第 40 卷 第 11 期 计 算 机 工 程 2014 年 11 月
Vol. 40 No. 11 Computer Engineering November 2014

·人工智能及识别技术· 文章编号：1000-3428(2014)11-0167-05 文献标识码：A 中图分类号：TP311.12

基于蚁群聚集信息素的半监督文本分类算法

杜芳华[1]，冀俊忠[1]，吴晨生[2]，吴金源[1]

（1. 北京工业大学计算机学院多媒体与智能软件技术北京市重点实验室，北京 100124；
2. 北京市科学技术情报研究所，北京 100048）

摘 要： 半监督文本分类中已标记数据与未标记数据分布不一致，可能导致分类器性能较低。为此，提出一种利用蚁群聚集信息素浓度的半监督文本分类算法。将聚集信息素与传统的文本相似度计算相融合，利用 Top-k 策略选取出未标记蚂蚁可能归属的种群，依据判断规则判定未标记蚂蚁的置信度，采用随机选择策略，把置信度高的未标记蚂蚁加入到对其最有吸引力的训练种群中。在标准数据集上与朴素贝叶斯算法和 EM 算法进行对比实验，结果表明，该算法在精确率、召回率以及 F1 度量方面都取得了更好的效果。

关键词： 文本分类；半监督学习；聚集信息素；自训练；Top-k 策略；随机选择策略

中文引用格式： 杜芳华，冀俊忠，吴晨生，等. 基于蚁群聚集信息素的半监督文本分类算法[J]. 计算机工程，2014，40(11)：167-171.

英文引用格式： Du Fanghua，Ji Junzhong，Wu Chensheng，et al. Semi-supervised Text Classification Algorithm Based on Ant Colony Aggregation Pheromone[J]. Computer Engineering，2014，40(11)：167-171.

Semi-supervised Text Classification Algorithm Based on Ant Colony Aggregation Pheromone

DU Fanghua[1]，JI Junzhong[1]，WU Chensheng[2]，WU Jinyuan[1]

（1. Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology，College of Computer Science and Technology，Beijing University of Technology，Beijing 100124，China；
2. Beijing Institute of Science and Technology Information，Beijing 100048，China）

【Abstract】 There are many algorithms based on data distribution to effectively solve semi-supervised text categorization. However，they may perform badly when the labeled data distribution is different from the unlabeled data. This paper presents a semi-supervised text classification algorithm based on aggregation pheromone，which is used for species aggregation in real ants and other insects. The proposed method，which has no assumption regarding the data distribution，can be applied to any kind of data distribution. In light of aggregation pheromone，colonies that unlabeled ants may belong to are selected with a Top-k strategy. Then the confidence of unlabeled ants is determined by a judgment rule. Unlabeled ants with higher confidence are added into the most attractive training colony by a random selection strategy. Compared with Naïve Bayes and EM algorithm，the experiments on benchmark dataset show that this algorithm performs better on precision，recall and Macro F1.

【Key words】 text classification；semi-supervised learning；aggregation pheromone；self-training；Top-k strategy；random selection strategy

DOI： 10.3969/j.issn.1000-3428.2014.11.033

1 概述

基于机器学习原理的有监督文本分类技术[1]，已在自然语言处理[2]、信息过滤、文本挖掘[3]、情感挖掘、舆情监控等领域具有广泛的应用，但是有监督的文本分类方法一般都需要事先拥有大量的已标记数据来完成分类器的训练。然而，随着大数据时代的到来，人们通常得到的数据往往是海量的且数据分布未知的未标记数据[4]。由于客观资源的限制和手工标记数据的主观性，因此在实际应用中准确地

基金项目： 国家自然科学基金资助项目(61375059，61332016)。

作者简介： 杜芳华(1988－)，男，硕士研究生，主研方向：数据挖掘，机器学习；冀俊忠，教授；吴晨生，研究员；吴金源，硕士研究生。

收稿日期： 2013-11-13 **修回日期：** 2014-01-07 **E-mail：** fanghuadu@126.com

获取大量的已标记数据非常困难。面对这一挑战，基于半监督学习的分类器迅速兴起，并成为有效解决该问题的一种途径[5-6]。半监督学习（semi-supervised learning）是指利用少量已标记数据和大量未标记数据来进行分类的学习技术[7]，其基本思想是基于数据分布上的模型假设[8-9]，建立学习器来对未标记数据进行标记。

