结合半监督学习和 LDA 模型的文本分类方法

韩 栋[1]，王春华[1]，肖 敏[2]

(1. 黄淮学院 信息工程学院，河南 驻马店 463000；
2. 武汉理工大学 计算机科学与技术学院，湖北 武汉 430063)

摘 要：针对样本集中具有较少标记样本情况下的文本分类问题，提出一种结合半监督学习(SSL) 和隐含狄利克雷分配(LDA) 主题模型的标记样本扩展方法(SSL-LDA)，并整合朴素贝叶斯(NB) 分类器构建一种文本分类方法。使用 LDA 主题模型生成主题分布，以表示所有样本；根据训练集中已标记样本，通过一种简化粒子群优化(SPSO) 算法获得 SSL-LDA 自训练模型的最优参数；基于 SSL-LDA 自训练模型对训练集中一些未标记样本进行标记，扩展训练集；基于扩展后的训练集，训练 NB 文本分类器。在 3 个数据集上的实验结果表明，该方法能够很好地应对标记样本较少的情况，获得了较高的分类精确度。

关键词：文本分类；半监督学习；LDA 主题模型；简化粒子群优化；标记样本扩展
中图法分类号：TP311 **文献标识号**：A **文章编号**：1000-7024 (2018) 10-3265-07
doi：10.16208/j.issn1000-7024.2018.10.045

Text categorization scheme based on semi-supervised learning and latent Dirichlet allocation model

HAN Dong[1], WANG Chun-hua[1], XIAO Min[2]

(1. School of Information Engineering, Huanghuai University, Zhumadian 463000, China; 2. School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: For the text classification problem of fewer labeled samples in the sample set, a labeled sample extension method (SSL-LDA) combining the semi-supervised learning (SSL) and the latent Dirichlet distribution (LDA) topic model was proposed, and naive Bayesian (NB) classifier was integrated to construct a text categorization method. The LDA topic model was used to generate a topic distribution to represent all samples. The optimal parameters of the SSL-LDA self-training model were obtained using a simplified particle swarm optimization (SPSO) algorithm according to the labeled samples in training set. The SSL-LDA self-training model was used to label some unlabeled samples in the training set. The NB text classifier was trained based on the expanded training set. Experimental results on three datasets show that the proposed method can deal with the less labeled samples and obtain high classification accuracy.

Key words: text categorization; semi-supervised learning; latent Dirichlet allocation model; simplified particle swarm optimization; labeled samples extension

0 引 言

为了使文本分类效率进一步提高，必须提供足够多的已标记样本来训练分类器。然而，在很多实际情况中，已标记样本可能很少，但却有大量的未标记样本。为此，可以使用半监督学习(semi-supervised learning，SSL) 方法[1]，通过学习将未标记样本进行归类，扩大初始标签集，用作传统机器学习算法的输入。

主题模型(topic modeling)[2]是一种对文字中隐含主题进行建模的方法，由于其考虑了上下文关系，能够显著减少特征的数量并能够压缩文本表示。有研究表明，在训练集数据量较少时，主题模型在监督学习环境中的性能优于

收稿日期：2017-08-23；**修订日期**：2018-03-15
基金项目：河南省科技厅科技计划基金项目 (172102210117)；河南省驻马店市科技计划基金项目 (17135)
作者简介：韩栋 (1979-)，男，河南驻马店人，博士研究生，讲师，研究方向为数据挖掘等；王春华 (1980-)，女，四川眉山人，博士，副教授，研究方向为数据挖掘。肖敏 (1979-)，女，河南南阳人，博士，副教授，研究方向为数据挖掘与信息安全。
E-mail：hhxyHanD@126.com

万方数据

传统表示方法[3,4]。其中，以隐含狄利克雷分配(latent Dirichlet allocation，LDA)[5]概率主题模型作为文本相似性度量时，比基于词频的 TF-IDF 表示方法具有更高的效率[6]。

基于上述分析，为了解决在较少初始标签集环境下的文本分类问题，提出一种结合 SSL 和 LDA 主题模型表示法的训练集构建方法(SSL-LDA)。使用 LDA 主题模型表示文本特征，使用一个基于 SSL 的自训练模型来学习分类未标记文本，扩展标记文本集，以此解决具有少量标记样本集的场景。另外，为了获得 SSL-LDA 模型的最优参数组合，首先，通过方差分析(analysis of variance，ANOVA) 统计测试方法来确定各种参数的影响能力，找出影响最小的参数并确定其值，以此降低参数维度。然后，通过一种简化粒子群算法(simplified particle swarm optimization，SPSO)来获得其它参数的最优组合。最后，基于扩展的训练集来训练朴素贝叶斯(NB) 分类器实现文本分类。

