LF-LDA:
A Supervised Topic Model for Multi-Label Documents Classification

Yongjun Zhang, Huaiyin Institute of Technology, Huaian, China & College of Computer and Information, Hohai University, Nanjing, China
Zijian Wang, College of Computer and Information, Hohai University, Nanjing, China
Yongtao Yu, Huaiyin Institute of Technology, Huaian, China
Bolun Chen, Huaiyin Institute of Technology, Huaian, China
Jialin Ma, Huaiyin Institute of Technology, Huaian, China & College of Computer and Information, Hohai University, Nanjing, China
Liang Shi, Jiangsu Vocational College of Business, Nantong, China

ABSTRACT
This article describes how text documents are a major data structure in the era of big data. With the explosive growth of data, the number of documents with multi-labels has increased dramatically. The popular multi-label classification technology, which is usually employed to handle multinomial text documents, is sensitive to the noise terms of text documents. Therefore, there still exists a huge room for multi-label classification of text documents. This article introduces a supervised topic model, named labeled LDA with function terms (LF-LDA), to filter out the noisy function terms from text documents, which can help to improve the performance of multi-label classification of text documents. The article also shows the derivation of the Gibbs Sampling formulas in detail, which can be generalized to other similar topic models. Based on the textual data set RCV1-v2, the article compared the proposed model with other two state-of-the-art multi-label classifiers, Tuned SVM and labeled LDA, on both Macro-F1 and Micro-F1 metrics. The result shows that LF-LDA outperforms them and has the lowest variance, which indicates the robustness of the LF-LDA classifier.

KEYWORDS
Function Terms, Gibbs Sampling, Graph Model, Multi Label, Parameter Estimation, Probability Generation Process, Text Classification, Topic Model

1. INTRODUCTION
With the development of Internet information technology, there is an explosive growth of data. Among the massive data, text documents are a major data structure. Effectively mining the knowledge behind text documents has become popular and attracts more and more attention. For labeled text documents, automatic text categorization is a powerful technology to process them. Traditional single label text classification assumes that each text document can be classified into only one label. Some state-of-the-art text categorization technologies, such as SVM, can achieve an accuracy over 90%. However, text documents are usually associated with multiple topics. For example, if a document focusing on transgenic technology may be mainly related to the technological topic, but may also be relevant to...
the health topic. The multi-topic properties of text documents make the traditional single label text classification technologies ineffective when they are applied to text documents with multi-labels (Yamamoto and Satoh, 2014). At present, multi-label text classification is a challenging task, and there is a huge room remained for improving its performance, especially for the text documents with unbalanced labels (Hmeidi et al., 2016).

To our knowledge, multi-label classification algorithms are usually extended from single label algorithms. According to the expansion way, they can be divided into two categories (Tsoumakas et al., 2009): algorithm adaption (AA) and problem transformation (PT). The AA algorithms reform the existing label classification algorithm to make them capable to be applied to multi-label tasks. The representative AA algorithms include: Multi-label decision tree (MLDT) (Zhang and Zhou, 2007; Piyatumrong et al., 2012), Multi-label k nearest neighbor (MLkNN) (Zhang et al., 2007; Brinker and Hüllermeier, 2007; Savvas and Sofianidou, 2015), improved Adaboost algorithm (Schapire and Singer, 2000), improved algorithm based on support vector machine (SVM), and neural network (Elisseeff and Weston, 2001; Vinod and M.P., 2015; Zhang et al., 2006; Zhang, 2009). By contrast, in PT algorithms, the multi-label classification task is transformed into some single label classification tasks, and then combine all single label classification results to determine the final labels. For example, binary relevance (BR) (Shoji et al., 2016; Tsoumakas and Taniar, 2007) tries to learn a binary classifier for each label, and then determines the associated labels for each unseen instance in terms of the results of all the binary classifiers. However, the BR algorithm doesn’t consider the correlations between labels, although it’s straightforward and easy to implement. The label powerset (LP) algorithm (Boutell et al., 2004) incorporates the correlations between labels to improve the classification performance. It estimates the correlations of labels by regarding each distinct combination of labels in the training data set as a new label. However, there may be an explosive increase of new label numbers with the considerable number of labels and training examples. The random k-labelset (RAkLE) algorithm (Tsoumakas et al., 2011) aims to solve the problem of excessive number of labels in the LP algorithm. It divides the initial set of labels into a small number of groups to reduce the number of new labels. There are some other well-known PT algorithms, such as classifier chains (CC) (Read et al., 2011), multi-label learning by exploiting label dependency (LEAD) (Zhang and Zhang, 2010) algorithm, and conditional dependency network (CDN) (Guo and Gu, 2011).

Recently, supervised topic models, which are derived from the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model, have become popular and seem promising to solve the supervised learning tasks in a distinct way, such as supervised LDA (sLDA) (Blei and Mcauliffe, 2010) model, discriminative LDA (Disc LDA) (Lacoste-Julien et al., 2009) model, maximum entropy discrimination LDA (Med LDA) model (Zhu et al., 2009), and Labeled LDA (L-LDA) model (Ramage et al., 2009). They simultaneously model textual contents and labels of text documents by the Bayes network model. The tokens of text documents are considered to be generated by topics. Each topic is a multinomial distribution over terms, and can be regarded as a semantic cluster of terms. The labels of text documents are usually embedded into the generation process of terms, which makes the posterior distribution over labels able to identify the associated labels of text documents.

Among supervised topic models, L-LDA is an effective model for the multi-label classification task. It builds a one-to-one correspondence between labels and topics, then regards each text document as a mixture of topics corresponding to its labels, rather than as a mixture of all topics that LDA adopts. Thus, the tokens of each document are generated by its associated labels. However, there are some noninformative terms in text documents, such as ‘get’, ‘why’, ‘properly’, which are not helpful for but disturb the recognition of the labels of text documents. These terms are called function terms in this paper. They are usually for the purpose of facilitating writing and have little explicit semantics. L-LDA assumes that function terms are generated by labels, thus associating function terms with labels wrongly would decrease the classification performance.

In this paper, a supervised topic model named labeled LDA with function terms (LF-LDA) is proposed to extract function terms from text documents. It uses a component to represent function terms. The component is shared by all the text documents in the corpus and modeled as a multinomial
Distributed Privacy Preserving Clustering via Homomorphic Secret Sharing and Its Application to (Vertically) Partitioned Spatio-Temporal Data
[www.igi-global.com/article/distributed-privacy-preserving-clustering-via/49640?camid=4v1a](www.igi-global.com/article/distributed-privacy-preserving-clustering-via/49640?camid=4v1a)

A New Spatial Transformation Scheme for Preventing Location Data Disclosure in Cloud Computing
[www.igi-global.com/article/a-new-spatial-transformation-scheme-for-preventing-location-data-disclosure-in-cloud-computing/117157?camid=4v1a](www.igi-global.com/article/a-new-spatial-transformation-scheme-for-preventing-location-data-disclosure-in-cloud-computing/117157?camid=4v1a)