SUBMODULAR FEATURE SELECTION FOR HIGH-DIMENSIONAL ACOUSTIC SCORE SPACES

Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, Jeff Bilmes

Department of Electrical Engineering, University of Washington
Seattle, WA, 98195, USA

ABSTRACT

We apply methods for selecting subsets of dimensions from high-dimensional score spaces, and subsets of data for training, using submodular function optimization. Submodular functions provide theoretical performance guarantees while simultaneously retaining extremely fast and scalable optimization via an accelerated greedy algorithm. We evaluate this approach on two applications: data subset selection for phone recognizer training, and semi-supervised learning for phone segment classification. Interestingly, the first application uses submodularity twice: first for score space sub-selection and then for data subset selection. Our approach is computationally efficient but still consistently outperforms a number of baseline methods.

Index Terms— feature selection, Fisher kernel, acoustic similarity, graph-based learning, submodularity

1. INTRODUCTION

Generative acoustic score spaces, such as the Fisher score space [1], are constructed by training generative models on acoustic data and taking the derivative of the log-likelihood of the data with respect to the parameters of the models. They have found multiple uses in speech processing, including acoustic event classification [2], acoustic-phonetic classification [3], segmental minimum Bayes risk decoding [4], or speaker verification [5, 6]. Generalizations of Fisher score spaces for speech were explored in [7]. The dimensionality of these score space is often very high, however, and depend on the number of model parameters (which can be many millions). Many of these dimensions might be uninformative or redundant; it is hence desirable to select the best subset of features in some computationally feasible way. Previously proposed feature selection methods either do not scale to very high-dimensional feature spaces and/or large data sets, or they do not provide theoretical guarantees about their performance.

In this work, we present a feature selection method based on submodular function optimization that both provides theoretical performance guarantees and also scales easily to high-dimensional spaces. The method is applied to subselecting high-dimensional Fisher vectors for the purpose of computing $O(n^2)$ pairwise similarity scores between variable-length acoustic segments. These scores are then themselves used for two tasks: (a) to instantiate submodular functions for selecting subsets of training data; and (b) for semi-supervised graph-based learning for phonetic segment classification. Thus, in the first case, submodular functions are used twice (for both feature and data subset selection); in the second case they are used for feature subset selection only.

We demonstrate that our method outperforms standard baseline feature selection methods using mutual information between feature and class variables. In fact, our method is applicable to any feature selection problem and thus has broad implication for any pattern recognition task that involves high-dimensional feature spaces.

2. GENERATIVE SCORE SPACES

Given a sequence of acoustic feature vectors X and a generative model (such as an HMM) with parameter vector θ that models the underlying generation process, the Fisher score vector U_X is the vector of derivatives of the log-likelihood of X with respect to the parameters θ:

$$U_X^\theta = \nabla_\theta log P(X|\theta) \quad (1)$$

When several models are involved, the resulting vectors of derivatives are stacked to form the total Fisher score space:

$$U'_X = ((U_X^{\theta_1})^\intercal, (U_X^{\theta_2})^\intercal, ..., (U_X^{\theta_n})^\intercal)^\intercal \quad (2)$$

To compute a dissimilarity between two acoustic sequences i and j we take the dot-product of the Fisher score vectors, normalized by F the Fisher information matrix:

$$K_{i,j} = U'_i F^{-1} U'_j. \quad (3)$$

Other dissimilarity scores are also possible to compute, for example a norm (e.g., $d_{ij} = ||U'_i - U'_j||_q$ for some non-negative q). These can also be converted into similarity measures [8].

Depending on the number of models and parame-

ters per model, Fisher score vectors can be very high-dimensional. Many of the dimensions might either carry little information or be redundant with other dimensions. Previous applications have either selectively used the derivatives of only some parameters, such as only mean vectors or mixture weights of Gaussian mixtures; or they have applied binary compression [9]. A useful goal, therefore, is to select a subset of dimensions that retain as much informative as possible but that are non-redundant.

