LDA 特征扩展的多类 SVM 短文本分类方法研究

郑 腾，吴雨川*

（武汉纺织大学 机械工程与自动化学院，湖北 武汉 430200）

摘 要：针对短文本信息量少、特征稀疏的特点，提出一种基于LDA主题扩展的多类SVM短文本分类方法。在短文本基础上，利用LDA主题模得到文档的主题分布，将主题中的词扩充到原短文本的特征中，在特征空间上使用基于经典权重计算方法的多类SVM分类器进行分类。实验结果表明，在各个类别上的查准率、查全率和F1值都有所提高，验证了该方法的可行性。

关键词：短文本分类；特征扩展；SVM；LDA

中图分类号：TP391.1 **文献标识码：**A **文章编号：**2095－414X(2019)02－0072－05

随着网络技术的快速发展，手机短信、微博、广告文本等以文本表示的信息内容以极快的内容增长，而这其中又有很大一部分是短文本内容，如何从短文本内容中发现有价值的信息成为信息处理急需解决的问题。

由于短文本具有信息量少、特征稀疏、依赖上下文等特点，传统的空间向量模型以及 KNN、贝叶斯等经典算法不能很好应用在短文本分类上。短文本的分类主要的难点在于特征非常稀疏[1]和上下文依赖性强。一些学者通过引入外部知识库来扩展文本的语义特征，丰富了词语间语义关系[2-3]，但是它的计算量大，耗时大。因此，本文采用 LDA 主题扩展可以将对应主题下的词扩充到原来短文本的特征中，作为新的部分特征词，在新的特征空间上使用基于经典权重计算方法的多类 SVM 分类器进行分类。

1 LDA 主题扩展的短文本分类

1.1 LDA 主题模型

LDA[4]是一种三层贝叶斯概率模型，由词项、主题、文档组成，LDA 三层模型如图 1 所示，若干个隐含主题随机组成一个文档，而每个主题又由文档中的若干个词语表示。因此，可将每篇文档代表为主题概率分布，而每个主题又代表词项概率分布。

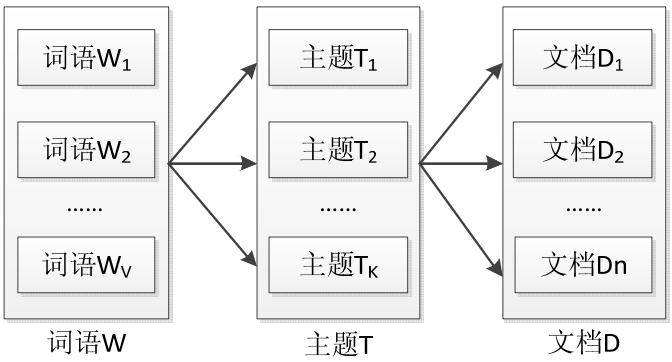

图 1 文档－主题－词语关系

LDA 模型如图 2 所示，该模型引入两个超参数 α，β，表示多维变量相互之间的权重关系。“文档—主题”的概率分布符合多项分布，“主题—词语”的概率分布也同样符合多项分布。

***通讯作者：**吴雨川（1960–），男，教授，研究方向：模式识别、智能检测与控制.
基金项目：国家自然科学基金面上项目（61271008）.

LDA 模型符号含义如下：M 表示文档集的文档总数，K 表示主题数，V 表示词汇数，α 表示主题分布的超参数，β 表示词项分布的超参数，θ_m 表示第 m 篇文档的主题分布，φ_k 表示 K 主题下的词分布，N_m 表示第 m 篇文档的词总数，Z_{mn} 表示第 m 篇文档中第 n 个词项的主题，W_{mn} 表示第 m 篇文档中第 n 个词语。

