

宫小翠,安新颖,单连慧. 基于 Labeled LDA 主题模型的医学文献自动分类法[J]. 中华医学图书情报杂志, 2018, 27(10): 53-58.

DOI:10.3969/j.issn.1671-3982.2018.10.009

· 信息组织与信息服务 ·

基于 Labeled LDA 主题模型的医学文献自动分类法

宫小翠,安新颖,单连慧

[摘要]提出了一种基于 Labeled LDA 主题模型的医学文献自动分类法。以 10 个医学领域的研究文献为案例,通过语料库的设置及参数设置调整模型为最佳,与 SVM 方法进行对比实验。结果显示,无论是准确率还是召回率,基于 Labeled LDA 主题模型的自动分类法均比 SVM 法高出 7.00% 左右,表明基于 Labeled LDA 主题模型的医学文献自动分类法具有较好的医学领域文本分类效果。

[关键词]Labeled LDA; 主题模型; 自动分类; SVM

[中图分类号]TP391.1; R-05

[文献标志码]A

[文章编号]1671-3982(2018)10-0053-06

Labeled LDA topic model-based automatic classification of medical literature

GONG Xiao-cui, AN Xin-ying, SHAN Lian-hui

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

Corresponding author: AN Xin-ying

[Abstract] A Labeled LDA topic model-based automatic classification of medical literature was proposed. The Labeled LDA topic model was adjusted to its best performance by establishing its language repository and setting its parameters with the literature in 10 medical fields as an example and compared with the SVM-based classification of medical literature, which showed that the precision rate and recall rate of Labeled LDA topic model-based automatic classification were about 7.00% higher than those of SVM-based classification, indicating that the efficacy of Labeled LDA topic model-based automatic classification is higher than that of SVM-based classification for the classification of medical literature.

[Key words] Labeled LDA; Topic model; Automatic classification; SVM

随着医疗大数据的爆炸性增长,医学文献共享和知识挖掘的需求越来越迫切。如何提高医学文献分类的效率和质量,快速准确地识别这些海量文献的类别信息,从而挖掘出有价值的信息,成为该领域

一个十分突出和重要的课题。虽然文本自动分类^[1]作为一种成熟、高效的文本分类方法,已经在文本挖掘领域得到了广泛的应用,但分类效果还不够理想。

目前,关于自动分类的研究主要是利用 SVM 或 KNN 等分类算法^[2-3],而基于主题模型的自动分类方法研究比较少,特别是针对医学领域文本自动分类方法的研究更少。

鉴于主题模型要考虑文本的语义信息可靠性较强,本文提出了一种基于 Labeled LDA^[4]主题模型的医学文献自动分类方法。它采用医学文本特定的方式构建训练文本,解决数据不平衡问题,调整模型的参数为最佳,选取了 10 个医学领域的文献进行实

[基金项目] 中国医学科学院中央级公益性科研院所基本科研业务费专项“科技创新环境下医学科研机构科技成果转化能力评价研究”(2017PT63004);国家重点研发计划子课题“精准医学本体语料库构建与扩展”(2016YFC0901902-2)

[作者单位] 中国医学科学院医学信息研究所,北京 100020

[作者简介] 宫小翠(1990-),女,河北保定人,硕士,研究实习员,研究方向为医学数据挖掘与知识发现。

[通讯作者] 安新颖(1978-),女,黑龙江大庆人,博士,研究员,研究方向为医学信息分析。E-mail: an.xinying@imicams.ac.cn

验,并用 Labeled LDA 主题模型与 SVM 方法进行对比实验,对实验结果进行分析总结,验证本文提出的自动分类方法的可靠性和效果。

1 实验设计

自动分类包括词典构建、语料库构建、算法对比与分析等阶段。

在语料库构建阶段,语料库数量庞大、各类别的覆盖度及类之间的不平衡问题突出,需要设计语料库构建规则,对语料库文档进行筛选,在保证各类完整性的基础上尽量减小语料库规模和计算规模,同时充分考虑类之间的平衡;在词典构建阶段需要对词进行规范,对已有词典文本进行处理,从而提高特征向量的纯度,同时还需要对多种分词算法进行对比分析和测试,选择合适的方法提高分词准确率和未登录词的识别率;在分类算法阶段,针对语料库及需要分类的文本特征选择合适的算法,提高分类效率和准确率。

