

Statistical topic models for multi-label document classification

Timothy N. Rubin · America Chambers ·
Padhraic Smyth · Mark Steyvers

Received: 1 October 2010 / Accepted: 9 November 2011 / Published online: 29 December 2011
© The Author(s) 2011

Abstract Machine learning approaches to multi-label document classification have to date largely relied on discriminative modeling techniques such as support vector machines. A drawback of these approaches is that performance rapidly drops off as the total number of labels and the number of labels per document increase. This problem is amplified when the label frequencies exhibit the type of highly skewed distributions that are often observed in real-world datasets. In this paper we investigate a class of generative statistical topic models for multi-label documents that associate individual word tokens with different labels. We investigate the advantages of this approach relative to discriminative models, particularly with respect to classification problems involving large numbers of relatively rare labels. We compare the performance of generative and discriminative approaches on document labeling tasks ranging from datasets with several thousand labels to datasets with tens of labels. The experimental results indicate that probabilistic generative models can achieve competitive multi-label classification performance compared to discriminative methods, and have advantages for datasets with many labels and skewed label frequencies.

Keywords Topic models · LDA · Multi-label classification · Document modeling · Text classification · Graphical models · Probabilistic generative models · Dependency-LDA

Editors: Grigorios Tsoumakas, Min-Ling Zhang, and Zhi-Hua Zhou.

T.N. Rubin (✉) · M. Steyvers
Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697, USA
e-mail: trubin@uci.edu

M. Steyvers
e-mail: mark.steyvers@uci.edu

A. Chambers · P. Smyth
Department of Computer Science, University of California, Irvine, Irvine, CA 92697, USA

A. Chambers
e-mail: ahollowa@uci.edu

P. Smyth
e-mail: smlyth@ics.uci.edu