3. BACKGROUND

Submodular functions are a class of discrete functions that have the property of "diminishing returns." Given a finite set V, $f : 2^V \rightarrow \mathbb{R}$ is said to be submodular if $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$ holds $\forall A \subseteq B \subseteq V$ and $v \notin B$. In words, the incremental value (or "gain") of element v decreases as the context in which v is considered grows from A to B. Powerful guarantees exist for specific subtypes of submodular function optimization. For instance, a function is *monotone submodular* if $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B), \forall A \subseteq B, v \in V$. In such case, the problem of maximizing $f(S)$ subjected to a cardinality constraint can be approximately solved using a simple greedy algorithm that easily scales to extremely large data sets [10]. Submodular functions have already yielded superior results in various practical settings, such as environmental monitoring, activity recognition, sensor placement and information extraction [11], and document summarization [12].

Herein, we formulate our feature subset selection problem as monotone submodular function maximization subject either to a cardinality or to a knapsack constraint:

$$\arg\max_{S \subseteq V} \{f(S) : c(S) \leq K\} \tag{4}$$

where V is the set of features, K is max number of features to be selected, $c(\cdot) \geq 0$ is a feature cost (if $c(S) = |S|$, this constitutes a cardinality constraint, and if $c(S) = \sum_{s \in S} c(s)$ then this is a knapsack constraint), and the submodular function $f(\cdot)$ valuates the selected feature subset.

In this work, we evaluate two submodular functions each instantiated from $O(n^2)$ pairwise similarity scores — both, it turns out, significantly outperform standard baseline feature selection methods using mutual information between feature and class. The first objective is the uncapacitated facility location function defined as

$$f_1(S) = \sum_{i \in V} \max_{j \in S} w_{i,j}, \tag{5}$$

where $w_{i,j}$ is a similarity between feature i and j — this function valuates S by choosing and then accumulating for each $i \in V$ a single representative's similarity within S that is closest to i.

A second function we evaluate is the "saturated coverage function", defined as:

$$f_2(S) = \sum_{i \in V} \min\{C_i(S), \beta C_i(V)\}, \tag{6}$$

where $C_i(S) = \sum_{j \in S} w_{i,j}$ measures the degree that i is "covered" by S. $\beta \in [0, 1]$ is a hyperparameter that determines a global saturation threshold. The minimum within each term keeps features from being over-represented by subset S. Both objectives are monotone submodular functions, hence, the formulated optimization problem for feature selection can be solved near-optimally using a greedy algorithm. In fact, submodularity has another advantage, namely an accelerated greedy algorithm [13] having complexity $O(K \log |V|)$, allowing the scaling to very high-dimensional cases. A traditional greedy algorithm having complexity $O(K|V|)$ would not be as scalable to very large sizes.

Note that both f_1 and f_2 have been successfully used for extractive document summarization in the context of learning mixtures [14], and f_1 has been used for training data subset selection [15] with cardinality constraints. The novelty of this present work is as follows: f_1 has never before been used for training data subset selection under a knapsack contraints (we show that they both perform well). Moreover, never before has submodular subset selection been used simultaneously for acoustic score space selection and training data subset selection (we show this also performs quite well). Lastly, submodular subset selection for acoustic score space selection to produce similarity scores in a graph-based learning system is novel to this work as well — we show state-of-the-art results here as well.

4. TASKS, DATA AND BASELINE SYSTEMS

We evaluate our methods on TIMIT data. Two applications are considered: 1) data subset selection for phone recognizer training, and 2) semi-supervised learning for phone segment classification. The first application uses submodular optimization twice (for score space subset selection and for utterance subset selection for training analogous to [15]). Both applications make use of similarity scores between variable-length segments and necessitate the computation of the Fisher kernel.

