# Word Way in Classificational MedLDA

## Abstract

本文探讨分类MedLDA模型如何进行后续的单词选择。为了实现这个目标，我们增强了原来的分类MedLDA模型，对每个单词赋予一个权重，因此每个文档的向量表示不再是所包含单词向量的简单加和，而是所包含的不同权重的单词向量的加和。在EM算法的M-Step我们使用梯度下降的方法优化这些权重。模型训练好之后，按照权重从大到小，我们选取不同比例的单词向量来对新文档进行分类。我们在真实的网页分类集上测试所提出的方法，结果总结为以下三点：1) 所有类别的平均分类效果在所有的单词向量被使用时最好。2) 很小一部分单词(在我们试验中小于2%)能够保留全部单词使用时几乎相同的效果。3) 这部分单词的比例随着样本中不同单词数量的增加而变小。这些结果表明在使用分类MedLDA模型进分类时一个更廉价的方式，仅保留一小部分权重高的单词向量。

## Introduction

随着越来越多的有监督的样本可以被收集和处理，建立有监督的主题模型引起了更多的关注。这包括有监督的LDA(sLDA) (Blei & McAuliffe,2007)，判别LDA(DiscLDA) (Lacoste-Jullien et al., 2008)等。这些样本中大量的单词使有监督的主题模型用来预测时面临挑战，对于每个单词，我们要存储一个K(例如取值150)维浮点数组成的向量。存储所有单词向量需要的空间非常大，从而使模型变得笨拙和低效。这些单词向量相互依赖，即使去掉其中的一小部分向量，也会使整个模型崩溃。我们必须重新思考这些监督主题模型，找到有效的单词选择方式。

最大熵判别LDA(MedLDA) (Zhu et al., 2012) 将判别最大边缘技术运用到一个统一的概率框架中。Classificational MedLDA是用来进行分类预测的MedLDA模型。 它不同于传统的两阶段启发式： 先使用主题模型估计出每篇文档的主题向量，然后用主题向量替代文档作为分类预测模型的输入。它将最大边缘预测模型(例如svm)内含机制与层次贝叶斯主题模型(例如LDA)内含机制融合到一个整体的约束最优化框架中。因为包含的主题模型机制，它比最大边缘预测模型具有更好的泛化能力。

(Classificational MedLDA 的被应用情况介绍)

然而，关于Classificational MedLDA模型如何进行单词选择的研究还较少。在(Taira & Haruno，1999) 中，他们探讨了svm 文本分类如何进行先验单词选择。他们提出两种方法，基于交互信息和基于词性标注。在他们的实验中，基于词性标注的单词选择要优于基于交互信息的单词选择。从而得出一种比较通用的单词选择方法：只保留词性标注为名词的单词。这种先验特征选择的方法也适用于Classificational MedLDA。然而这种方法能够过滤掉的单词比例较小，在一般的文本中，词性标注为名词的单词所占的比例超过50%。SVM优于Classificational MedLDA的一点是，如果kernel为线性的话，所有的支持向量可以被压缩为一个支持向量，这样就会节约大量的存储空间。而在Classificational MedLDA中，为了保持模型的完整性，我们必须将所有的单词向量都存储起来。

为了解决上述问题，本文探讨如何进行Classificational MedLDA的后续特征选择。首先使用全部单词向量进行训练Classificational MedLDA模型，使用训练好的模型进行分类预测时，只有被选择的单词向量起作用。为了使后续单词选择有参考的依据，我们对每个单词赋予一个权重。该权重反应出单词的向量在类别区分中的贡献度。有些单词(例如虚词)对类别区分作用甚微，我们希望这部分单词的权重越小越好。