本文主要是针对医学文本自动分类的问题设计实验,通过对比 Labeled LDA 和 SVM 的分类效果,选取好的自动分类方案,提高分类的准确率和效率。

1.1 数据预处理

数据预处理需要解决去除数据中的噪音和数据不均衡性两个关键的问题。

数据下载后,需要进行文本的切词和去停用词处理。由于本文是对医学文本内容进行分类,加之医学词汇有许多特有词和专有词,因此需要构建医学专用词词典,辅助切词过程,提高切词的准确率;需要下载临床医学疾病库中的词汇以及 MeSH 词汇,经人工筛选、规范同义词,并补充如药物、基因等其他词汇构成切词词典。文本切词后,需要构建停

用词词典,将意义不大的词去除。分词工具选用 AnsjSeg。

实际分类数据中的数据往往是非平衡的。如何正确分类非平衡数据集成为文本分类的挑战。非平衡数据集是指在同一个人数据集中某些类的样本数远大于其他类的样本数,其中样本少的类为少数类,样本多的类为多数类^[5]。目前,解决不平衡分类问题的策略分为两大类:一类是从训练集入手,通过改变训练集样本分布降低不平衡程度;另一类是从学习算法入手,根据算法解决不平衡问题中的缺陷,适当地改进算法使之适应不平衡分类问题。本文着眼于使训练样本各类数量基本一致,并从特征选取的角度尽量保留少数类的特征^[6]。特征选取的方法为选取各类别中出现频次较高的词保留,结合人工判断保留频次较低但属该类的专有词,并补充漏掉的各类别特征词和专有词。

1.2 Labeled LDA 模型

1.2.1 模型的训练

实现 Labeled LDA 模型在 JGibbLDA^[7]的基础上进行。当 Labeled LDA 模型在训练阶段对词进行主题采样的时候,不再计算该词在所属文档上未标记类别的概率,避免了 LDA 模型中词会在所有主题上进行分配的问题。将词的主题范围限定在所属文档标记的主题之内,很好地利用了人工标记的主题信息,分类效果比 LDA 模型好。

Labeled LDA 模型训练文本的输入格式与 JGibbLDA 类似,每一行表示一个文档,词与词之间使用空格分隔。Labeled LDA 模型需要在每行文本前面使用“[]”的标识,显示每个文档所属的类别。

实验中 10 个类别的训练文本如图 1 所示。

- [0] 急性胰腺炎 肠炎 肠黏膜 螺旋杆菌 胃癌 肝硬化 肝癌 食管癌 肝炎病毒 大肠癌
- [1] 小细胞肺癌 肺腺癌 支气管扩张 肺损伤 非小细胞 肺栓塞 气道炎症 呼吸道 通气
- [2] 皮炎 硬皮病 银屑病 红斑狼疮 汗孔角化症 白癜风 荨麻疹 瘢痕疙瘩 带状疱疹
- [3] 肾病 肾损害 肾功能衰竭 肾小球 间质性肾炎 肾功能不全 肾小管 多囊肾病 肾盂
- [4] 前列腺 膀胱炎 射精 性功能 输尿管梗阻 泌尿系结石 精子 肾移植 尿路 肾切除术
- [5] 鳞状细胞癌 肺癌 胃癌 卵巢癌 结直肠癌 癌 肿瘤 骨肉瘤 宫颈癌 淋巴瘤 癌症
- [6] 卵巢癌 产妇 怀孕 雌激素 宫颈癌 卵母细胞 卵巢早衰 阴道 胚胎 促卵泡激素 流产
- [7] 下丘脑 帕金森氏病 海马 原发性失眠 星形胶质 脑梗塞 脑脊液 神经 周围神经
- [8] 心血管风险 心脏 心肌梗塞 左心室肥厚 心动过速 心脑血管 心房颤动 血流储备
- [9] 白血病 血清 血小板减少 抗血小板 淋巴 骨髓 红细胞 血栓 血液病 造血干细胞

图 1 实验中 10 个类别的训练文本

图中每 1 行代表一个医学国标类别,第 1 行代表消化病学,第 2 行代表呼吸病学,第 3 行代表皮肤病学,第 4 行代表肾脏病学,第 5 行代表泌尿外科学,第 6 行代表肿瘤学,第 7 行代表妇产科学,第 8 行代表神经病学,第 9 行代表心血管病学,第 10 行代表血液病学。这些文本词由相关的医学文本经过分词、筛选并人工分类后得到。