目前，已有许多学者对半监督文本分类进行了研究。文献[10-11]将 EM 算法与朴素贝叶斯分类器（Naïve Bayes classifier）相结合实现了一种半监督的文本分类。该算法首先利用已标记样本来训练一个初始的分类器，然后用这个分类器对未标记样本进行不确定性标记，从而使每个未标记样本都部分地属于某一类别。再利用所有的样本训练新的分类器，这一步骤迭代进行直到分类器稳定为止。文献[12]提出了基于自训练的改进 EM 算法。在利用 EM 算法训练中间分类器的过程中，引入了利用中间结果的自训练机制，即将置信度高的未标记样本加到已标记样本集中，能够使未标记数据集不断缩小，从而加快迭代速度，提高分类器的训练性能。但是由于训练出的中间分类器也可能存在比较大的误分类，因此该方法并不适用于已标记类别比较相似的情况。文献[13]提出了一种基于紧密度的半监督文本分类方法。这种方法首先在负例集合中提取出一些置信度较高的负例，然后再根据未标记集合中样本与正例以及抽取出负例的相应紧密度来对可信的负例集合进行相应的扩展，得到用于分类训练所用的负例集合。最后结合训练集中原来给出的正例集，进行文本分类。由于此方法只扩展了负例集合，因此并不适用于正例集合本身就比较少的情况。上述这些半监督文本分类算法本质上都是基于数据统计理论上的分布假设，即将已标记样本与未标记样本视为具有独立同分布的文本数据。但是，这一假设在现实世界中一般难以成立，这是因为大量未标记样本可能来自于与已标记样本的分布不同或完全未知的环境，而且这些未标记样本可能带有噪声。当未标记样本的分布与已标记样本不同时，使用未标记样本可能会降低分类器的性能。因此，如何在已标记样本的分布与未标记样本的分布有较大差异的情况下，提高半监督文本学习器的性能，已经成为半监督学习领域的一个研究热点。

本文借鉴了基于蚁群聚集信息素浓度的半监督分类算法（Aggregation Pheromone Density Based Semi-Supervised Classification，APSSC）的基本思想，提出一种基于蚁群聚集信息素浓度的半监督文本分类算法。针对文本数据，将文本相似度融合到聚集信息素的计算之中，从而扩展了聚集信息素浓度的计算模型。

2 基于蚁群聚集信息素浓度的半监督分类算法

文献[14]将蚁群聚集信息素引入到半监督分类中，提出了 APSSC 算法，它是一种自训练的算法，并在非文本数据集上取得了比较好的效果。聚集信息素（aggregation pheromone）[15]是蚂蚁等昆虫分泌的一种化学物，可用于招引同种个体一起栖息，共同取食，攻击异种对象，并最终形成群集智能的整体行为。

APSSC 算法把每一个样本都看作一只蚂蚁，每个类别当作蚂蚁可能归属的种群，其中已标记蚂蚁 $a_i^{l_k}$ 表示类别 k 中的已标记样本 $x_i^{l_k}\in C_k^0$，未标记蚂蚁 a_i^u 表示未标记样本 $x_i^u\in U$。初始时，训练种群中只包含已标记蚂蚁，各训练种群对未标记蚂蚁释放的聚集信息素浓度均为 0。在每一代的学习过程中，需要对每一只未标记蚂蚁进行如下操作：

（1）计算训练种群 k 向未标记蚂蚁 a_j^u 释放的聚集信息素浓度：

$$\overline{\Delta\tau}_{jk}^{t}=\frac{1}{|C_k^t|}\sum_{x_i^{l_k}\in C_k^t}\Delta\tau^t(x_i^{l_k},x_j^u)\quad\forall j,\forall k \tag{1}$$

其中，$|C_k^t|$ 为训练种群 k 包含的蚂蚁数量；$\Delta\tau^t(x_i^{l_k},x_j^u)$ 为已标记蚂蚁 $a_i^{l_k}$ 向未标记蚂蚁 a_j^u 释放的信息素浓度，计算模型为：

$$\Delta\tau^t(x_i^{l_k},x_j^u)=\frac{1}{(2\pi)^{d/2}(\det(\boldsymbol{\Sigma}_k^t))^{\frac{1}{2}}}\cdot \exp\left(-\frac{1}{2}(x_j^u-x_i^{l_k})^{\mathrm{T}}(\boldsymbol{\Sigma}_k^t)^{-1}(x_j^u-x_i^{l_k})\right) \tag{2}$$

