of the selected subset of features. We define it as

$$\mathcal{R}(S) = \sum_{n=1}^{N} g(\sum_{j \in P_n \cap S} r_j) \tag{5}$$

where g is a monotone non-decreasing concave function, $P_n, n = 1, ..., N$ is a partition of the ground set V according to some clustering, and r_j measures the importance of the selected element j to the whole subset S. The above function measures the diversity of a subset, since the function encourages selecting features from different clusters. In our work, we adopt the following instantiation of $\mathcal{R}(S)$: the concave function g is defined as the square-root function; the ground set is partitioned into N blocks using the k-means algorithm, where N is chosen from $\{4, 8, 16, 32, 64\}$, and the best N is tuned on the development set; r_j is defined as $r_j = \sum_{i \in V} w_{ij}/|V|$. The diversity term is added to either the facility location or saturated graph cut function, weighted by a λ parameter.

All objectives are non-negative monotone submodular, hence the optimization problem for feature selection can be solved near-optimally in terms of f using a greedy algorithm. Submodularity has another advantage, namely the accelerated greedy algorithm [11] that in practice significantly speeds up greedy selection. Submodularity is equivalent to the gain in f being monotone nonincreasing: $f(S_i \cup \{k\}) - f(S_i) \geq f(S_j \cup \{k\}) - f(S_j), \forall S_i \subseteq S_j$. In the accelerated greedy algorithm, a list of uppers bounds $\rho(i), i \in V$ is maintained in decreasing order (often implemented using a priority queue) and during each greedy iteration, the list pops up the top element k^*. Once the gain associated with k^* is greater than any other elements $k \in V \setminus S$, we can safely add k^* to S. Algorithm 1 gives the accelerated greedy algorithm [11].

Algorithm 1 The accelerated greedy algorithm [11].

Given: features $\{v_i\}_{i \in [N]}$**, a desired number of features** $K < N$**, and a similarity graph** W **where** w_{ij} **is the pairwise mutual information between** i **and** j**.**
Initialize $S \leftarrow \emptyset$**, a priority queue** $\rho \leftarrow \emptyset$
for $i = 1, 2, \cdots, N$ **do**
  $\delta \leftarrow f(v_i)$
  $\rho.push(tuple(i, \delta))$
end for
while $|S| \leq K$ **do**
  $k^* \leftarrow \rho.top().key$
  $\rho.pop()$
  $\delta = f(S \cup \{k^*\}) - f(S)$
  if $\delta > \rho.top().value$ **then**
    $S \leftarrow S \cup \{k^*\}$ {submodularity guarantees that $\delta \geq f(S \cup \{k\}) - f(S), k \in V \setminus S$ }
  else
    $\rho.push(tuple(k^*, \delta))$ {re-sort otherwise}
  end if
end while
return S

These functions require a similarity measure w. We use the estimated pairwise mutual information between features i and j. We compute this by discretizing the continuous features into 50 equal-width bins and computing discrete mutual information:

$$MI(i; j) = \sum_i \sum_j p(i, j) \frac{p(i, j)}{p(i)p(j)} \tag{6}$$

We compare submodular feature selection against our baseline, a traditional feature selection approach using mutual information (MI) between individual features and target labels, ranking features in descending order, and selecting the top k features — this is a "modular" approach since it treats each feature entirely independently (Eq. (1) is satisfied everywhere with equality).

3. Data and Classifiers

We use the standard feature set provided for the Autism Sub-Challenge [12], consisting of 6,373 acoustic-prosodic features. As classifiers we train multilayer perceptrons (MLPs) using the QuickNet software[1], with one input layer, one hidden layer, and one output layer. The number of output units is 2 for the Typicality task and 4 for the Diagnostic task. The number of units in the input layer is identical to the number of features utilized in the system. For each of our experiments, we investigate 6 different feature set sizes (500, 1000, 2000, 3000, 4000, and 5000). For each feature set size, different numbers of hidden units (100, 200, 300, 400, 500, 800, 1000, 2000, 3000, and 4000) are investigated, leading to $10 \times 6 = 60$ classification

[1] `http://www1.icsi.berkeley.edu/Speech/qn.html`, but we modified this to support L2-norm regularization.