1 基于 LDA 主题模型的文本表示

仅仅依靠特定单词不足以描述一个文本，也不能有效地对文本进行区别。也就是说，具有相同内容的两个文本可以使用包含相似含义的不同单词来进行表述。因此，需要在一个共同的语义空间中表示文本，这种表示的最基本技术是潜在语义分析(latent semantic analysis，LSA)[7]。LSA 方法中，对文本矩阵进行奇异值分解，并将其表示为低维语义空间中的潜在概念。概率潜在语义分析(PLSA)是一种改进 LSA 方法，这种方法提升了对主题的解释，将其考虑为多词分布。由于 PLSA 方法基于对给定文本的极大似然估计，所以容易出现过拟合现象。

为了解决上述问题，学者引入 LDA 方法。使用 LDA 方法，可以将文本表示为多个主题的分布，主题可以作为类似集群的更高层次的概念。这一算法基于这样一个假设：集合中的每个文本都是由几个潜在主题创建的，其中每个主题都以混合的单词呈现[8]。文本的主题表示过程描述如下：

(1) 对于每个文本 $m \in M$，主题分布 θ_m 都是 Dirichlet 分布 $Dir(\alpha)$ 的一个抽样；

(2) 对于文本 m 中的每个单词占位符 n：

1) 根据抽样得到的主题分布 θ_m 随机选择一个主题 $z_{m,n}$；

2) 从主题 $z_{m,n}$ 的多项式分布 ϕ_k 中，随机选择一个词 $w_{m,n}$。

在 LDA 模型中，为了获得文本中各种主题的分布，可采用期望最大化(expectation maximization，EM) 方法、期望变分法以及 Gibbs 抽样方法[9]。然而，EM 方法容易陷入局部最优。为此，本文采用了 Gibbs 抽样方法。这种方法基于 Markov 链蒙特卡洛算法(MCMC)，估计出文本中每个单词的主题分布(即文本-主题分布 θ) 和文本集中所有单词的主题分布(主题-单词分布 ϕ)，并以此来计算单词属于某个主题的概率，从而更新该词的主题。Gibbs 抽样方法步骤描述如下：

(1) 设定训练文本集中文本数量为 M，单词数为 I，主题数为 T，Gibbs 抽样迭代次数为 N。

(2) 初始化主题分布，即对于每个单词 w_i，$i \in I$，将其随机赋给一个主题 t，表示为 $z_i = t, t = random(T)$。

(3) 在每次迭代中，根据下式计算单词 w_i 属于主题 t 的后验概率

$$p(z_i = t \mid z_{-i}, w_i) \propto \frac{n_{-i,j}^{(w)} + \beta}{n_{-i,j}^{(\cdot)} + M\beta} \times \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{d_i} + T\alpha} \tag{1}$$

式中：z_{-i} 表示除了词汇 w_i，所有词汇 $w_k (k \neq i)$ 的主题分配；$n_{-i,j}^{(w)}$ 表示整个文本集中，除了当前词汇 w 外的所在词汇 $w_k (k \neq i)$ 分配为主题 j 的数量；$n_{-i,j}^{(\cdot)}$ 表示整个文本集中，除了词汇 w_i，分配为主题 j 的所有词汇数量；$n_{-i,j}^{(d_i)}$ 表示文档 d_i 中，除了词汇 w_i，分配为主题 j 的所有词汇数量；$n_{-i,\cdot}^{d_i}$ 表示文档 d_i 中，除了词汇 w_i，所有分配了主题的词汇数量。

(4) 通过对每个文本中主题数量的统计，计算 θ 和 ϕ，表示如下

$$\theta^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}, \quad \phi_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + M\beta} \tag{2}$$

式中：$n_j^{(d)}$ 表示文档 d 中分配为主题 j 的所有词汇数量，$n_\cdot^{(d)}$ 则表示文档 d 中分配主题的词汇总数。$n_j^{(w)}$ 表示词汇 w 分配为主题 j 的次数，$n_j^{(\cdot)}$ 则表示分配为主题 j 的词汇总数。