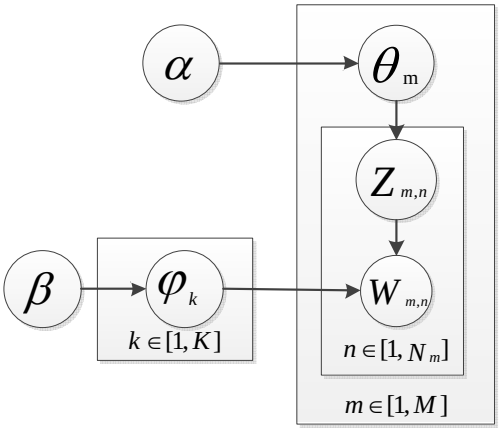

图 2　LDA 主题模型

利用 LDA 主题模型对文档集主题模型的生成过程可以看作模型一种概率取样的过程，具体步骤如下：

（1）对文档集中的第 m 篇文档，N_m 服从泊松分布，其参数为 ξ：N_m：$Poisson(\xi)$；

（2）对第 m 篇文档,按概率生成其主题分布：θ_m：$Dirichlet(\alpha)$；

（3）每个主题 z，按概率生成其词项分布：φ_k：$Dirichlet(\beta)$；

（4）第 m 篇文档中每个词 W_n 的生成过程有：

（i）依据主题分布 θ_m 生成文档 m 词项 W_n 主题：$Z_{m,n}$：$Multinomial(\theta_m)$；

（ii）根据词项分布 φ_k 生成所选主题词项：$W_{m,n}$：$Multinomial(\varphi_k)$。

不断重复上述过程，完成 M 篇文档的生成。依据 LDA 主题模型，可以写出所有变量的联合分布：

$$P(w_m, Z_m, \vartheta_m, \phi \mid \alpha g\beta) = \prod_{n=1}^{N} p(w_{m,n} \mid \varphi_{Z_{m,n}}) p(Z_{m,n} \mid \vartheta_m) g p(\vartheta_m \mid \alpha) g p(\phi \mid \beta) \tag{1}$$

通过对 ϑ_m 和 ϕ 积分以及 $Z_{m,n}$ 求和，可以求得 $w_{m,n}$ 的分布：

$$P(w_m \mid \alpha g\beta) = \iint p(\vartheta_m \mid \alpha) g p(\phi \mid \beta) g \prod_{n=1}^{N} \sum_{Z_{m,n}} p(w_{m,n} \mid \varphi_{Z_{m,n}}) p(z_{m,n} \mid \vartheta_m) d\phi d\vartheta_m =$$

$$\iint p(\vartheta_m \mid \alpha) g p(\phi \mid \beta) g \prod_{n=1}^{N} p(w_{m,n} \mid \vartheta_m, \phi) d\phi d\vartheta_m \tag{2}$$

整个文档集 W 的分布为：

$$p(w \mid \alpha g\beta) = \prod_{m=1}^{M} p(w_m \mid \alpha g\beta) \tag{3}$$

1.2　SVM 简介

SVM 是在高维特征空间使用线性函数假设空间的学习系统，它集成了最大间隔超平面、Mercer 核、凸二次规划、稀疏矩阵核松弛变量等多项技术[5]。SVM 在解决小样本、非线性以及高维模式识别中所具有的优势，取得丰硕成果。但是，传统的 SVM 是针对二分类问题提出的，现实中遇到的大多数问题是属于多分类问题。所以，在保证精度的前提下提出合理的多分类策略就成为当前的一个研究热点。按照采取策略不同，将多类 SVM 分类器的构建方法分为“一对一”[6]、“一对多”、二叉树算法、直接非循环图 SVM 等。其中，“一对一”方法是对多类中的任意两类都构建一个二类分类器，理论最严密，也是实际中使用最多的方式。具体做法是：选取 2 个不同类别构成一个 SVM 子分类器，共有 k(k−1)/2 个 SVM 子分类器，通过组合这些子分类器，利用特征权重进行分类投票，票数最多的一类即为该样本所属的类别。文献[7]比较了常见的几种多类分类支持向量机。

1.3　参数估计方法

LDA 主题模型有两种主要的参数估计方法：变分推算方法和吉布斯抽样方法（Gibbs 抽样方法）。由于 Gibbs 采样的直接、易于理解和运行速度快的特性，成为最常用的估计参数的方法。本文中采用它对 LDA 模型进行估计，其采样公式如下：

$$p(z_i = j \mid z_{-i}, w_i) = \left(\frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(g)} + V_\beta} g \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T_\alpha}\right) / \left(\sum_{j=1}^{T} \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(g)} + V_\beta} g \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T_\alpha}\right) \tag{4}$$