System	Acc (%)	UAR (%)	# features	# hus
Official - Typ	92.6	92.8	N/A	N/A
Official - Diag	69.8	51.7	N/A	N/A
Baseline - Typ	93.5	93.7	6373	2000
Baseline - Diag	76.9	51.6	6373	100
Modular - Typ	92.7	92.7	2000	300
Modular - Diag	76.8	54.2	2000	1000

Table 1: Accuracy (Acc) and Unweighted Average Recall (UAR) rates on the development set for the official baseline, our own in-house baseline systems: MLPs trained on the full feature set, and the best systems using modular feature selection. Typ = typicality task, Diag = diagnostic task. Column 3 indicates the number of input features and column 4 gives the number of hidden units (hus) in the MLP classifiers.

experiments for each classification task (typicality vs. diagnostic) and feature selection method. In each case, the optimal number of features and hidden units is optimized on the development set. The hidden layer uses the sigmoid activation function; the output layer uses the softmax function. The MLPs are trained using backpropagation to minimize the Kullback-Leibler divergence between the predicted and the true probability distributions over the output classes, plus a penalty term that implements a form of L2 regularization and discourages weights from becoming too large:

$$F(x, \theta) = KL(p(c|x)||\hat{p_\theta}(c|x)) + \lambda||\theta||_2 \quad (7)$$

where θ are the parameters, and λ is a coefficient indicating the weight of the penalty term. Different λ's are used for the two layers in the MLPs; their values are tuned on the development set in each case. An iterative learning schedule with decreasing learning rate is used. Training stops when the accuracy on a held-out cross-validation set starts to decrease. We use the official definitions of the training and development sets for training and cross-validation, respectively.

4. Experiments and Results

As baseline systems we trained MLPs on the full feature set, and on feature sets selected using the modular method described above in Section 2.2. Table 1 shows the impact on classification performance on the development set. We see that modular feature selection deteriorates the performance for the Typicality task compared to the baseline, though it does improve the UAR for the Diagnosis task by almost 3 points. The number of system parameters, computed as $P = (N_f \times N_{hu}) + N_{hu} + (N_{hu} \times N_o) + N_o$ where N_f is the number of features, N_{hu} is the number of hidden units, and N_o is the number of output units, decreases quite drastically for the Typicality task but increases for the Diagnostic task.

Next we investigated the submodular feature selection functions, i.e. the facility location and saturated graph cut functions, in each case combined with the diversity function. For each experimental condition (defined by the λ weight and number of clusters N for the diversity function), we varied the number of selected features and hidden units as explained in the previous section. The best performance on the development set was obtained by the facility location function, with parameters $\lambda = 5, N = 8$ for the Typicality task, and $\lambda = 20, N = 32$ for the Diagnosis task. The performance of the submodular systems

System	Acc (%)	UAR (%)	# features	# hus
Typicality - devel	93.7	94.1	3000	400
Typicality - test	92.7	92.5		
Typicality - test*	93.8	92.6		
Diagnosis - devel	78.6	56.5	3000	800
Diagnosis - test	79.5	57.4		
Diagnosis - test*	83.9	64.4		

Table 2: Accuracy and Unweighted Average Recall (UAR) rates on the development and test sets for the best systems using submodular feature selection. Systems marked by $*$ were retrained on the joint training and development set.

Rank	Feature
1	pcm_Mag_spectralCentroid_sma_minPos
2	pcm_Mag_psySharpness_sma_percentile99.0
3	audSpec_Rfilt_sma[12]_lpc0
4	pcm_Mag_spectralRollOff75.0_sma_maxPos
5	pcm_Mag_spectralRollOff75.0_sma_de_pctlrange0-1
6	audSpec_Rfilt_sma[24]_lpc0
7	audSpec_Rfilt_sma[19]_lpc0
8	pcm_Mag_spectralSkewness_sma_maxPos
9	audSpec_Rfilt_sma[5]_lpc0
10	audSpec_Rfilt_sma[10]_flatness
11	pcm_Mag_psySharpness_sma_segLenStddev
12	pcm_Mag_spectralKurtosis_sma_pctlrange0-1
13	audSpec_Rfilt_sma[15]_lpc0
14	audSpec_Rfilt_sma[8]_lpc0
15	audSpec_Rfilt_sma[1]_pctlrange0-1
16	pcm_Mag_fband1000-4000_sma_rqmean
17	pcm_Mag_psySharpness_sma_peakRangeAbs
18	logHNR_sma_amean
19	pcm_Mag_fband250-650_sma_pctlrange0-1
20	audspecRasta_lengthL1norm_sma_de_maxPos

Table 3: Features ranked highest by the submodular feature selection method – Typicality task.

on the development and test sets is shown in Table 2. On the Typicality task, the development set results are significantly better than those of the modular system or the Challenge baseline, and slightly better than the in-house baseline system while using less than half of the full feature set, and an order of magnitude fewer parameters than the in-house baseline. On the Diagnosis task submodular feature selection significantly improves the performance on the development set over the modular system, the in-house baseline, and the Official baseline. Test set results for the Challenge baseline were reported after re-training the system on the joint training and development set; we therefore report not only the test performance for the original system optimized on the development set, but also after having undergone similar retraining. Note that in our case we still held out 10% of the data to determine the early stopping point during training. On the Typicality task our system outperforms the 90.7% UAR reported for the Challenge system by almost 2 points. However, it remains below the 67.1% UAR of the Challenge baseline on the Diagnosis task, possibly because we were not able to make use of all of the development data.