Labeled LDA 模型使用 Gibbs 抽样方法对主题词分布进行迭代抽取,其最终的准确度受 Gibbs 抽样迭代次数的影响。实际使用 Labeled LDA 模型时,会用某个间隔次数阶梯状地训练 Labeled LDA 模型,然后需要用人工评估模型的训练结果,找最合适的迭代次数,以保证 Gibbs 抽样过程已经收敛,也不进行多余的迭代。

1.2.2 模型的预测

预测文本格式与训练文本格式类似,预测时需要设定相关参数而不需要标记类别号。如预测文本的位置,预测类别数目,迭代次数,选择哪种迭代模型进行结果推断。预测完成后结果在输出文件“.theta”中,每行代表预测集中的一个文档,每一行中的值表示该文档属于不同主题的概率,输出最大值和次最大值对应的类别作为最终的分类结果,次最大值数目过多或过少则不保留此分类。

1.3 SVM 模型

支持向量机(SVM)已在手写数字的识别、扬声器识别、人脸识别、文本分类等许多领域展示了较好的效果,并在解决小样本、非线性及高维模式识别问题中表现出了特有的优势。它的原始设计是针对两类模式识别的问题及现实世界中的多类问题。

经典的 SVM 多类分类算法^[8]主要有一对一方法(1-vs-1)、一对多方法(1-vs-all)、有向无环图方法(DAG-SVM)和二叉树法(BT-SVM)等。从分类器的复杂度来看,一对多和二叉树算法对 n 类分类问题分别需要构造 n 和 $n-1$ 个分类面。构造分类面最多的是一对一和 DAG 算法,对 n 类问题需要构造 $n(n-1)/2$ 个分类面。从分类的精度方面分析,一对一和 DAG 算法比一对多和二叉树算法高。因此在目标分类问题中,分类的精度和分类器的复杂度为一对互相矛盾的指标,因而在解决实际问题中究竟采用哪种分类器,应根据实际的需求而定。当

分类精度要求比较高时可以采用一对一和 DAG 算法,若是需要同时考虑分类精度和分类速度,可以使用一对多和二叉树算法,当然也可以将不同的分类器组合成混合的分类器来协调二者间的需求矛盾^[9]。

SVM 使用时相对简单,首选需要构建数据的输入格式,然后优化最佳参数并选择核函数,利用训练数据构建模型,最后利用模型对测试数据进行分类预测。数据的输入格式如下:

```
[label][index1]:[value1][index2]:[value2]
...
[label][index1]:[value1][index2]:[value2]
...
```

一行即为一条数据记录,label 是种类,即这一条记录所属的类别,通常为整数;index 是有顺序的索引,通常是连续的整数,代表这个类别的特征;value 是特征的值,通常是一些实数。

1.4 结果的评估算法

分类器的性能通常采用评估指标衡量,评估指标是指在测试过程中所使用的一些用来评价分类准确度的量化指标^[10],包括准确率(Precision)、召回率(Recall)和 F1 标准等。

2 实验流程与结果分析

2.1 数据预处理

在 CBM 数据库中,选取 2015 年发表的部分中文数据作为训练数据和测试数据,根据标题和关键词对文献进行自动分类实验。

下载的数据中包括以 R 开头的分类号—cn_auo 字段(图 2)。这是作者给文献的分类号,可将其对应成国标分类号和分类名,作为文献的标准分类结果,并根据需要补充分类名称。

数据需要根据标题和关键词进行分词和去停用词处理,使用 AnsiSeg 分词包时需要准备医学词表和停用词表,分词的结果只限定于词表中包含的词。分词的词典目前已有 30 多万词汇,需要用临床医学知识库中的疾病库词汇及 MeSH 词汇进一步规范和完善词典,然后进行同义词归并,并补充一些较新且属专指的词汇。

对标题和关键词分词后的结果为图 2 中的“ti_kw”字段和“tw_kw”字段,将这两个字段合并的结果

构建训练语料。利用消化病学、肿瘤学、皮肤病学、肾脏病学、泌尿外科学、呼吸病学、妇产科学、神经病学、心血管病学、血液病学等 10 个类别进行自动分类的测试,每个类别下的词汇需要从标题和关键词