4.1. Data Subset Selection

The goal of our first task is to identify a subset of the training data that provides as much information about the full data set as possible while being much smaller. We use a training set of 4620 utterances, the TIMIT core test set of 192 utterances, and 200 utterances from the remaining test set as development data. The data is preprocessed into 39-dimensional MFCC vectors extracted every 10ms. In order

to generate Fisher score vectors for the data subset selection task we train 16-component diagonal-covariance Gaussian mixture HMMs for each of the 48 phone classes in TIMIT. Derivatives are taken of all mean and variance vectors as well as the mixture weights, resulting in a score vector dimensionality of 186,577. Mean and variance normalization is applied to the Fisher score vectors to handle the different dynamic ranges of different dimensions. We then construct all 4620×4620 similarities between all pairs of training utterances. We use the facility location function with these similarities to select 2.5%, 5%, 10%, 20%, 30%, 40% and 50% and of the data (measured as percentage of non-silence speech frames). 3-state monophone HMMs with N-component Gaussian mixtures are trained for each of the 48 phone classes on the full training set as well as the subsets. N is optimized based on the size of the training subset (we set $N = 4, 8, 8, 16, 32, 32, 32$ respectively for each setting). Recognition accuracy is computed after mapping the 48 phones to 39 classes as described in [16].

4.2. Segment Classification

The second task consist of classifying the segments in the TIMIT core test set using the time boundaries given by the annotated labels (i.e. it is distinct from phone recognition). For this task we use a training set without the *sa* sentences (= 3686 utterances), the core test set, and a development set of 210 sentences. We use the standard phone set of 48 phones for training and collapse them to 39 classes for evaluation. Glottal stop segments are excluded. The total number of segments for training, development and testing is 121385, 7416, and 6589, respectively. Speaker-dependent mean and variance normalization of the acoustic features was applied. We use two semi-supervised graph-based learning approaches, label propagation (LP) [17] and measure propagation (MP) [18], both of which use pairwise similarity scores and were previously used for frame-based phone classification [19]. In this paper, we classify variable-length acoustic segments and thus use Fisher kernels. We utilize the Fisher scores generated by a baseline HMM system, comprised of 48 3-state HMMs with 16 Gaussian mixture components with diagonal covariance matrices per phone class. The dimensionality of the score vectors is 182,017. Our key goal (which we have achieved below) is to demonstrate the viability of submodular feature selection to improve Fisher kernel based similarity measures. In order to compute the Fisher kernel we stack the derivatives of all Gaussian components into one vector per segment; Eq. (3) is then computed on these vectors. We do not use the full Fisher information matrix to reduce computational demands. We also normalize the values as: $\tilde{K}_{ij} = K_{ij}/\sqrt{K_{ii}K_{jj}}$. Using this similarity measure, 10-nearest neighbor graphs were constructed for use by the two graph-based learning algorithms.

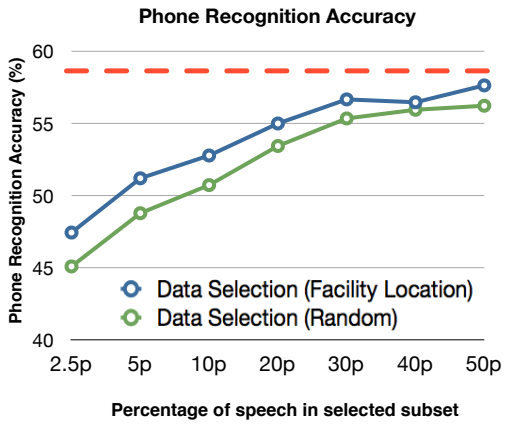

Fig. 1. Phone accuracy rates for random subset selection and submodular subset selection using the entire Fisher score space. All (100%) of data (red), facility location (blue), and average (of 100) random selection (green).