Rank	Feature
1	pcm_Mag_spectralCentroid_sma_minPos
2	pcm_Mag_psySharpness_sma_percentile99.0
3	audSpec_Rfilt_sma[12]_lpc0
4	pcm_Mag_spectralRollOff75.0_sma_maxPos
5	pcm_Mag_spectralRollOff75.0_sma_de_pctlrange0-1
6	audSpec_Rfilt_sma_de[2]_minPos
7	audSpec_Rfilt_sma[24]_lpc0
8	audSpec_Rfilt_sma[19]_lpc0
9	audSpec_Rfilt_sma[5]_lpc0
10	audSpec_Rfilt_sma[10]_flatness
11	audSpec_Rfilt_sma[1]_pctlrange0-1
12	logHNR_sma_amean
13	audSpec_Rfilt_sma[15]_lpc0
14	pcm_Mag_spectralKurtosis_sma_pctlrange0-1
15	pcm_Mag_fband250-650_sma_pctlrange0-1
16	logHNR_sma_de_percentile99.0
17	audSpec_Rfilt_sma[2]_peakRangeAbs
18	pcm_Mag_fband1000-4000_sma_rqmean
19	pcm_RMSenergy_sma_quartile2
20	pcm_Mag_psySharpness_sma_segLenStddev

Table 4: Features ranked highest by the submodular feature selection method – Diagnosis task.

5. Analysis

An advantage of explicit feature selection (as opposed to feature transformation) is that the resulting ranking of features is interpretable. In order to determine which features are chosen preferentially, we chose the two best-performing parameter settings for our submodular feature selection algorithm for each classification task, averaged their ranks, and selected the top 20 features with the highest average rank. The results are shown in Tables 3 and 4. These are to be interpreted as the top features of the feature subsets that collectively express most of the information contained within the full feature set. A detailed description of each feature was not provided. However, it appears that for the Typicality task, most features are related to auditory spectrum or psychoacoustic spectral sharpness, or they characterize extreme points of the distribution of energy throughout the spectrum. For the Diagnosis task, the top features were identical – it is likely that these features are responsible for separating typical from non-typical speakers. Other features that are not present in Table 3 and thus are more informative for the other classes in the Diagnostic task include e.g. *logHNR_sma_de_percentile99.0* or *pcm_RMSenergy_sma_quartile2*.

In order to determine whether our feature selection method is equally useful in combination with other classifiers we re-ran the SVM classifier of the Official baseline with our subselected feature sets. However, results were below those obtained with the full feature set. This may be due to the use of a linear kernel in the SVM classifier, which prevents it from taking advantage of the interactions between features. Increasing the degree of the polynomial kernel to 2 or 3 to achieve non-linearity did not lead to any different results. Thus, it may be the case that the ability of MLPs to learn another implicit feature representation from the input feature set is of key importance here. One of our future goals is to investigate other combinations of subselected features and classifiers, such as deep neural networks, and SVMs with other kernels.

6. Summary

We have described our system developed for the Autism Subchallenge for Interspeech 2013. The main contribution of this system is the use of a novel feature selection technique which simultaneously improves classification results, reduces the number of features used, and provides an explicit ranking of features that is amenable to human inspection. We have demonstrated significant improvements over the official baseline system on the development set for both tasks and on the test set for the Typicality task. Most of the selected features relate to the computation of the auditory spectrum, psychoacoustic spectral sharpness, or the global energy distribution.

7. Acknowledgments

This material is partially based on research sponsored by Intelligence Advanced Research Projects Activity (IARPA) under agreement number FA8650-12-2-7263. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Intelligence Advanced Research Projects Activity (IARPA) or the U.S. Government.

8. References

[1] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Department of Computer Science, University of Waikato, 1999.

[2] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(8), p. 12261238, 2005.

[3] A. Krause and C. Guestrin, "Submodularity and its applications in optimized information gathering," *ACM Transactions on Intelligent Systems and Technology*, vol. 2(4), 2011.

[4] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proceedings of ACL*, 2011.

[5] A. Krause and V. Cevher, "Submodular dictionary selection for sparse representation," in *Proc. International Conference on Machine Learning (ICML)*, 2010.

[6] H. Lin and J. A. Bilmes, "How to select a good training-data subset for transcription: Submodular active selection for sequences," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, UK, September 2009.

[7] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using document summarization techniques for speech data subset selection," in *Proceedings of NAACL*, 2013.

[8] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," *arXiv preprint arXiv:1102.3975*, 2011.

[9] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular functions-I." *Math. Program.*, vol. 14, pp. 265–294, 1978.

[10] Y. Liu, K. Wei, K. Kirchhoff, Y. Song, and J. Bilmes, "Submodular feature selection for high-dimensional acoustic score spaces," in *Proceedings of ICASSP*, 2013.

[11] M. Minoux, "Accelerated greedy algorithms for maximizing submodular functions," in *Lecture Notes in Control and Information Sciences*, vol. 7, 1978, pp. 234–243.

[12] B. S. et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.