Orthogonal Matching Pursuit for Text Classification

Konstantinos Skianis, Nikolaos Tziortziotis, Michalis Vazirgiannis
LIX, École Polytechnique, France

Introduction

Text is hard:

- high dimensionality of text → overfitting remains
- not all words are useful → sparsity

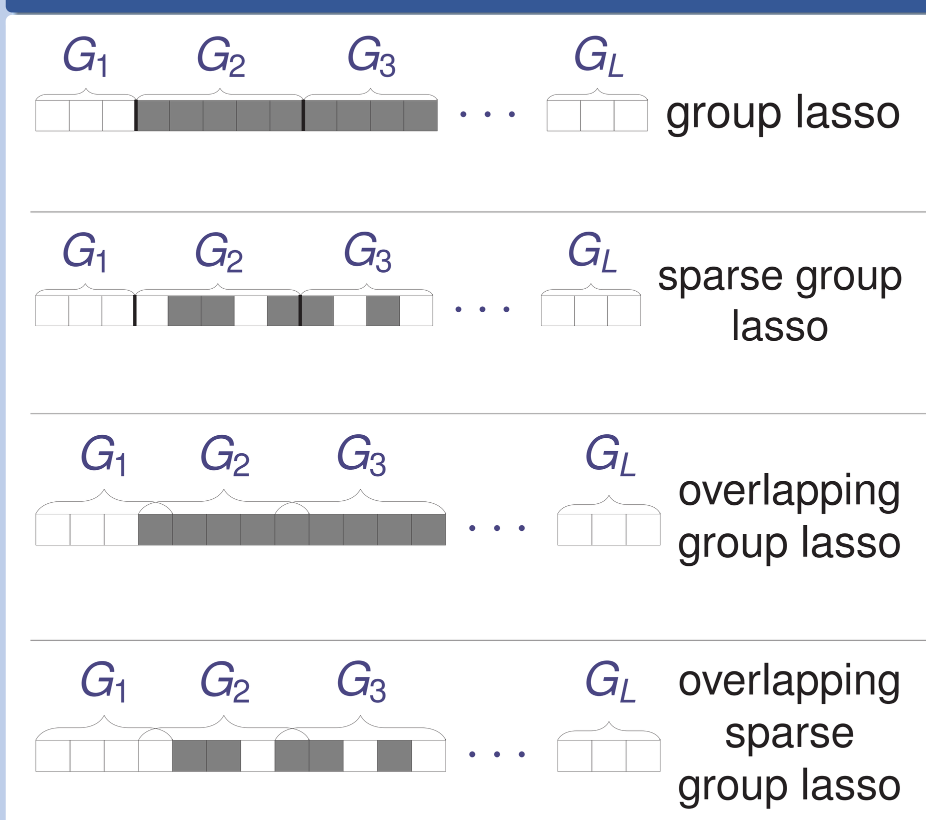
Regularization:

- critical for text classification, opinion mining, noisy text normalisation
- group lasso can fail to create sparse models
- groups are not always available

Contribution:

1. apply OMP to text classification;
2. introduce overlapping GOMP, moving from disjoint to overlapping groups;
3. analyze their efficiency in accuracy and sparsity (vs. group lasso & deep learning).

I. Structured Regularization



Where?

- removing unnecessary words along with their weights
- Text normalization → machine learning problem (Ikeda, Shindo, and Matsumoto 2016)

Methods

- ℓ_1, ℓ_2, Elastic net regularization
- Group lasso (Yuan and Lin 2006)
- Linguistic structured regularization (Yogatama and Smith 2014)

II. Orthogonal Matching Pursuit

Algorithm Logistic Overlapping GOMP

Input: $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times d}$, $\boldsymbol{y} \in \{-1, 1\}^N$, $\{G_1, \dots, G_J\}$ (groups), K (budget), ϵ (precision), λ.