在 Gibbs 抽样算法中，需要设定 α 和 β 参数、主题数量 T 以及估计 θ 和 ϕ 的迭代次数 N。根据相关研究[10]，参数 α 和 β 可以设定为固定值，即：$\alpha = 50/K$，$\beta = 0.01$。其它参数对主题模型的性能影响较大，为此需要获得最优参数组合值。这将在第三章中进行详细描述。

2 提出的 SSL-LDA 标记文本扩展模型

2.1 半监督学习

半监督学习(SSL) 方法是有监督学习方法的延伸，其将未标记样本作为整个训练集的一部分，而不是仅用已标记样本，从而获得一个更好的分类器。首先，使用少量的已标记样本 D^l 训练得到一个基分类器。然后，使用该分类器对未标记样本 D^u 进行分类。根据分类预测结果，将一些可信度最高的未标记样本归类到特定标签中。重复整个过程直到达到停止条件[11]。

在文本分类中，由于 SSL 方法是基于标记文本和未标记文本之间的相似性度量对文本进行划分，因此文本的表示是至关重要的。对于非结构化内容的表示，通常使用向量空间模型(vector space model，VSM) 进行表示[12]。其中，特征是以单词为基本单元来构建的，特征值由不同的

万方数据

加权算法计算得到，例如常用的词频-逆文本频率(TF-IDF)方法[13]。但是，这种方法忽略了单词的顺序及其含义。此外，这种方法得到的单词矢量是稀疏的，具有非常高的维度。所以，为了提高SSL方法扩展训练集的准确性，需要一种有效的内容表示模型。为此，本文将LDA与SSL相结合。文本的LDA表示能够产生数量较少且包含语义信息的主题特征，能够为SSL提供高效输入特征，从而可提高其对未标记样本进行标记的准确性。另外，由于LDA是一种无监督学习过程，其主题与文本类别的对应关系存在不确定性，将SSL与LDA结合，构成半监督LDA可以降低这种不确定性。

2.2 提出的SSL-LDA模型

为了实现在标记样本数量较少情况下的样本分类，本文构建了一种用于扩展标记样本集的自训练模型：SSL-LDA模型。SSL-LDA模型由3个部分组成：①文本LDA主题模型表示，其是系统的基础，产生了整个系统的输入；②用于扩大初始标记样本集的自训练模型，其是整个系统的核心，实现较少样本集的扩展；③基于SPSO的参数优化模型，其是系统的促进剂，能够提高自训练模型的性能。

基于LDA主题模型，利用主题分布来表示样本。使用少量已标记样本和更多非标记样本构建一个初始样本集，这些非标记样本具有和已标记样本相似的主题分布，作为SSL-LDA方法的输入。

SSL-LDA中，首先对给定的任意样本集合，使用LDA主题模型生成主题分布，并将所有样本用这一分布进行表示。然后，根据已标记样本来测试不同参数组合下的SSL-LDA性能，通过参数优化模型来获得最优参数。接着，将初始标记样本集合以及未标记样本集合(规模相比标记样本集合大得多)一并提供给最优参数的SSL-LDA自训练模型。之后，执行自训练过程，自训练模型的输出为扩大后的标记样本集合。最后，基于该扩大后的标记样本集合，通过任何监督分类方法训练分类器，并将该分类器对其它未标记样本进行分类。整个过程如图1所示。

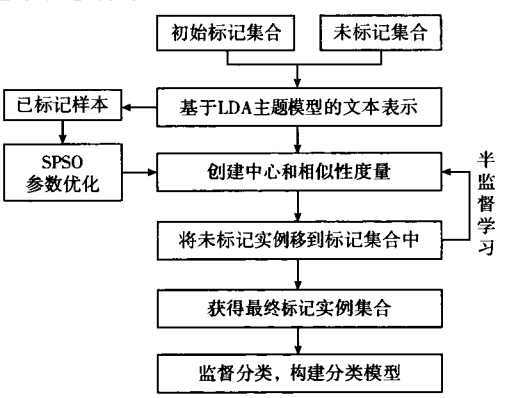

图1 所提出的SSL-LDA方法框架

自训练模型由两个阶段组成，用于扩大初始标记样本集合。第一阶段的目标是获得标记和未标记样本基于主题的表示。将所有的样本组合成一个集合，然后在该集合上进行主题建模。构建LDA模型时使用了Gibbs采样方法，使得每个样本都可以用LDA主题分布表示。