的分词结果中遴选,并且辅助人工判断进行补充。选取各个类别中出现频次较高的词保留,并结合人工判断,将频次较低但属该类的专有词保留,并补充漏掉的专有词汇。

ui	ti	ti_kw	tw	tw_kw	cn_auo	dp
2015161229	经桡动脉行PCI治	经桡动脉 pci治疗 术中配合	冠状动脉介入治疗	冠状动脉 介入治	*R473.5	2015
2015161230	外来器械实行集	外来器械 集中管理 存在的	外来器械;集中管理	外来器械 集中管	*R197.39	2015
2015161231	小剂量利妥昔单抗	小剂量 利妥昔单抗 难治性	免疫性血小板减少	免疫性 血小板	*R558.2	2015
2015161232	糖尿病教育对糖	糖尿病教育 糖耐量异常	糖耐量异常;糖尿病	糖耐量异常 糖尿	*R587.1	2015
2015162664	八段锦操练联合	八段锦 常规疗法 神经根型	神经根型颈椎病;八	神经根型 颈椎	*R255.6	2015
2015162667	当归拈痛汤与金	当归拈痛汤 金黄膏 西医常	当归拈痛汤;金黄膏	当归拈痛汤 金	*R255.6	2015
2015162674	人参皂苷Rh2联合	人参皂苷 rh2 肝动脉栓塞	人参皂苷Rh2;肝动	人参皂苷 rh2 肝	*R273	2015
2015162682	针灸干预对重症	针灸干预 重症 手足口病 模	重症手足口病;针灸	重症 手足口病	*R285.5	2015
2015162686	多指标正交试验	多指标 正交试验 乳痛 凝胶	凝胶育剂;基质;多	凝胶育剂 基质	*R283.6	2015

图 2 CBM 数据预处理

2.2 Labeled LDA 模型应用

2.2.1 训练文本的构建

应用 Labeled LDA 模型对文本内容进行分类,首先需要利用训练语料库构建分类模型。分类模型构建的好坏直接影响预测数据分类的准确性,因此模型训练阶段十分关键,其中训练语料库的构建又是模型训练阶段的关键。从语料库的构建入手,选取消化病学、肿瘤学、皮肤病学、肾脏病学、泌尿外科学、呼吸病学、妇产科学、神经病学、心血管病学、血液病学等 10 个类别、300 篇中文数据(共 3 000 条),构建 Labeled LDA 模型的输入文本。首先对这 10 个类别的词汇出现频率进行统计,选取词频大于 5 的词,并通过人工判断,除去对类别判断没有意义的词,如病毒、注射、复发、早期治疗等,保留频次较低但属该类别的特征词,还需要补充漏掉的专有词汇或特征词。最终选取的特征词汇数量见表 1。

表 1 特征词统计数量

学科名称	特征词数量/个	学科名称	特征词数量/个
消化病学	310	呼吸病学	267
肿瘤学	312	妇产科学	310
皮肤病学	256	神经病学	285
肾脏病学	225	心血管病学	254
泌尿外科学	266	血液病	210

2.2.2 模型的训练

运用 Labeled LDA 进行模型训练,设置相关的参数,类目数设为 10,每个类别选取前 50 个最大概率词项,生成的模型文件见图 3。

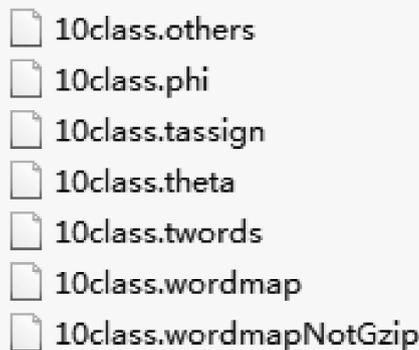


图 3 Labeled LDA 模型文件

模型训练后得到几个文件,其中“. twords”文件为“词项-主题”预测文件,列出每个主题下的特征词及其概率;“. theta”文件为“文档-主题”预测文件,每行代表训练数据集中的一个文档属于不同主题的概率;“. tassign”文件为“文档-词项-主题”预测文件,文件的每行代表一个文档。如“. twords”文件生成 10 个主题,每个主题下包含 50 个词。每个主题的前 10 个词见表 2。