Initialize: $\mathcal{I} = \emptyset$, $\boldsymbol{r}^{(0)} = \boldsymbol{y}$, $k = 1$;

while $|\mathcal{I}| \leq K$ **do**
  $j^{(k)} = \arg\max_j \frac{1}{|G_j|} \left\| X_{G_j}^\top \boldsymbol{r}^{(k-1)} \right\|_2^2$
  break if $\left\| X_{G_{j^{(k)}}}^\top \boldsymbol{r}^{(k-1)} \right\|_2^2 \leq \epsilon$
  $\mathcal{I} = \mathcal{I} \cup \{G_{j^{(k)}}\}$
  for $i = 1$ **to** J **do**
    $G_i = G_i \setminus G_{j^{(k)}}$
  end for
  $\boldsymbol{\theta}^{(k)} = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^N \mathcal{L}(\boldsymbol{x}_i, \boldsymbol{\theta}, y_i) + \lambda \|\boldsymbol{\theta}\|_2^2$ s.t. $supp(\boldsymbol{\theta}) \subseteq \mathcal{I}$
  $\boldsymbol{r}^{(k)} = \frac{1}{1+\exp\{-X\boldsymbol{\theta}^{(k)}\}} - \mathbb{1}_{\{\boldsymbol{y}\}}$
  $k \mathrel{+}= 1$
end while

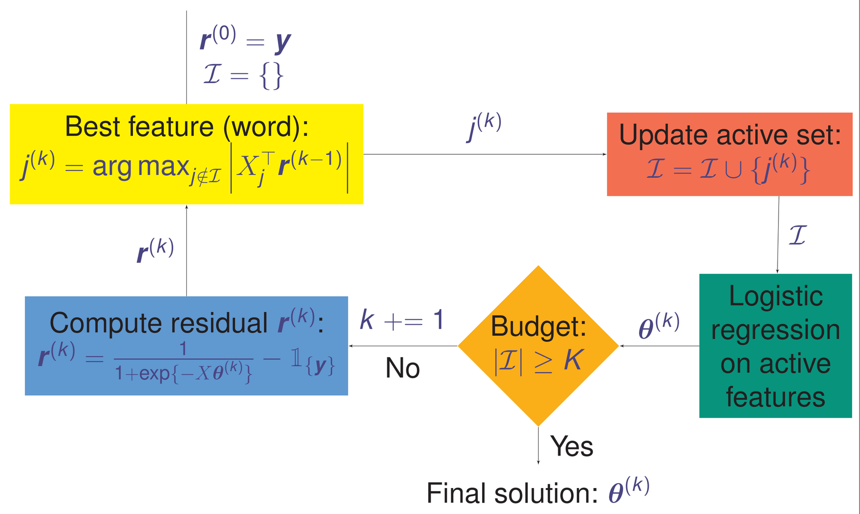


Figure: $X \in \mathbb{R}^{N \times d}$: design matrix, $\boldsymbol{y} \in \mathbb{R}^N$: response vector, K: budget, $\mathcal{I}$: set of active features.

III. Datasets & Setup

Data

- Topic categorization on 20NG dataset
  - Four binary classification tasks
- Sentiment analysis
  - Floor speeches by U.S. Congressmen deciding "yea"/"nay" votes on the bill under discussion (Thomas, Pang, and Lee 2006)
  - Movie reviews (Pang and Lee 2004)
  - Product reviews from Amazon (Blitzer, Dredze, and Pereira 2007)

	dataset	train	dev	test	# words	# sents
20NG	science	949	238	790	25787	16411
	sports	957	240	796	21938	14997
	religion	863	216	717	18822	18853
	comp.	934	234	777	16282	10772
Sentiment	vote	1175	257	860	19813	43563
	movie	1600	200	200	43800	49433
	books	1440	360	200	21545	13806
	dvd	1440	360	200	21086	13794
	electr.	1440	360	200	10961	10227
	kitch.	1440	360	200	9248	8998

Table: Descriptive statistics of the datasets

Settings

- Parameter tuning on development set
- Minibatch K-Means clustering on word2vec with max 2000 clusters.

IV. Results

	dataset	no reg.	lasso	ridge	elastic	OMP	group lasso LDA	LSI	sen	GoW	w2v	GOMP
20NG	science	0.946	0.916	0.954	0.954	0.964*	**0.968**	**0.968***	0.942	0.967*	**0.968***	0.953*
	sports	0.908	0.907	0.925	0.920	0.949*	0.959	0.964*	**0.966**	0.959*	0.946*	0.951*
	religion	0.894	0.876	0.895	0.890	0.902*	0.918	0.907*	**0.934**	0.911*	0.916*	0.902*
	computer	0.846	0.843	0.869	0.856	0.876*	0.891	0.885*	0.904	0.885*	**0.911***	0.902*
Sentiment	vote	0.606	0.643	0.616	0.622	0.684*	**0.658**	0.653	0.656	0.640	0.651	**0.687***
	movie	0.865	0.860	0.870	0.875	0.860*	**0.900**	0.895	0.895	0.895	0.890	0.850
	books	0.750	0.770	0.760	0.780	0.800	0.790	0.795	0.785	0.790	0.800	**0.805***
	dvd	0.765	0.735	0.770	0.760	0.785	0.800	0.805*	0.785	0.795*	0.795*	**0.820***
	electr.	0.790	0.800	0.800	0.825	**0.830**	0.800	0.815	0.805	0.820	0.815	0.800
	kitch.	0.760	0.800	0.775	0.800	0.825	0.845	**0.860***	0.855	0.840	0.855*	0.830