在第二阶段，未标记样本通过迭代逐渐地被移到标记样本集合中，直至达到预定的阈值。为了使最可靠的未标记样本被移动到标记集合中，定义了一种语义相似性度量，这种度量基于主题分布和余弦相似性度量来计算。由于在训练过程中计算每个未标记样本和标记样本的距离是相当耗时的，因此，本文为每个类计算得到一个质心，并基于这些质心来测量相似性距离。每次迭代依次执行以下两个步骤：

(1) 对于每个类别，都分别创建一个质心向量，其值为给定类别中已标记样本的平均值。然后，计算未标记样本与质心向量间的余弦距离，如下式所示

$$d_{\cos}(x,y)=1-\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}} \tag{3}$$

(2) 将未标记样本按其可靠性进行排序，可靠性根据未标记样本与其两个最近质心距离的差定义。其值越大，表示样本更加接近其中一个质心。然后，将最可靠的未标记样本移动到标记集合中，并基于距离度量将这一样本标记为最近质心所代表的类别。在选择未标记样本时，还考虑了类的平衡/不平衡性质，使得在可能的情况下，标记集合服从均匀分布。为此，首先计算所有类的不平衡率。然后将平衡参数值(R)减去每个类别的不平衡率(r)，以便可以使得$R-r$个样本能够移动到特定的类中。

对于所以未标记样本，当其对于两个最近质心的距离之间的差都小于预定义的相似性阈值(ST)时，迭代过程结束，得到最终的标记集合。通过使用平衡参数，在一次迭代中移动的样本数量也被确定下来。该过程的如算法1所示。

算法1：提出的SSL-LDA自训练模型

输入：

D^l：标记样本；

D^u：未标记样本；

$D=D^u\cup D^l$；$D^u\gg D^l$；//对于D中的每个分类，都至少具有一个样本在D^l中(即D中至少有1个标记样本)

ST：相似度阈值；

R：平衡参数；

T：主题数；

GI：Gibbs迭代次数；

输出：

最终标记集合$\hat{D}^l$

初始化：

万方数据

基于参数 T 和 GI，将 D 中所有样本以主题分布的方式表示。

自训练过程：

While $\varepsilon > ST$ **do**

(1) 为标记样本 D^l 的每个类创建质心 $C = c_1, \cdots, c_m$，每类都具有一个类标签 l_{c_i}；

(2) 对于每个 $d_k^{(u)}$，计算他们与每个质心的余弦距离；$d_{\cos}(d_k^{(u)}, c_i)$，$c_i \in C$；

(3) 对于每个 $d_k^{(u)}$，计算其与两个最小质心距离的差值，并将其按差值由大到小排序；

$\overline{C} = \{c_x \in C \mid \exists c_y \in C: d_{\cos}(d_k^{(u)}, c_x) \geqslant d_{\cos}(d_k^{(u)}, c_y)\}$；

$c_{\min 1} = \operatorname{argmin}_{c_i \in C} d_{\cos}(d_k^{(u)}, c_i)$；

$c_{\min 2} = \operatorname{argmin}_{c_i \in C} d_{\cos}(d_k^{(u)}, c_i)$；

$dif_k = d_{\cos}(d_k^{(u)}, c_{\min 2}) - d_{\cos}(d_k^{(u)}, c_{\min 1})$；

(4) 定义 ε 的值：$\varepsilon = \max\{dif_1, dif_2, \cdots, dif_n\}$；

(5) **If** ($\varepsilon > ST$)

1) 计算每类中 D^l 样本所占比率 r，为每类选择 $R - r$ 个 dif_k 最大的未标记样本 D^u 放入各类中；

2) 将符合 $l(d_k^{(u)}) = l_{c_{\min 1}}$ 的样本 $d_k^{(u)}$ 从 D^u 移动到 D^l 中；

End if

End while

3 SSL-LDA 中的最优参数选择

3.1 ANOVA 参数重要性分析

在提出的 SSL-LDA 模型中，涉及 4 个参数，即相似度阈值 ST；平衡参数 R；主题数 T 和 Gibbs 迭代次数 GI。为了使 SSL-LDA 模型具有最好的性能，需要设定合适的参数组合。参数组合可以使用传统网格搜索方法来获得。在这种情况下，对于每种可能的参数组合，都分别进行评估。然而，如果在不同样本集下，对每种参数组合都进行实验验证，则需要消耗大量的资源，这是不切合实际的。因此，本文首先采用了 Castillo 等[14]提出的方差分析(analysis of variance，ANOVA) 统计测试方法，来确定各参数对分类方法性能的影响，以此来减少参数组合中参数数量。

