Hybrid SVD for Text Mining

Anastasiia Baiandina   Iurii Kolomeitsev

Skolkovo Institute of Science and Technology

NLA/Optimization Course Project

Document classification

Document-term matrix

matrix of weighted word occurrences in documents (e.g. TF-IDF)

- sparse
- high-dimensional
- low-rank

⇒ dimensionality reduction using Singular Value Decomposition (e.g. Latent Semantic Analysis)

- words in different documents share their meaning
- we may know relations between documents

⇒ incorporate additional information in SVD

Problem Statement

Notation

$R \in \mathbb{R}^{D \times T}$ — document-term matrix

$K \in \mathbb{R}^{D \times D}$ — document similarity matrix

$S \in \mathbb{R}^{T \times T}$ — term similarity matrix

Model

$$A = RR^T = DCD$$

$$c_{ij} = \cos(i,j) \sim r_i^T r_j \quad \Rightarrow \quad \mathrm{sim}(i,j) \sim r_i^T S r_j$$

$$\begin{cases} RSR^T = U\Sigma^2 U^T \\ R^T K R = V\Sigma^2 V^T \end{cases} \quad \Rightarrow \quad \hat{R} = K^{\frac{1}{2}} R S^{\frac{1}{2}} = \hat{U}\Sigma\hat{V}^T$$

$\hat{U} = K^{\frac{1}{2}} U$, $\hat{V} = S^{\frac{1}{2}} V$ — matrices with orthonormal columns

$\Sigma \in \mathbb{R}^{r \times r}$ — diagonal matrix with first r principal values

Computation

Model

$K^{\frac{1}{2}} R S^{\frac{1}{2}} = \hat{U} \Sigma \hat{V}^T$

Similarity

require K, S to be diagonal dominant

$$K = I + \alpha K', \quad S = I + \beta S',$$

where K', S' — original zero-diagonal similarity matrices

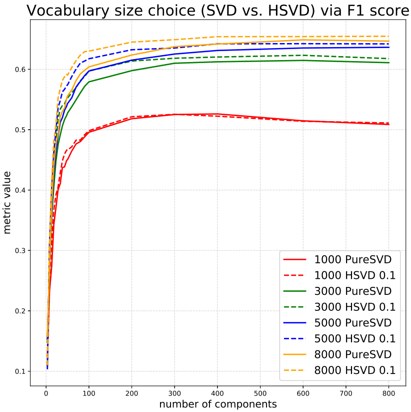
$\Rightarrow$ square root replaced with Cholesky decomposition
$K = L_k L_k^T$, $S = L_s L_s^T$

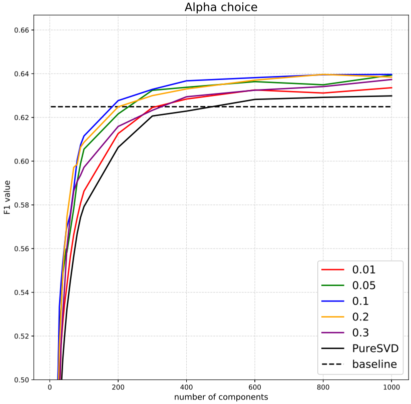
20 Newsgroups

dataset	num docs	avg doc len	initial sparsity, %	sparsity, %
20 Newsgroups	18846	181.6	0.066	0.858

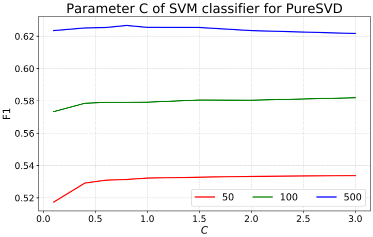
- language: English
- 20-class classification: news topics
- term similarity: cosine between FastText word representations
- classifier: linear SVM
- baseline: on the full TF-IDF matrix

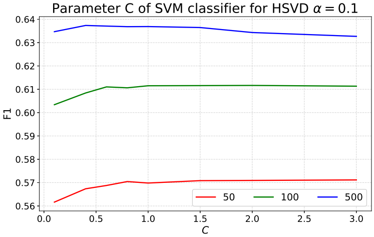
20 Newsgroups

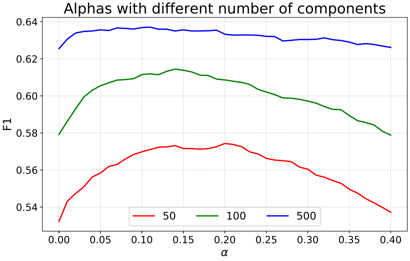

Vocabulary size choice (SVD vs. HSVD) via F1 score

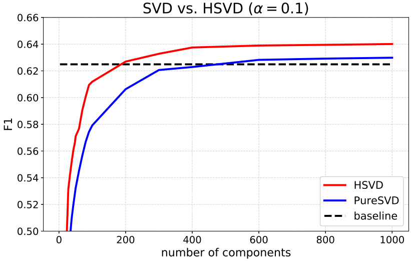

Alpha choice

20 Newsgroups


Parameter C of SVM classifier for PureSVD


Parameter C of SVM classifier for HSVD $\alpha = 0.1$


Alphas with different number of components

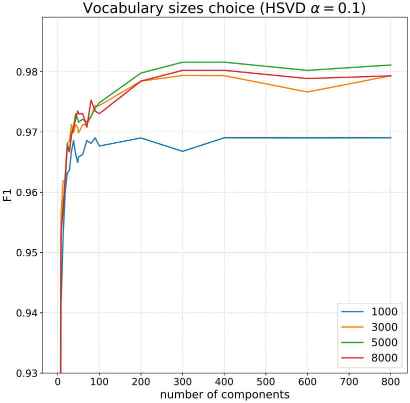

SVD vs. HSVD ($\alpha = 0.1$)