表 2 实验中 10 个主题的前 10 个主题词

主题	主题词
消化病学	伪膜性小肠结肠炎 肠镜 肝囊肿 肝炎 消化性食管炎 胃炎 肝衰竭 蛋白丢失性胃肠病 胆汁性肝硬化 热带性口炎性腹泻
呼吸病学	肝肺综合征 肺鳞癌 肺血栓栓塞 慢性阻塞性肺病 肺真菌病 肺缺血 肺腺癌 肺浆细胞瘤 呼吸道疾病 肺结节
皮肤病学	毛囊炎 药疹 头癣 斑秃 红斑狼疮 皮损 天疱疮 黄褐斑 卟啉病 脚癣
肾脏病学	肾炎 肾癌 肾结石 肾小管 肾损伤 肾囊肿 多囊肾 糖尿病肾病 尿毒症 肾小球
泌尿外科学	包茎 Wilms 瘤 附睾 勃起障碍 包皮损伤 睾丸扭转 精子 前列腺 尿路 尿道癌
肿瘤学	卡铂 顺铂 放射治疗 消化道肿瘤 肠镜 肺癌 乳腺癌 肿瘤 癌症 宫颈癌
妇产科学	产妇 产道异常 白带 闭经 产前筛查 宫腔粘连症 妇产 宫腔积液 分娩 阴道炎
神经病学	帕金森 脑梗死 脊髓神经 下丘脑 大脑海马 美多巴 面肌痉挛 周围神经 神经炎 脑卒中
心血管病学	心室颤动 心脏病 肺动脉高压 肺栓塞 窦性心律 二尖瓣狭窄 房室传导阻滞 肺心病 大血管错位 大动脉炎
血液病学	骨髓肿瘤 白血病 贫血 骨髓增生 白细胞减少 出血性障碍 低凝血酶原血症 低蛋白血症 骨髓纤维化 重链病

2.2.3 分类的预测

从实验的 10 个类别中选取 3 000 条中文文本数据进行分类的测试工作,预测数据的文本格式与训练文本格式类似,只是不需要标记预测数据的类别号。

运用 Labeled LDA 进行分类预测,设置相关参数,类目数设为 10,每个类别选取前 50 个最大概率词项,生成的结果文件见图 4。

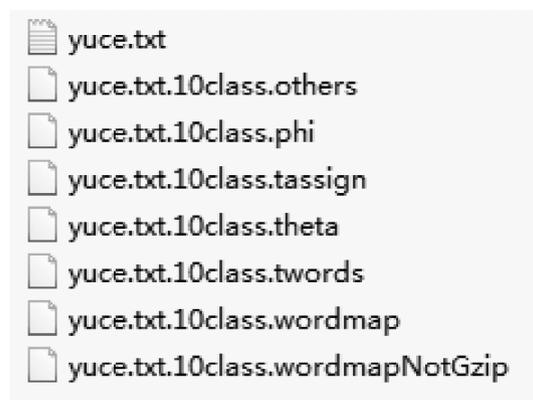


图 4 Labeled LDA 结果文件

读取 .theta 文件,选取其中的最大值分类和次最大值分类作为最终分类。如果次最大值个数大于 2 或是次最大值小于 0.02,则只保留最大值分类。

2.3 SVM 模型应用

选取以上处理过的消化病学、肿瘤学、皮肤病学、肾脏病学、泌尿外科学、呼吸病学、妇产科学、神经病学、心血管病学、血液病学等 10 个类别的 3 000 篇文献数据构建训练文本,每行表示一条中文文献记录,每列代表一个属性词。若是该词在文献中出

现则属性值为 1,没有出现则属性值为 0,属性值为 0 时可以用不用列出。部分文本见图 5。

```

0 1:1 2:1 9:1 10:1 11:1
0 11:1 12:1 13:1 14:1 15:1
0 10:1 13:1 16:1 17:1
1 200:1 203:1 205:1 206:1 211:1
1 202:1 205:1 206:1 209:1 213:1 216:1
1 213:1 214:1 215:1 216:1
2 404:1 409:1 430:1 450:1
2 450:1 451:1 453:1 456:1 457:1
2 453:1 454:1 455:1 458:1 459:1 460:1
2 460:1 461:1 462:1 463:1 464:1 467:1
3 650:1 651:1 656:1 657:1 658:1 659:1
3 462:1 653:1 654:1 656:1 659:1 660:1
3 202:1 656:1 657:1 666:1

```

图 5 SVM 训练文本

选用的 SVM 训练文本为 libsvm-3.20,应用其中的 Windows 版本执行命令 python easy.py svmtrainsvmtest 进行参数寻优,使用核函数 RBF 寻得最优的参数 c 为 0.03125, g 为 0.0001220703125。