ANOVA 是数据分析中常用的一种统计方法，其从结果变量的方差入手，研究诸多控制变量中哪些变量对结果变量有显著影响。ANOVA 依靠 F-分布为机率分布的依据，基于平方和(sum of square) 与自由度(degree of freedom) 计算的组间与组内均方值(mean of square)，并以此来估计 F 值。

在这一部分，所有的实验都是在预定义的训练集上进行的，其中所有样本都已被分配标签。为了测试 SSL-LDA 模型扩展样本标签的准确性，去除了已标记样本集中 80% 的样本标签，使其成为未标记样本。通过这种方式，在训练集上以每种可能参数组合的 SSL-LDA 模型来扩展样本标签，根据扩展准确性来验证参数组合的效果。

参数值的范围根据经验进行设定。对于相似性阈值 ST，设定范围为 [0.1，0.9]，增量为 0.1。对于平衡参数 R，设定范围为 [20，200]，增量为 10。对于主题数量 T，设定范围为 [20，200]，增量为 10；对于 Gibbs 采样迭代的次数 GI，设定为 500、750、1000、1250 和 1500。

通过这种方式，本文测试了 16 245 种不同的参数组合，用来确定哪些参数对结果影响最大。其中，使用未标记样本标签扩展的精度作为评估指标。

使用 ANOVA 统计测试来确定参数值的变化对标签扩展性能的影响是否显著。对于每种参数，分别统计了其平方和、自由度、均方值、F 统计量以及显著性水平，ANOVA 统计结果见表 1。

根据 ANOVA 统计分析结果，从 F 统计量来看，相似性阈值 ST 对标签扩展性能是最敏感的，主题数量 T 和平衡参数 R 对性能的影响也较高。而 Gibbs 迭代次数 GI 的重要性最低，且远远低于其它 3 个参数。尽管 Gibbs 迭代次数的增加可以提升主题模型的质量和稳定性，但它并不能显著的影响性能。另一方面，GI 参数对时间复杂度有着较大影响。因此，将 GI 参数值设定为可接受范围内的最小值是合理的。另外，在其它不同训练集上的 ANOVA 分析结果也表明，GI 的影响很小。

表 1 不同参数对标签扩展准确性影响的 ANOVA

参数	平方和	自由度	均方值	F 值	显著性水平
相似性阈值 ST	1.083×10^6	8	1.288×10^5	8.286×10^2	$<10^{-3}$
平衡参数 R	2.712×10^4	9	3.217×10^3	1.833×10^1	$<10^{-3}$
主题数量 T	2.815×10^4	9	3.182×10^3	1.851×10^1	$<10^{-3}$
Gibbs 迭代次数 GI	1.246×10^3	2	5.293×10^2	3.342	0.045

为此，本文将 Gibbs 迭代次数 GI 固定为 500。这样需要确定的 4 个参数就减少为 3 个参数，参数组合空间缩小为原来的 1/5，为 3249 种。这有助于提高空间搜索算法的收敛速度。

3.2 基于 SPSO 的参数优化

为了进一步获得剩余 3 个参数的最优组合，本文采用

万方数据

了一种收敛速度较快的简化粒子群优化算法(SPSO) 算法[15]。将粒子编码为 3 个参数的值，以基于该参数组合下 SSL-LDA 扩大标记样本集合的准确性作为粒子的适应度，以此进行寻优。

传统粒子群优化(PSO) 算法中，粒子位置是根据全局最优和当前最优位置来动态更新，表示为

$$v_{id}^{t+1} = w \cdot v_{id} + c_1 \cdot r_1 \cdot (p_{id} - x_{id}) + c_2 \cdot r_2 \cdot (p_{gd} - x_{id}) \tag{4}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{5}$$

式中：$i = 1,2,\cdots,N$ 表示粒子编号，x_{id} 表示粒子 i 中第 d 维的位置值，v_{id} 为移动速度，p_{id} 和 p_{gd} 分别为全局最优和当前最优位置。c_1 和 c_2 为比例参数，取值为 $c_1 = c_2 = 2$。$r_1()$ 和 $r_2()$ 为 [0，1] 内的随机值。w 为惯性权重。为了使其适应迭代收敛过程的变化，本文采用递减型动态惯性权重。