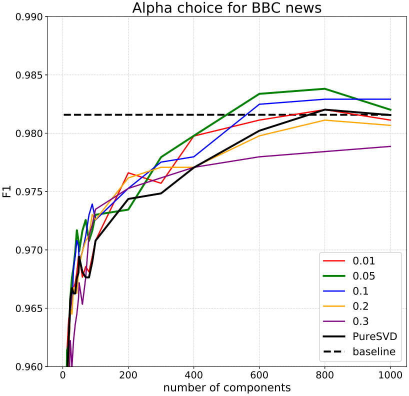
BBC News

dataset	num docs	avg doc len	initial sparsity, %	sparsity, %
BBC News	2225	2274.7	0.504	3.318

- language: English
- 5-class classification: BBC news topics
- term similarity: cosine between FastText word representations
- classifier: linear SVM
- baseline: on the full TF-IDF matrix

BBC news


Vocabulary sizes choice (HSVD $\alpha = 0.1$)
F1
0.93
0.94
0.95
0.96
0.97
0.98
0
100
200
300
400
500
600
700
800
number of components
1000
3000
5000
8000

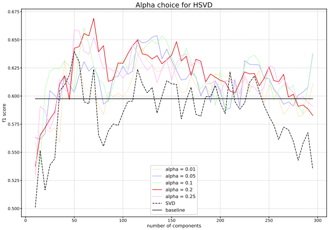

Alpha choice for BBC news
F1
0.960
0.965
0.970
0.975
0.980
0.985
0.990
0
200
400
600
800
1000
number of components
0.01
0.05
0.1
0.2
0.3
PureSVD
baseline

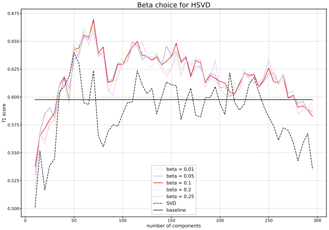
Paper Reviews

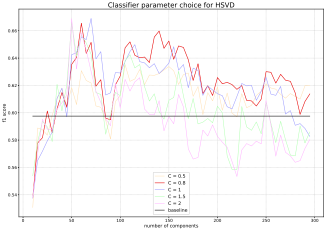
dataset	num docs	avg doc len	initial sparsity, %	sparsity, %
Paper Reviews	388	66	1.104	1.33

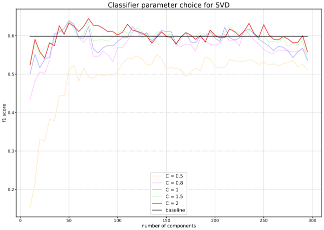
- language: Spanish
- binary classification: whether the review is positive or negative
- vocabulary size: 5000
- term similarity: cosine between word2vec word representations
- document similarity: two reviews are considered similar if they refer to the same article
- classifier: linear SVM
- baseline: SVM on the full TF-IDF matrix

Paper Reviews

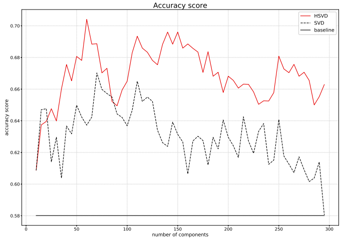




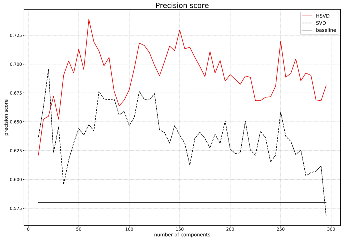


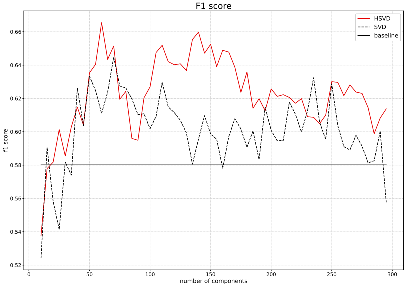


Paper Reviews



	SVD	HSVD
accuracy	0.670 ± 0.079	0.704 ± 0.066
precision	0.676 ± 0.111	0.740 ± 0.110
f1	0.645 ± 0.086	0.665 ± 0.072





Summary

- hybrid SVD model incorporating side information
- the model has been tested on the datasets from different application domains
- the model outperforms baseline and SVD in all cases

Future Work

- explore different term similarity measures
- develop approaches to the other text mining problems (e.g. clustering, comparison)
- work on the modifications of folding-in
- end-to-end solution where S and K are part of optimization process

References

A. N. Nikolakopoulos, V. Kalantzis and J. D. Garofalakis, EIGENREC: An Efficient and Scalable Latent Factor Family for Top-N Recommendation. arXiv preprint arXiv:1511.06033, 2015.

E. Frolov and I. Oseledets, PureSVD with Side Information for top-N Recommendations, *in process*, 2017.