其中，$\boldsymbol{\Sigma}_k^t$ 是训练种群 k 的协方差矩阵；$\det(\boldsymbol{\Sigma}_k^t)$ 是协方差矩阵的行列式；d 是数据集的维度。

（2）更新训练种群 k 向 a_j^u 释放的聚集信息素浓度：

$$\tau_{jk}^t=(1-\rho)\tau_{jk}^{t-1}+\rho\overline{\Delta\tau}_{jk}^{t} \tag{3}$$

其中，$\rho\in[0,1]$ 是蒸发常量。

（3）当所有的训练种群都已经向未标记蚂蚁 a_j^u 释放完聚集信息素时，需要对更新后的 τ_{jk}^t 进行归一化处理，得到归一化的聚集信息素浓度：

$$\mu_{jk}^t=\frac{\tau_{jk}^t}{\sum_{k=1}^{K}\tau_{jk}^t}\quad\forall j,\forall k \tag{4}$$

其中，K 为训练种群总数。

（4）根据算法 1 选取置信度高的未标记蚂蚁加入到对其最有吸引力的一个训练种群 h 中，它将在 $(t+1)$ 代的自训练过程中作为种群 h 的已标记蚂蚁。

（5）当训练种群全部稳定时，自训练过程结束，此时训练种群已经得到了扩展；否则，进入 $t+1$ 代学习。

可见,APSSC 算法的自训练过程是一个利用大量置信度高的未标记蚂蚁对训练种群进行扩展的过程。因此,如何选取置信度高的未标记蚂蚁即算法 1,是 APSSC 算法的核心。自训练过程完成以后,进入测试过程。

在测试过程中,根据种群 k 对测试蚂蚁 a_n 释放聚集信息素浓度 $\overline{\Delta\tau}_{nk}$(见式(5)),把测试蚂蚁 a_n 分到对它释放了最多聚集信息素的种群 h 中,即测试蚂蚁 a_n 归属于种群 h。

$$\overline{\Delta\tau}_{nk} = \frac{1}{|C_k^t|}\sum_{x_i\in C_k^t}\frac{1}{(2\pi)^{\frac{d}{2}}(\det(\sum_k^t))^{\frac{1}{2}}}\cdot \exp\left(-\frac{1}{2}(x_n-x_i)^{\mathrm{T}}(\sum_k^t)^{-1}(x_n-x_i)\right) \tag{5}$$

算法 1　选取置信度高的未标记蚂蚁

输入　未标记蚂蚁 a_j^u

输出　扩展后的训练种群 h

```
1:for  μ_jk^t (k≠h)
2:  { if(μ_jk^t / μ_jh^t = max(μ_jk^t) ≤ 1/K) then
3:  {flag_variable = 1;}
4:  else
5:  {flag_variable = 0;
6:  break;}
7:}
8:if( flag_variable = =1) then
9:  {把未标记蚂蚁 a_j^u 加入到训练种群 h 中;}
10:else
11:  {未标记蚂蚁 a_j^u 不加入到任何一个训练种群中;}
```

3　本文算法

3.1　APSSC 算法在处理文本时存在的问题

APSSC 算法在 Ionosphere、Balance Scale、Sonar 等来自 UCI 的数据集上取得了比较好的效果,证明了蚁群聚集信息素在半监督学习中的有效性,但由于该算法存在着如下不足,使其无法完成文本分类。

(1)由式(2)可以看出,其聚集信息素计算公式中数据集的维度 d 作为 2π 的指数出现在分母中,由于 $2\pi > e$,当 d 取值比较大时,式(2)趋向于 0,因此该算法不能处理像文本这样的高维度数据。

(2)式(2)中 $\det(\sum_k^t)$ 作为一个因子出现在分母当中,这就要求数据集的协方差矩阵 $\sum_k^t$ 必须是可逆的,因此该算法很难适用于具有稀疏性的文本数据。

(3)在算法 1 中,选取置信度高的蚂蚁的判定条件如下:

$$\frac{\mu_{jk}^t}{\mu_{jh}^t = \max(\mu_{jk}^t)} \leqslant \frac{1}{K} \tag{6}$$

也就是说,当遇到类别比较多且相似性较大的文本数据集时,该算法可能会学习不到置信度高的蚂蚁,从而无法达到扩展训练种群的目的。

3.2　SSTCACAP 算法的主要思想

本文对 APSSC 算法进行了扩展,提出了基于蚁群聚集信息素浓度的半监督文本分类算法(Semi-Supervised Text Classification Based on Ant Colony Aggregation Pheromone,SSTCACAP),使其能有效处理文本数据。