简化粒子群优化(SPSO) 算法中，其认为传统 PSO 中寻优过程的收敛性与粒子速度无关。在一些情况下，粒子速度的更新会增加算法复杂度和优化时间。为此，省略了传统 PSO 位置更新中的速度项，将二阶位置更新方程简化成为一阶形式，表示为

$$x_{id}^{t+1} = w \cdot x_{id}^{t} + c_1 \cdot r_1 \cdot (p_{id} - x_{id}^{t}) + c_2 \cdot r_2 \cdot (p_{gd} - x_{id}^{t}) \tag{6}$$

简化后的位置更新过程加快了收敛过程，使其能够快速的在解空间中搜索最优参数组合。

4 实验及分析

4.1 实验设置

将提出的分类算法在 WEKA 开源机器学习环境上，使用 JAVA 语言编程实现。为了建立 LDA 主题模型，还使用了 MALLET 工具箱。

本文使用了 3 个公共数据集进行了实验：Newsgroups、Reuters-10 和 Ohscal 数据集。Newsgroups 数据集由大约 20 000个样本组成，分为 20 个类别的新闻样本。这些样本都是从 UseNet 上收集的。Reuters-10 数据集来源于基本的 Reuters-21578 数据集，其由 Reuters-21578 中样本数量最多的 10 个类的样本组成。Ohscal 数据集是一个包含医学期刊中文献标题和摘要的数据集。其包含了 9121 个样本，分为 10 个不同的医学领域类别。

对于每个数据集，都进行了一些简单的预处理，包括将所有字母都转换为小写，删除了停用词以及长度少于 3 个字符的单词，并将其余单词都使用 Porter Stemmer 算法进行修剪。各个数据集的样本数量、类别数量以及预处理后的单词数量见表 2。

基于上述数据集，构建训练集和测试集，其大小比例为 6∶4，并尽可能使训练集和验证集中各类别样本的均匀分布。其中，训练集中已标记的样本比例为 20%，其它样本都为未标记样本。测试集中都为未标记样本。

表 2 实验数据集的基本属性

样本集	样本数量	类别数量	特征词数量
Newsgroups	18 828	20	33 489
Reuters-10	7951	10	7236
Ohscal	10 162	10	10 036

4.2 性能度量

(1) 准确性(accuracy)

准确性由查准率(precision) 和召回率(recall) 这两个度量计算得到，用来表示算法的整体正确分类性能。表示如下

$$pre = \frac{t_{pos}}{pos},\ rec = \frac{t_{neg}}{neg} \tag{7}$$

式中，t_{pos} 为正确分类的阳性样本，pos 为阳性样本总数，t_{neg} 为正确分类的阴性样本，neg 为阴性样本总数。准确性表示为

$$Accuracy = \left(pre * \frac{pos}{pos + neg}\right) + \left(rec * \frac{neg}{pos + neg}\right) \tag{8}$$

(2) AUC 面积

感受性曲线 (ROC) 是以真阳性率和假阳性率为坐标的曲线，曲线与 X 坐标轴之间的面积则为 AUC 面积，取值为 0.5 到 1 之间，用来反映分类器的效果。其值越大说明分类效果越好。

4.3 参数优化的性能验证

在提出的 SSL-LDA 模型中，需要合理设定 3 个参数。为了验证提出的参数优化方法的有效性，将估计出的最优参数与通过网格搜索方法获得最优参数进行比较。网格搜索方法是遍历所有 3249 种参数组合，根据该参数组合下的分类准确率来找到最优参数，该方法非常耗时，但较为准确，所以可作为基准。在两种方法获得参数组合下，在 3 个数据集上的训练集上进行标签样本扩展实验，其中以扩展样本的标签准确性作为性能指标，结果见表 3。

表 3 基于两种参数优化方法的样本标签扩展准确性/%

分类方法	数据集		
	Newsgroups	Reuters-10	Ohscal
SSL-LDA+NB (SPSO 参数优化)	96.27	95.02	97.73
SSL-LDA+NB (网格搜索参数优化)	96.46	95.39	97.57

可以看出，基于 SPSO 参数优化获得的扩展训练集与基于网格搜索的参数所获得扩展训练集的准确性相近，这说明了提出的参数优化算法所获得的参数组合几乎为最优参数组合，验证其有效性。另外，SPSO 算法只需要迭代搜索约 50 次左右即可得到最优解，而网格搜索需要遍历所有