预测得到的最终结果保存到 svmtest.predict 文件中,预测得到的总准确率为 79.9333%。

2.4 结果的对比

Labeled LDA 模型与 SVM 模型分类结果如表 3 所示。利用 Labeled LDA 模型分类结果的平均准确率为 87.00%、平均召回率为 87.00%、平均 F1 值为 86.97%,利用 SVM 模型分类结果的平均准确率为 79.93%、平均召回率为 79.95%、平均 F1 值为 79.94%。由此可见,利用 Labeled LDA 模型分类的 3 个测评指标均高于利用 SVM 模型分类的结果,特别是对肾脏病学的分类结果差别更加明显,其

Labeled LDA 模型的 F1 值高出 SVM 模型的 F1 值 8.59%。实验表明,通过训练文档集的改善及参数的调整,主题模型的分类效果要优于 SVM 模型,显示出更好的分类效果。

表 3 Labeled LDA 模型与 SVM 模型的分类结果对比

学科	分类方法	准确率/%	召回率/%	F1 值/%
消化病学	Labeled LDA	86.96	86.67	86.81
	SVM	80.54	80.00	80.27
肿瘤学	Labeled LDA	86.44	85.00	85.71
	SVM	79.47	80.00	79.73
皮肤病学	Labeled LDA	87.46	88.33	87.89
	SVM	80.27	80.00	80.13
肾脏病学	Labeled LDA	87.67	87.67	87.67
	SVM	78.18	80.00	79.08
泌尿外科学	Labeled LDA	87.25	89.00	88.12
	SVM	81.91	79.67	80.77
呼吸病学	Labeled LDA	86.38	86.67	86.52
	SVM	80.47	80.00	80.23
妇产科学	Labeled LDA	87.84	86.67	87.25
	SVM	79.93	80.00	79.96
神经病学	Labeled LDA	87.25	86.67	86.96
	SVM	80.54	79.67	80.10
心血管病学	Labeled LDA	86.67	86.67	86.38
	SVM	79.47	80.00	79.74
血液病学	Labeled LDA	86.09	86.67	86.38
	SVM	78.69	80.00	79.34

3 结论

本文提出了一种基于 Labeled LDA 主题模型的医学文献自动分类法,通过构建医学领域文本的特定训练集,解决数据的不平衡问题,然后进行模型的

训练与测试,并与 SVM 模型进行了对比实验。结果显示,基于 Labeled LDA 主题模型自动分类方法的准确率和召回率均优于 SVM 模型的准确率和召回率,表明此方法在医学领域文本分类方面具有一定的可靠性。但是,本文训练文档集的设定受人为主观因素的干扰较多,下一步工作需要加强研究训练文档集的自动构建方案,进一步提高分类的效率。

【参考文献】

- [1] 苏金树,张博锋,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.
- [2] 刘悦婷,李晓霞,李思璇,等. 基于新改进的 SVM 不平衡数据集分类算法[J]. 石河子大学学报:自然科学版,2018,36(5):637-643.
- [3] 文 武,培 强. 基于 K 中心点和粗糙集的 KNN 分类算法[J]. 计算机工程与设计,2018,39(11):3389-3394.
- [4] 李 伟,马永征,沈 一. 一种解决“中心主题淹没问题”的基于图模型的 Labeled-LDA 文本分类算法[J]. 计算机科学,2014,41(3):223-227.
- [5] CHNWL A N V, JAPKOWICZA. Editorial: special issue on learning from imbalanced data sets[J]. SIGKDD Explorations Newsletters, 2004,6(1):1-6.
- [6] 张玉芳,王 勇,熊忠阳. 不平衡数据集上的文本分类特征选择新方法[J]. 计算机应用研究,2011,28(12):4532-4534.
- [7] 宫小翠,赵迎光,安新颖. 研究前沿识别方法探析[J]. 医学信息学杂志,2015,36(9):47-51.
- [8] 吴恩英,吕 佳. 基于二叉树支持向量机多类分类算法的研究[J]. 重庆师范大学学报:自然科学版,2016,33(3):102-106.
- [9] 郭显娥,武 伟,刘春贵,等. 多类 SVM 分类算法的研究[J]. 山西大同大学学报,2010,26(3):6-8.
- [10] 黄正伟,唐芳艳. 基于 SVM 分类模型的垃圾文本识别研究[J]. 2016,46(7):144-153.

[收稿日期:2018-09-30]

[本文编辑:徐必清]