首先,根据文本数据高维稀疏的特点,结合文本相似度与聚集信息素之间的关系,将文本相似度作为蚁群聚集信息素浓度计算公式的一个影响因子,扩展了蚁群聚集信息素浓度的计算模型:

$$\Delta\tau^t(x_i^{l_k},x_j^u) = \exp^{-\frac{[1-Similarity(x_i^{l_k},x_j^u)]^2}{2\delta^2}} \tag{7}$$

其中,$Similarity(x_i^{l_k},x_j^u)$ 是余弦相似性公式;δ 是高斯函数的扩散速度系数。

相应地扩展了测试过程中种群 k 对测试蚂蚁释放的聚集信息素浓度计算模型:

$$\overline{\Delta\tau}_{nk} = \frac{1}{|C_k^t|}\sum_{x_i\in C_k^t}\exp^{-\frac{[1-Similarity(x_i,x_u)]^2}{2\delta^2}} \tag{8}$$

其次,根据文本数据集类别多且相似性高的特点,利用 Top-k 策略和随机选择策略扩展了 APSSC 算法中选取置信度高的未标记蚂蚁算法即算法 1,得到扩展后的算法 2,具体的扩展过程如下:

在半监督学习过程中,因为文本数据本身可能携带多个类别特征,并且初始的训练种群包含的种群特征信息非常有限,所以各个训练种群对其释放的聚集信息素的差别很小,甚至可能出现几个种群对其释放的聚集信息素浓度相同的情况。因此,在判定未标记蚂蚁置信度的过程中,采用在 Web 搜索引擎中被广泛使用的 Top-k 查询策略,选取对其最有吸引力的 k 个种群,组成一个候选种群集合 $T_{\mathrm{Top-k}}$。根据集合 $T_{\mathrm{Top-k}}$ 中包含的种群数量来判定未标记蚂蚁的置信度,判定规则为:

$$\frac{|T_{\mathrm{Top-k}}|}{K} \leqslant \beta \tag{9}$$

其中,$|T_{\mathrm{Top-k}}|$ 是 $T_{\mathrm{Top-k}}$ 包含的种群数;K 是训练种群总数;β 是设定的阈值。

当未标记蚂蚁满足判定式(9)时,就判定它的置信度高。这种 Top-k 策略可以有效处理因训练种群对未标记蚂蚁释放的聚集信息素差别较小而可能学习不到置信度高的未标记蚂蚁的情况。

一旦未标记蚂蚁 a_j^u 的置信度被判定为高时,就需要将其加入到 $T_{\mathrm{Top-k}}$ 中的一个种群中,来达到扩展

训练种群的目的。由于 $T_{\text{Top-k}}$ 中的训练种群对未标记蚂蚁的吸引力相差不大,也就是说置信度高的未标记蚂蚁 a_j^u 归属于 $T_{\text{Top-k}}$ 中任何一个训练种群的概率是相等的,因此在 $T_{\text{Top-k}}$ 中随机选择一个种群 k,把 a_j^u 加入到种群 k 中。那么 a_j^u 就可以在下一代的学习过程中作为种群 k 的已标记蚂蚁,从而扩展了训练种群 k。

3.3 SSTCACAP 算法描述

在 SSTCACAP 算法的每一代学习中,需要判定每一只未标记蚂蚁的置信度,并利用置信度高的未标记蚂蚁扩展训练种群。算法描述如下:

算法 2 选取置信度高的未标记蚂蚁

输入 未标记蚂蚁 a_j^u

输出 扩展后的训练种群 h

1:$T_{\text{Top-k}} = \{h\}$

2:for μ_{jk}^t $(k \neq h)$

3:{ if $(\dfrac{\mu_{jk}^t}{\mu_{jh}^t} \geqslant \alpha)$ then

4: {将种群 k 加入到候选种群集合 $T_{\text{Top-k}}$ 中;}

5: }

6:if$(\dfrac{|T_{\text{Top-k}}|}{K} \leqslant \beta)$ then

7: {在候选种群集合 T 中随机选取一个种群 k;/* 当 T 中只有一个训练种群时,k = h */

8: 将未标记蚂蚁 a_j^u 加入到选取的训练种群 k 中;}

9:else

10: {未标记蚂蚁不加入到任何一个训练种群中;}

每一代学习完成以后,便得到了扩展后的训练种群。

4 实验结果与分析

采用著名的语料库 20 Newsgroups,从中挑选 5 个comp. * 类数据来进行实验。对文本的预处理采用了去除停用词(stop word)和基于信息增益(IG)的特征选择等方法。在实验中随机地从每个类别中选取 20% 的文本作为测试样本集,然后再从剩余的文本中选取 50% 作为固定的未标记样本集。算法的评价度量采用精确率(*precision*)、召回率(*recall*)和宏平均 F1 度量(*Macro F*1),其中,宏平均 *F*1 度量的计算公式为:

