Yue-peng ZOU, Ji-hong OUYANG, Xi-ming LI, 2018. Supervised topic models with weighted words: multi-label document classification. *Frontiers of Information Technology & Electronic Engineering*, 19(4):513-523.
https://doi.org/10.1631/FITEE.1601688

Supervised topic models with weighted words: multi-label document classification

Key words: Supervised topic model; Multi-label classification; Class frequency; Labeled latent Dirichlet allocation (L-LDA); Dependency-LDA

Corresponding author: Xi-ming LI
E-mail: liximing86@gmail.com
ORCID: https://orcid.org/0000-0001-8190-5087

Motivation

In supervised topic modeling for multi-label classification, each word has equal importance. This is problematic for classification. More specifically, the class frequency information for words—the number of classes where a word has occurred in the training data—is neglected.

Main idea

1. The class frequency is a discriminative knowledge for classification. Inspired by this, we propose a word weighting method, namely class frequency weight (CF-weight).

2. The proposed CF-weight can be easily applied to existing supervised topic models. In this work, we apply CF-weight to L-LDA and Dependency-LDA.

Method

1. Design a computation of CF-weight, which provides low/high class frequency words with large/small weights.

2. Apply CF-weight to L-LDA and Dependency-LDA, and use two popular algorithms, i.e., variational inference and Gibbs sampling, to train models.

Major results

Compared to baseline models, our models can achieve higher performance in multi-label classification.

Table 4 Experimental performance of Micro-F1

Dataset	Micro-F1						
	WL-LDA$_V$	WL-LDA$_G$	L-LDA	WD-LDA	Dep-LDA	SVMs	RAkLE
Yahoo! arts	0.451 ±0.0039	0.437 ±0.0019	0.412 ±0.0059	**0.468 ±0.0014**	0.454 ±0.0031	0.435 ±0.0025	0.446 ±0.0121
Yahoo! health	0.603 ±0.0058	0.591 ±0.0096	0.578 ±0.0129	**0.626 ±0.0073**	0.616 ±0.0098	0.617 ±0.0132	0.614 ±0.0095
Medical	0.795 ±0.0108	0.782 ±0.0091	0.769 ±0.0139	**0.805 ±0.0125**	0.792 ±0.0174	0.791 ±0.0102	0.793 ±0.0121
Enron	0.393 ±0.0168	0.401 ±0.0124	0.381 ±0.0228	0.523 ±0.0114	0.514 ±0.0132	0.526 ±0.0183	**0.554 ±0.0091**
Rcv1subset1	0.257 ±0.0034	0.248 ±0.0085	0232 ±0.0101	**0.269 ±0.0028**	0.248 ±0.0084	0.258 ±0.0045	0.239 ±0.0095
Bibtex	0.376 ±0.0065	0.369 ±0.0103	0.361 ±0.0085	**0.411 ±0.0021**	0.375 ±0.0025	0.398 ±0.0046	0.404 ±0.0068
Bookmarks	0.225 ±0.0113	0.218 ±0.0057	0.193 ±0.0124	**0.239 ±0.0065**	0.211 ±0.0095	0.212 ±0.0086	0.216 ±0.0093

Dep-LDA is the abbreviation for dependence-LDA. The best results of each dataset are in boldface

Major results

Table 5 Experimental performance of Macro-F1

Dataset	Macro-F1						
	WL-LDA$_V$	WL-LDA$_G$	L-LDA	WD-LDA	Dep-LDA	SVMs	RAkLE
Yahoo! arts	0.302	0.306	0.287	**0.322**	0.318	0.301	0.314
	±0.0017	±0.0033	±0.0014	**±0.0029**	±0.0032	±0.0053	±0.0025
Yahoo! health	0.271	0.269	0.218	**0.319**	0.310	0.288	0.289
	±0.0051	±0.0067	±0.0083	**±0.0153**	±0.0143	±0.0072	±0.0059
Medical	0.341	0.334	0.329	0.356	0.324	0.358	**0.369**
	±0.0242	±0.0153	±0.0295	±0.0192	±0.0348	±0.0246	**±0.0183**
Enron	0.123	0.127	0.094	**0.149**	0.113	0.142	0.147
	±0.0097	±0.0137	±0.0101	**±0.0057**	±0.0062	±0.0089	±0.0064
Rcv1subset1	0.136	0.132	0.131	**0.142**	0.133	0.139	0.137
	±0.0057	±0.0072	±0.0094	**±0.0068**	±0.0102	±0.0088	±0.0053
Bibtex	0.272	0.283	0.221	**0.301**	0.293	0.282	0.268
	±0.0062	±0.0048	±0.0083	**±0.0037**	±0.0049	±0.0068	±0.0054
Bookmarks	0.114	0.115	0.098	**0.132**	0.112	0.082	0.088
	±0.0086	±0.0052	±0.0136	**±0.0084**	±0.0102	±0.0077	±0.0126

Dep-LDA is the abbreviation for dependence-LDA. The best results of each dataset are in boldface

Conclusions

1. Propose a novel CF-weight for supervised topic models in multi-label classification.

2. Apply CF-weight to L-LDA and Dependency-LDA.

3. Experimental results show that our models can perform better than existing supervised topic models.