其中，w_i 是词标号，$z_i = j$ 表示将 w_i 分配给主题 j，$n_{-i,j}^{w_i}$ 表示 w_i 具有相同主题 j 的词项数，$n_{-i,j}^{(g)}$ 表示分配主题为 j 的词项数，$n_{-i,j}^{(d_i)}$ 表示文档 d_i 具有相同主题的词项数。

Gibbs 抽样方法的步骤如下：

（1）随机初始化：训练语料库中任一篇文档中的每个单词 w 被随机分配一个主题号 z；

（2）重新扫描语料库，根据每个单词 w 的采样公式（4）重新对其主题进行采样，并在语料库中对其进行更新；

（3）重复（1）到（2）步骤，直到 Gibbs 采样收敛；

（4）统计语料库的“主题–词”共出现的频率矩阵，该矩阵就是 LDA 模型。

1.4 卡方检验

卡方检验是一种假设检验方法，它是比较两个分类变量的关联性分析。词项 t 与类别 c 之间的卡方统计模型为[8]：

$$CHI(t, c_i) = x^2(t, c_i) = \frac{N g (AD - BC)^2}{(A+C) g (B+D) g (A+B) g (C+D)} \tag{5}$$

其中，N 表示文档数，A 表示包含词项 t 且属于 c 类的文档数，B 是包含词项 t 且不属于 c 类的文档数，C 是不包含词项 t 且属于 c 类的文档数，D 是不包含词项 t 且不属于 c 类的文档数。$CHI(t, c_i)$ 值越大代表词项 t 和类别 c_i 有关联的可能性越大。

2 实验分类步骤

2.1 分类框架

具体分类框架如图 3 所示。

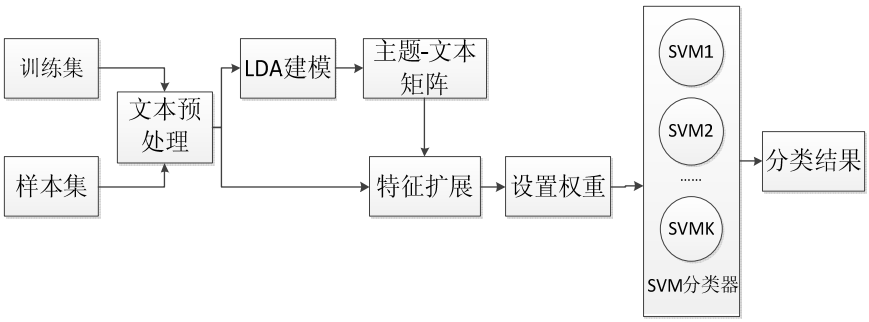

图 3 基于 LDA 特征扩展的短文本分类框架

2.2 文本预处理

文本预处理是文本分类的第一步。首先，去除指定无用的符号，比如：数字、空格、标点等，可以通过 Python 的正则表达式（re）删除，让文本只保留汉字。然后，用 jieba 分词(结巴分词)工具进行短文本分词。最后，通过停用词表去除停用词，过滤文本中很多无效的词。

2.3 特征选择与向量表示

目前大多数中文分类系统都采用词作为特征项，如果把所有的词项都作为特征项，将会导致特征向量的维数非常高，快速完成文本分类将非常困难。特征选择是在表达文本关键信息的时候，能够使特征向量维度处于在一个合适的范围，使得在处理文本分类时候效率得到提高。文本特征选择的方法主要有：互信息[9]、信息增益[10]、文本证据权、卡方检验[11]等。本文选择卡方检验作为文本特征选择。对训练集中的每一篇文档经过文本预处理后，进行向量化，得到特征词典。