$$Macro\ F1 = \frac{1}{K}\sum_{k=1}^{K}(F1)_k \tag{10}$$

其中,K 为类别总数。实验中 SSTCACAP 算法使用的参数配置 $\delta = 1.0$,$\alpha = 0.9$,$\beta = 0.6$。EM 算法采用文献[10]的终止条件。本文比较了 SSTCACAP、EM 以及 Naïve Bayes 算法的分类效果。通过不断改变已标记样本集的大小,得出算法的精确率、召回率以及宏平均 F1 度量与已标记样本集大小之间的关系图,如图 1 ~ 图 3 所示。

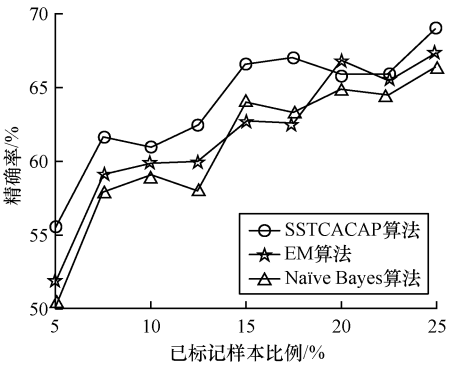

图 1 精确率比较

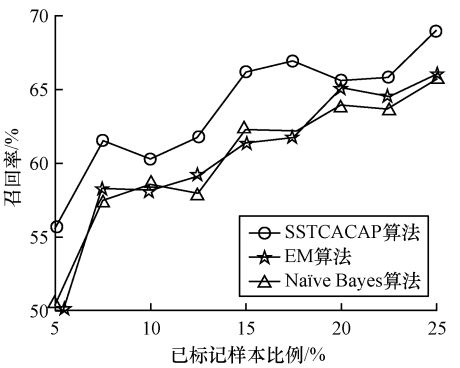

图 2 召回率比较

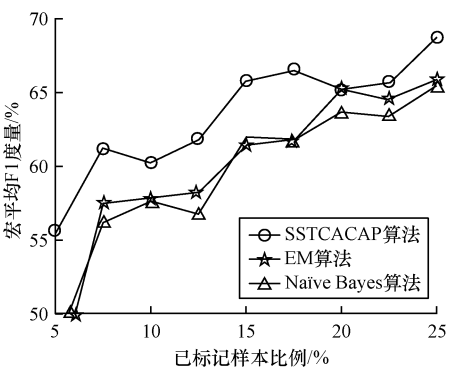

图 3 宏平均 F1 度量比较

由图 1 可以看出,2 种半监督学习方法 SSTCACAP 算法和 EM 算法在引入未标记样本后,都在一定程度上提高了分类的精确性,尤其在已标记样本较少的情况下,半监督学习方法的分类性能表现更好,且本文 SSTCACAP 算法的精确率比传统有监督的方法 Naïve Bayes 算法提高了 5% 。SSTCACAP 算法在大部分情况下,分类精确率都高于 EM 算法和 Naïve Bayes 算法,只有在 20% 已标记样本时,精确率低于 EM 算法。例外产生的一个可能原因是由于某个训练种群的吸引力远大于其他的训练种群,使得这个训练种群学习到了大部分置信度高的未标记蚂蚁,从而出现了相对较多的误分情况。随着已标记样本数量的增加,出现了 Naïve Bayes 算法的分类精确率高于 EM 算法的情况。其中一个可能的原因是此时

已标记样本的数据分布与未标记样本的数据分布存在较大的差异,导致 EM 算法学习到的分类器的性能下降。而本文 SSTCACAP 算法没有出现分类性能低于 Naïve Bayes 的情况,表明 SSTCACAP 算法的稳定性要优于 EM 算法。