万方数据

3249 种参数组合，这就大大降低了参数优化过程的计算时间。

4.4 样本分类的性能比较

本文使用 SSL-LDA 方法构建训练集，之后使用朴素贝叶斯(NB) 分类器进行分类训练，构建称为 SSL-LDA+NB 的样本分类方法。将提出的 SSL-LDA 训练集构建方法与其它相关方法进行了比较，分别为基于 TF-IDF 加权文本表示的自训练方法(SSL-TF-IDF)、基于期望最大化(EM) 的自训练方法。SSL-TF-IDF 方法中的学习过程与提出的 SSL-LDA 相似，区别在于其使用了词包和 TF-IDF 权重，而不是本文采用的 LDA 主题表示。EM 方法是一种用于进行不完整数据优化的方法，使用分类器从已标记样本中估计参数，然后将概率加权的类标签分配给未标记样本，迭代执行直到收敛。与 SSL-LDA 算法不同，这种方法的结果完全是确定的，但其缺点是可能收敛于局部最优点。为了公平比较，这些方法都采用 NB 作为分类器。另外，还和只包含 NB 分类器的监督学习算法进行了比较，用来验证本文融合半监督学习算法的可行性。

在每个数据集上执行 10 次实验，并计算分类准确性和 AUC 面积的平均值，结果如表 4 和表 5 所示，其中最佳性能值由粗体突出表示。

表 4 各种方法在数据集上的分类准确度/%

数据集	NB	SSL-LDA+NB	SSL-TF-IDF+NB	EM+NB
Newsgroups	38.86	**73.27**	68.67	65.14
Reuters-10	50.69	**85.02**	79.94	74.60
Ohscal	35.68	**71.73**	66.12	62.52

表 5 各种方法在数据集上的 AUC 面积

数据集	NB	SSL-LDA+NB	SSL-TF-IDF+NB	EM+NB
Newsgroups	0.4855	**0.8032**	0.7461	0.7124
Reuters-10	0.5262	**0.8804**	0.8143	0.7553
Ohscal	0.4359	**0.7863**	0.7266	0.6618

结果表明，提出的 SSL-LDA+NB 方法的准确性和 AUC 面积结果都明显优于其它方法。在初始标记样本较小的情况下，传统监督学习方法(如 NB 分类器) 则不能实现良好性能，准确性很低。而本文通过 SSL-LDA 模型扩展训练集后，则能够明显提升分类性能，这说明了 SSL-LDA 模型的有效性。另外，与 SSL-TF-IDF 训练集扩展方法的结果表明，LDA 主题表示样本比 TF-IDF 权重更为有效。

5 结束语

在标签集较小情况下的文本分类中，半监督学习方法比监督学习方法更为合适。本文提出了一种基于主题模型的半监督学习模型：SSL-LDA，对一些未标记样本进行分类，以此来扩展标记样本集。为了实现更好的 SSL-LDA 模型，通过 ANOVA 分析了模型中各种参数的重要性，并通过 SPSO 算法来获得最优参数组合。最后，在扩展后的训练集上训练 NB 分类器。在只具有 20%已标记样本的训练集上进行了实验，根据与其它相关方法的分类结果比较，证明了提出方法的优越性。

在今后的工作中，将研究其它主题模型来表示文本，并结合半监督学习，对未标记样本做出更准确的标记，得到一个更加完美的标记样本集。

参考文献：

[1] SUN Xuechen, GAO Zhiqiang, QUAN Zhibin, et al. Short text classification based on semi-supervised learning [J]. Journal of Shandong University of Technology (Science and Technology), 2012, 26 (1): 1-4 (in Chinese). [孙学琛, 高志强, 全志斌, 等. 基于半监督学习的短文本分类方法 [J]. 山东理工大学学报 (自然科学版), 2012, 26 (1): 1-4.]

[2] YANG Mengmeng, HUANG Hao, CHENG Luhong, et al. Short text classification based on LDA topic model [J]. Computer Engineering and Design, 2016, 37 (12): 3371-3377 (in Chinese). [杨萌萌, 黄浩, 程露红, 等. 基于 LDA 主题模型的短文本分类 [J]. 计算机工程与设计, 2016, 37 (12): 3371-3377.]

[3] Kemaiaia A, Merouani HF. Clustering with probabilistic topic models on Arabic texts: A comparative study of LDA and K-means [J]. International Arab Journal of Information Technoklogy, 2016, 13 (2): 133-138.