2.4 特征扩展

首先使用一个数量比较大的文档集训练 LDA 模型，得到“主题–词”分布矩阵。将训练好的 LDA 模型

应用于语料集中的某一篇文档进行预测，得到“文档–主题”概率分布，将概率最大主题下的词语扩展到短文本初始特征中，形成新的特征向量。

2.5　多类 SVM 分类

在进行多类 SVM[12]训练前，需要组合这些子分类器，利用特征权重进行分类投票，也即需要对 LDA 特征扩展的空间向量模型中特征矩阵设置权重。一般最经典的常用方法是 TF–IDF，具体计算公式如下：

$$TFIDF(d,t)=TF(d,t)\mathrm{g}\log_2\frac{N}{DF(t)} \tag{6}$$

其中，$TF(d,t)$ 为文档 d 中词项 t 出现的频率数；N 为训练文档总数；$DF(t)$ 为包含词项 t 的其他训练文档数目。

根据 LDA 特征扩展的空间向量模型，首先依据权重计算公式(6)计算主题特征词的权重，通过 SVM 分类器（LIBSVM）进行分类。

2.6　LDA 模型主题数确定

将复旦大学提供的文本数据作为 LDA 模型的语料库。测试语料集与训练语料集的比例为 1:2。先在训练语料集上训练出不同主题下的 LDA 模型，通过训练生成的 LDA 模型推断出测试语料集的“主题–词”矩阵分布。并在测试语料集上计算其困惑度。困惑度计算公式如下：

$$perplexity=\exp\{-(\sum_{m=1}^{M}\sum_{n=1}^{N_m}\log(\sum_{k=1}^{K}p(w_n\mid z_k)p(z_k\mid d_m)))/(\sum_{m=1}^{M}N_m)\} \tag{7}$$

其中，M 为测试语料集中短文本总数，N_m 为第 m 篇文本词项总数，K 为主题数，$p(w_n\mid z_k)$ 表示词项 w_n 在主题 z_k 下的概率，$p(z_k\mid d_m)$ 表示测试文本 d_m 在主题 z_k 的概率。

重复多次实验可得到，随着主题数量的增加，困惑度逐渐降低。当主题数增加到 46 之后，困惑度变化并不明显。所以最终选取主题数 $T=46$。

3　实验结果与分析

3.1　实验数据

本文数据来源于复旦大学提供的中文分类数据集，将其分为 8 类：教育、体育、科技、财经、艺术、旅游、历史、政治。训练语料集与测试语料集的比例为 2:1。每一类随机抽取 1200 个文本，共 9600 个文本组成训练语料集。每一类筛选 600 个短文本，共 4800 个短文本组成测试语料集。

3.2　性能评估

使用准确率 P_m、召回率 R_m 和 F_1 值对实验结果进行评估。具体公式如下：

$$P_m=A_m/B_m\times100\% \tag{8}$$

$$R_m=A_m/C_m\times100\% \tag{9}$$

$$F_1=\frac{P_m\times R_m\times2}{P_m+R_m}\times100\% \tag{10}$$

3.3　实验结果

特征选择中，利用卡方检验公式（5）计算出每类词项的卡方值，将每一类前 600 个词项作为该类的特征，并合并为特征词典中。实验结果如表 1 所示：

表 1　实验结果对比

类别		教育	体育	科技	财经	艺术	旅游	历史	政治
VSM+SVM	P_m	0.739	0.794	0.813	0.821	0.758	0.772	0.804	0.782
	R_m	0.726	0.767	0.824	0.794	0.743	0.765	0.817	0.769
VSM+LDA+SVM	P_m	0.797	0.825	0.841	0.835	0.776	0.819	0.812	0.837
	R_m	0.804	0.818	0.848	0.848	0.752	0.804	0.819	0.829
VSM+LDA+KNN	P_m	0.754	0.835	0.834	0.819	0.763	0.792	0.793	0.809
	R_m	0.761	0.828	0.827	0.808	O.759	0.790	0.806	0.795