由图 2 可以看出,本文 SSTCACAP 算法在召回率方面的表现是最好的,大部分情况下召回率高于 60% ,要优于 EM 算法和 Naïve Bayes 算法。且在 5% 已标记样本时,本文提出的 SSTCACAP 算法分类性能比 EM 算法和 Naïve Bayes 算法分别提高了 7.5% 和 5.1% 。而另一种半监督学习方法 EM 算法和传统的有监督的学习方法 Naïve Bayes 算法在已标记样本低于 20% 时,分类性能出现了交替上升的情况。出现这种情况的原因是基于样本数据分布理论假设的半监督学习方法在无法保证已标记样本与未标记样本的分布一致的情况下,导致半监督学习方法学到的分类器的性能下降。

由图 3 可以看出,本文提出的 SSTCACAP 算法在宏平均 F1 度量(*Macro F*1)方面的分类性能也是最好的。只有在 5% 已标记样本的情况下,SSTCACAP 算法的宏平均 F1 度量低于 60% ,但是 SSTCACAP 算法比 EM 算法和 Naïve Bayes 算法在性能上分别提高了 10.7% 和 8.1% 。且只有在 20% 已标记数据时,另一种半监督学习方法 EM 算法的分类性能达到了本文 SSTCACAP 算法的水平。而 EM 算法在已标记样本低于 20% 时,分类性能与 Naïve Bayes 算法相当,并没有明显的提升。

5　结束语

本文提出一种基于蚁群聚集信息素浓度的半监督文本分类算法。在计算蚁群聚集信息素浓度时,把文本相似度作为其中的一个重要因素,扩展了聚集信息素浓度计算模型。在半监督学习过程中,首先使用 Top-k 策略找出未标记蚂蚁可能归属的种群集合,然后根据判断规则计算并判定未标记蚂蚁的置信度,最后利用随机选择的策略,把符合置信度条件的未标记蚂蚁加入到一个训练种群中,达到对训练种群进行扩展的目的。实验结果验证了本文算法的有效性。

在本文提出的基于蚁群聚集信息素浓度的半监督文本分类算法中,有一个重要的步骤就是需要对算法中的参数进行人工设置,并且参数选择的好坏也会影响分类的效果,因此,下一步的工作是对算法中的参数优化进行研究,以期进一步提高该算法的性能。

参考文献

[1] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002, 34(1):1-47.

[2] 苏金树,张博峰,徐　昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859.

[3] 王建会,王洪伟,申　展,等. 一种实用高效的文本分类算法[J]. 计算机研究与发展,2005,42(1):85-93.

[4] 周志华,王　珏. 机器学习及其应用[M]. 北京:清华大学出版社,2007.

[5] Zhu Xiaojin. Semi-supervised Learning Literature Survey[R]. University of Wisconsin, Technical Report: CS-1530,2008.

[6] Zhu Xiaojin, Goldberg A B. Introduction to Semi-supervised Learning[M]. [S. l.]: Morgan & Claypool Publishers,2009.

[7] Cohen I, Cozman F G, Sebe N. Semi-supervised Learning of Classifiers: Theory, Algorithm, and Their Application to Human-computer Interaction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(12):1553-1567.

[8] Blum A, Chawla S. Learning from Labeled and Unlabeled Data Using Graph Mincuts[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, USA:[s. n.],2001:19-26.

[9] Li Ming, Zhou Zhihua. SETRED: Self-training with Editing[C]//Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hanoi, Vietnam:[s. n.],2005:611-621.

[10] Nigam K, McCallum A K, Thrun S. Text Classification from Labeled and Unlabeled Documents Using EM[J]. Machine Learning,2000,39(2/3):103-134.

[11] Nigam K. Using Unlabeled Data to Improve Text Classification[D]. [S. l.]: Carnegie Mellon University, 2001.

[12] 张博峰,白　冰,苏金树. 基于自训练 EM 算法的半监督文本分类[J]. 国防科技大学学报,2007,29(6):65-69.

[13] 郑海清,林　琛,牛军钰. 一种基于紧密度的半监督文本分类方法[J]. 中文信息学报,2007,21(3):54-60.

[14] Halder A, Ghosh S, Ghosh A. Aggregation Pheromone Metaphor for Semi-supervised Classification[J]. Pattern Recognition,2013,46(8):2239-2248.

[15] Tsutsui S. Ant Colony Optimization for Continuous Domains with Aggregation Pheromones Metaphor[C]// Proceedings of the 5th International Conference on Recent Advances in Soft Computing. Nottingham, UK: [s. n.],2004:207-212.

编辑　任吉慧