Statistics > Machine Learning

Statistical Topic Models for Multi-Label Document Classification

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

(Submitted on 13 Jul 2011 (v1), last revised 10 Nov 2011 (this version, v2))

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.

Comments: 44 Pages (Including Appendices). To be published in: The Machine Learning Journal, special issue on Learning from Multi-Label Data. Version 2 corrects some typos, updates some of the notation used in the paper for clarification of some equations, and incorporates several relatively minor changes to the text throughout the paper

Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)

Cite as: **arXiv:1107.2462 [stat.ML]**

(or **arXiv:1107.2462v2 [stat.ML]** for this version)

Submission history

From: Timothy Rubin [view email]

[v1] Wed, 13 Jul 2011 04:28:32 GMT (958kb,D)

[v2] Thu, 10 Nov 2011 04:24:38 GMT (958kb,D)

Which authors of this paper are endorsers?