Table: Accuracy in test subsets. *: statistical significance over lasso at $p < 0.05$ using micro sign test.

	Dataset	CNN (20eps)	FastText (100eps)	Best OMP or GOMP	Best Lasso
20NG	science	0.935	0.958	0.964	**0.968**
	sports	0.924	0.935	0.951	**0.966**
	religion	**0.934**	0.898	0.902	**0.934**
	computer	0.885	0.867	0.902	**0.911**
Sentiment	vote	0.651	0.643	**0.687**	0.658
	movie	0.780	0.875	0.860	**0.900**
	books	0.742	0.787	**0.805**	0.800
	dvd	0.732	0.757	**0.820**	0.805
	electr.	0.760	0.800	**0.830**	0.820
	kitch.	0.805	0.845	0.830	**0.860**

Table: Comparison with state-of-the-art classifiers: CNN (Kim 2014), FastText (Joulin et al. 2017) with no pre-trained vectors.

	dataset	no reg.	lasso	ridge	elastic	OMP	group lasso LDA	LSI	sen	GoW	w2v	GOMP
20NG	science	100	**1**	100	63	2.7	19	20	86	19	21	5.8
	sports	100	**1**	100	5	1.8	60	11	6.4	55	44	7.7
	religion	100	**1.1**	100	3	1.5	94	31	99	10	85	1.5
	computer	100	1.6	100	7	**0.6**	40	35	77	38	18	4.9
Sentiment	vote	100	**0.1**	100	8	5	15	16	13	97	13	1.5
	movie	100	1.3	100	59	**0.9**	72	81	55	90	62	2.3
	books	100	**3.3**	100	14	4.6	41	74	72	90	99	8.3
	dvd	100	**2**	100	28	2.8	64	8	8	58	64	9
	electr.	100	**4**	100	6	6.3	10	8	43	8	9	12
	kitch.	100	4.5	100	79	**4.3**	73	44	27	75	46	6.5

Table: Fraction (in %) of non-zero feature weights in each model for each dataset. Bold for best, blue for best group.

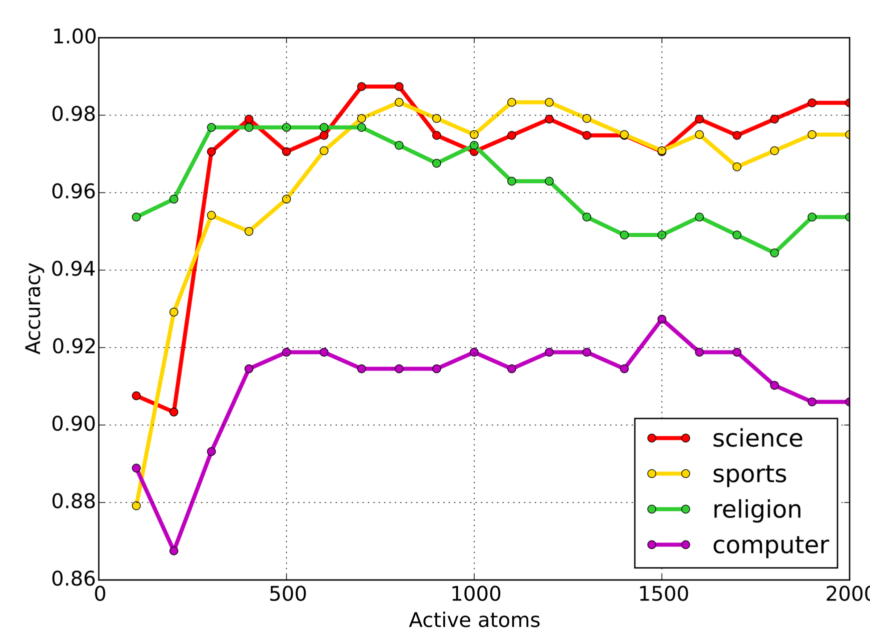


Figure: Accuracy vs. number of active atoms/features for OMP.