[4] Triguero I, Garcia S, Herrera F. SEG-SSC: A framework based on synthetic examples generation for self-labeled semi-supervised classification [J]. IEEE Transactions on Cybernetics, 2015, 45 (4): 622-634.

[5] JIANG Yuyan, LI Ping, WANG Qing. An improved Labeled latent Dirichlet allocation model for multi-label classification [J]. Journal of Nanjing University (Natural Sciences), 2013, 49 (4): 425-432 (in Chinese). [江雨燕, 李平, 王清. 用于多标签分类的改进 Labeled LDA 模型 [J]. 南京大学学报 (自然科学), 2013, 49 (4): 425-432.]

[6] LIU Hongbing, LI Wenkun, ZHANG Yangsen. Microblog topic detection based on LDA model and multi-level clustering [J]. Computer Technology and Development, 2016, 26 (6): 25-30 (in Chinese). [刘红兵, 李文坤, 张仰森. 基于 LDA 模型和多层聚类的微博话题检测 [J]. 计算机技术与发展, 2016, 26 (6): 25-30.]

[7] Cai D, Chang L, Ji D. Latent semantic analysis based on space integration [C] //International Conference on Cloud Computing and Intelligent Systems. IEEE, 2013: 1430-1434.

[8] ZHANG Jinrui, CHAI Yumei, ZAN Hongying, et al. Weakly

万方数据

supervised text classification method based on LDA [J]. Computer Engineering and Design, 2017, 38 (1): 86-91 (in Chinese). [张金瑞，柴玉梅，昝红英，等. 基于LDA的弱监督文本分类方法[J]. 计算机工程与设计，2017，38 (1)：86-91.]

[9] Magnusson M, Jonsson L, Villani M, et al. Parallelizing LDA using partially collapsed gibbs sampling [J]. Statistics, 2015, 24 (2): 301-327.

[10] Guo H, Liang Q, Li Z. An improved AD-LDA topic model based on weighted Gibbs sampling [C] //Advanced Information Management, Communicates, Electronic and Automation Control Conference. IEEE, 2017: 1978-1982.

[11] DU Fanghua, JI Junzhong, ZHAO Xuewu, et al. Semi-supervised text classification algorithm based on a feature mapping [J]. Journal of Beijing University of Technology, 2016, 42 (2): 230-235 (in Chinese). [杜芳华，冀俊忠，赵学武，等. 基于特征映射的半监督文本分类算法 [J]. 北京工业大学学报，2016，42 (2)：230-235.]

[12] Geng J, Lu Y, Chen W, et al. An improved text categorization algorithm based on VSM [C] //International Conference on Computational Science and Engineering. IEEE, 2014: 1701-1706.

[13] Trstenjak B, Mikac S, Donko D. KNN with TF-IDF based framework for text categorization [J]. Procedia Engineering, 2014, 69 (1): 1356-1364.

[14] Castillo PA, Arenas MG, Rico N, et al. Determining the significance and relative importance of parameters of a simulated quenching algorithm using statistical tools [J]. Applied Intelligence, 2012, 37 (2): 239-254.

[15] Gao W, Song C, Jiang J, et al. Simplified particle swarm optimization algorithm based on improved learning factors [C] //International Symposium on Neural Networks. Springer, 2017: 321-328.

(上接第3264页)

[9] Ren S, He K, Girshick R. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39 (6): 1137-1149.

[10] Joseph Redmon, Santosh Divvala. You only look once: Unified, real-time object detection [C] //IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779-788.

[11] HE Sihua, SHAO Xiaofang, YANG Shaoqing, et al. Analysis for the cylinder image quality of hyperbolic-catadioptric panorama image system [J]. Laser and Infrared, 2012, 42 (2): 187-191 (in Chinese). [何四华，邵晓方，杨绍清，等. 双曲面折反射全景成像柱面展开图像质量分析 [J]. 激光与红外，2012，42 (2)：187-191.]

[12] SUN Yu, LIU Guiquan. Face recognition method based on HOG and LBP feature [J]. Computer Engineering, 2015, 41 (9): 205-208 (in Chinese). [孙玉，刘贵全. 基于HOG与LBP特征的人脸识别方法 [J]. 计算计工程，2015，41 (9)：205-208.]

[13] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger [C] //IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017.

万方数据