5. EXPERIMENTS AND RESULTS

For feature subset selection, we let $V = \{v_1, v_2, \cdots, v_n\}$ be the set of all features in the Fisher score vectors. We first prune away an initial feature set P such that $\forall i \in P, I(v_i; C) < \tau$ where $I(v_i; C)$ is the mutual information between feature i and the class variable C, ranging over all phone classes, and τ is a cutoff threshold (0.01 in our case). Then we build a fully connected graph using the remaining set of features $V \setminus P$. In the data selection task, the remaining set has 90,422 features; in the classification task, it has 73,978 features. Graph edges are weighted with the mutual information $I(v_i; v_j)$ score for pairs v_i, v_j of features. We then select subsets of this set using the accelerated greedy algorithm using either the facility location or the saturated coverage functions. We compare against the baseline method of computing mutual information between **only** the feature and class variable, ranking features by that score and selecting the top N features. This procedure is only *modular* and does not allow for any direct interaction between the features (unlike submodular functions which do). In order to compute mutual information on continuous Fisher scores, they are first quantized into 50 equal-width bins.

For the data subset selection task we compare three experimental conditions: (a) the full Fisher vectors are used for constructing the utterance similarity graph; (b) the Fisher vectors are reduced by modular feature selection; (c) the Fisher vectors are reduced by submodular feature selection. For the reduced vectors we used 1k, 2k, 5k,

	2.5p	5p	10p	20p	30p
1k-s	**1.99%**	**2.17%**	**2.48%**	**1.48%**	**1.93%**
1k-m	-0.20%	-0.49%	1.48%	0.52%	1.52%
2k-s	**0.11%**	**1.60%**	**2.33%**	**1.46%**	**1.97%**
2k-m	0.04%	-0.59%	0.45%	1.24%	1.45%
5k-s	0.84%	1.15%	**2.64%**	**1.12%**	**1.37%**
5k-m	**1.06%**	**1.46%**	-0.41%	-0.30%	-0.18%
10k-s	**5.21%**	**3.05%**	**3.08%**	**1.16%**	**2.84%**
10k-m	2.79%	1.29%	1.03%	0.97%	0.92%
20k-s	**6.45%**	**3.36%**	**4.34%**	**2.58%**	**1.46%**
20k-m	3.77%	2.34%	3.61%	0.79%	1.16%
50k-s	**4.79%**	**5.27%**	**3.92%**	**2.79%**	**2.71%**
50k-m	2.55%	3.85%	2.48%	1.38%	2.11%
all	5.21%	4.96%	4.04%	2.92%	2.38%

Table 1. Using the saturated coverage function to select features in Fisher kernel vectors. In 28 out of 30 settings, submodular feature selection leads to a greater improvement over the baseline, and even outperforms it in 5 settings.

	400		800		2k		4k		10k	
	LP	MP	LP	MP	LP	MP	LP	MP	LP	MP
mod	42.07	42.95	63.83	64.23	62.80	64.09	68.05	68.99	63.48	64.30
sub	**66.46**	**67.16**	**69.43**	**70.45**	**68.87**	**69.25**	**69.24**	69.48	**67.04**	**67.14**

Table 2. Accuracy rates for segment classification, with modular (mod) and submodular (sub) feature selection. The baseline model (monophone HMMs, without graph-based learning) has a classification accuracy of 68.02%. Bold-face numbers are significant ($p < 0.05$) improvements over the modular MI-based method.

10k, 20k, and 50k. For each feature subset we re-build the graph using the reduced feature set, and then follow the accelerated greedy procedure mentioned above. The comparison of relative improvements over the random selection baseline is shown in Table 1 (Figure 1 shows absolute results when the entire score space is used). We see that in all these cases the submodular feature selection method outperforms the modular feature selection method in almost all cases and even outperforms the full feature vector under some conditions.

Table 2 shows the results of the segment classification experiments. Again, we compare submodular feature selection against modular selection. Here, the baseline HMMs have an accuracy of 68.02%. With only 400 features, the modular MI-based selection leads to a significant drop in performance, resulting in $\sim 42\%$ classification accuracy; however, using the top 400 features selected from the submodular method we achieve $\sim 67\%$. With 800 features selected by the submodular method, we obtain a significant improvement over the baseline model with almost an order of magnitude fewer features. This shows that our method based on submodular optimization is highly effective in selecting the most useful features. Importantly, the selection procedures can be done quite rapidly.