V. Discussion & Future Work

- Group based regularizers **better** than the baseline ones.
- GOMP requires some "good" groups along with single features.

Conclusion

- Introduce OMP and GOMP for the text classification task
- Extending the standard GOMP algorithm was also proposed, which is able to handle overlapping groups
- Simple (greedy feedforward feature selection) → accurate models with high sparsity

Future Work

- Examine the theoretical properties of overlapping GOMP
- Learning automatically the groups → Simultaneous OMP (Szlam, Gregor, and LeCun 2012)
- Sparse Group OMP

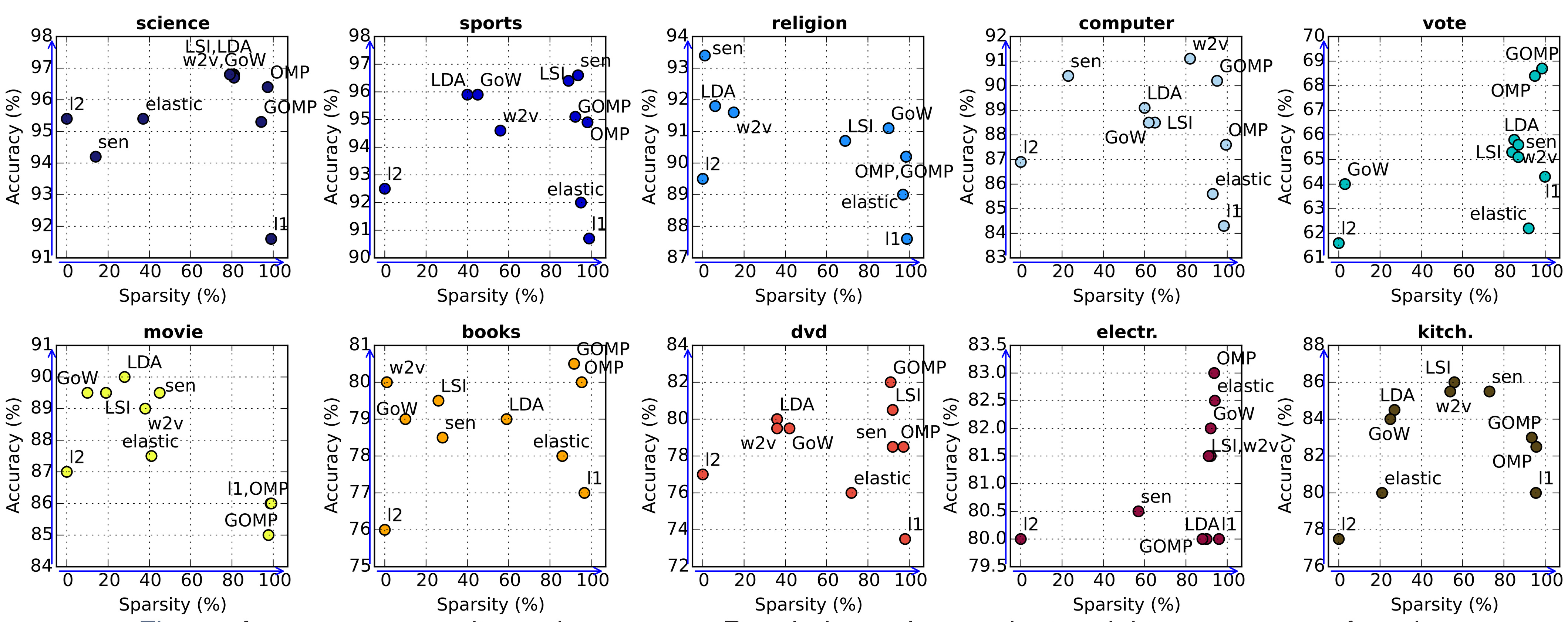


Figure: Accuracy vs sparsity on the test sets. Regularizers close to the top right corner are preferred.

4th Workshop on Noisy User-generated Text (W-NUT), EMNLP, 2018
kskianis@lix.polytechnique.fr