从表 1 中可以看出，本文所采用的 VSM+LDA+SVM 分类方法优于 VSM+SVM 和 VSM+LDA+KNN 方法。说明本文的方法是切实可行的。

F_1 值对比如图 4 所示，VSM+LDA+SVM 方法的 F_1 大于其他两种方法，也说明本文方法是可行的。

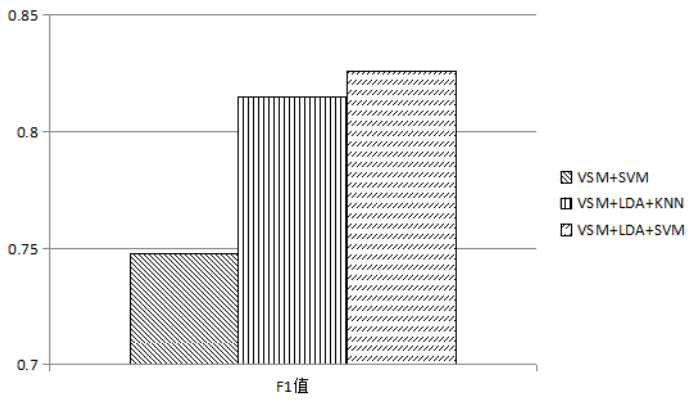

图 4 三种方法的 F_1 值对比

4 结论

文本分类涉及许多复杂的技术，如文本表示，特征稀疏处理和算法决策。本文研究并改进了传统的特征选择方法，利用 LDA 模型进行特征扩展，丰富了短文本的语义信息，解决了短文本数据长度短、信息弱的问题。在多类 SVM 特征权重设置是基于 TF-ID，在分类实现上还有很多不足地方，在后续的工作中，可以改进特征提取方法，设置新的特征权重，以期得到更好的分类效果。

参考文献：

[1] 张虹. 短文本分类技术研究[D]. 大连：辽宁师范大学，2015. 2-3.
[2] 朱征宇，孙俊华. 改进的基于知网的词汇语义相似度计算[J]. 计算机应用，2013，33(8)：2276-2279+2288.
[3] 王荣波，谌志群，周建政，等. 基于 Wikipedia 的短文本语义相关度计算方法[J]. 计算机应用与软件，2015，32(1)：82-85+92.
[4] Blei D M，Ng A Y，Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning R esearch，2003，(3)：993-1022.
[5] 刘秀松. 带有云化核函数的 SVM 文本分类方法[J]. 科技情报开发与经济，2007，17(30)：13-15.
[6] Ulrich Krebel. Pairwise classification and support vector machines[M]. In B. Schuolkopf，Burges C J C，Smola A J，editors，Advances in Kernal Methods：Support Vector Learning，Pages，MITPress，Cambrige，MA，1999. 255-268.
[7] HSU C W，LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks，2002，13(2)：415-425.
[8] 闫健卓. 基于 X2 统计的改进文本特征选择方法[J]. 计算机应用研究，2012，29(7)：2454-2456.
[9] 刘海峰，姚泽清，苏展. 基于词频的优化互信息文本特征选择方法[j]. 计算机工程，2014，40(7)：179-182.
[10] 刘庆河，梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用，2011，47(12)：130-134.
[11] 裴英博，刘晓霞. 文本分类中改进 CHI 特征选择方法的研究[J]. 计算机工程与应用，2011，47(4)：128-130.
[12] 霍颖瑜，王晓峰. 一种新的 SVM 多类分类算法[J]. 佳木斯大学学报（自然科学版），2006，24(4)：476-478.

Research on the Classification Methods of Multiple SVM Short Texts based on LDA Feature Extension

ZHENG Teng, WU Yu-chuan
(School of Mechanical Engineering and Automation, Wuhan Textile University, Wuhan Hubei 430200, China)

Abstract： Based on the short text and characteristics of sparse, this paper puts forward a short text classify method based on characteristics extend of LDA. This method used the LDA model to obtain the subject distribution of document, extended the word under the corresponding topic into the characteristics of the original short text as a new part of the feature word. A multi-class SVM classifier based on classical weight calculation was used. Experimental results show that the precision, recall and F1 values are improved in all categories. It verifies this model has some superiority in text categorization.

Key words： short text classification; feature expansion; Laten Dirichlet Allocation(LDA); SVM