6. RELATED WORK

Numerous methods have previously been proposed for feature selection. Two previous methods related to ours are correlation-based feature subset selection [20] and maximum-relevance-minimum-redundancy [21]. Both of them aim at selecting subsets that provide information about the class while minimizing the redundancy among features in the subset. We attempted to compare against [21] (using the C++ implementation provided by the authors) and [20] as implemented in the WEKA package. However, both methods have quadratic complexity and use traditional (non-accelerated) greedy algorithms for optimization and do not scale to feature sets of our size. Since neither of the two methods use a monotone submodular objective, the accelerated greedy algorithm [13] is inapplicable.

7. CONCLUSIONS

We have presented a novel submodular feature selection method that scales to high-dimensional spaces and provides optimality guarantees. On both a data subset selection and a segment classification task it was shown to outperform baseline modular feature selection; in some cases it even outperformed the full feature vector. The method is general and applies to large feature sets in a variety of application; it is not restricted to acoustic score spaces. Future work will involve further applications on a variety of big data sets and the investigation of other submodular functions.

Acknowledgment This material is 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] T. Jaakola and D. Haussler, "Exploiting generative models in discriminative classifiesr," in *Advances in Neural Information Processing Systems 11*, 1999, pp. 487–493.

[2] A. Temko, E. Monte, and C. Nadeu, "Comparison of sequence discriminant support vector machines

for acoustic event classification," in *Proceedings of ICASSP*, 2006.

[3] N.D. Smith, M.J.F. Gales, and M. Niranjan, "Data-dependent kernels in SVM classification of speech patterns," Tech. Rep. CUED/F-INFENG/TR387, Cambridge University Eng. Dept., 2001.

[4] V. Venkata Ramani and B. Byrne, "Support vector machines for segmental minimum Bayes risk decoding of continuous speech," in *Proceedings of ASRU*, 2003, pp. 1–6.

[5] V. Wan and S. Renals, "Evaluation of kernel methods for speaker verification," in *Proceedings of ICASSP*, 2002.

[6] J. Mariéthoz, Y. Granvalez, and S. Bengio, "Kernel-based text-independent speaker verification," in *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, J. Keshet and S. Bengio, Eds. 2009, pp. 197–223, Wiley.

[7] M. Layton and M. Gales, "Acoustic modeling using continuous rational kernels," in *Proceedings of MLMSP*, 2005.

[8] M.M. Deza and E. Deza, *Encyclopedia of distances*, Springer, 2009.

[9] F. Peronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proceedings of CVPR*, 2010, pp. 3384–3391.

[10] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher, "An analysis of approximations for maximizing submodular set functions i," *Mathematical Programming*, vol. 14, pp. 265–294, 1978.

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

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

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

[14] Hui Lin and Jeff Bilmes, "Learning mixtures of submodular shells with application to document summarization," in *Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, USA, July 2012, AUAI.

[15] Hui Lin and Jeff 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.

[16] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using Hidden Markov Models," *IEEE Trans. ASSP*, vol. 37, pp. 1641–1648, 1989.

[17] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., CMU-CALD-02, 2002.

[18] A. Subramanya and J. Bilmes, "Semi-supervised learning with measure propagation," Tech. Rep. UWEE-TR-2010-0004, University of Washington, 2010.

[19] A. Alexandrescu and K. Kirchhoff, "Phonetic classification by controlled random walks," in *Proceedings of Interspeech*, 2011.

[20] M.A. Hall, *Correlation-based Feature Selection for Machine Learning*, Ph.D. thesis, Department of Computer Science, University of Waikato, 1999.

[21] H.C. 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), pp. 12261238, 2005.