ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 2,Mar. 2021,pp. 240－247
DOI:10. 16337/j. 1004-9037. 2021. 02. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

MSDL-IEW:面向文本分类的密集度感知主动学习算法

TRAN Baphan[1,2,3]，马菲菲[4]，明晶晶[1]，余秦勇[2,3]，杨　辉[2,3]，李全兵[5]，王永利[1]

(1. 南京理工大学计算机科学与工程学院，南京 210094;2. 中电科大数据研究院有限公司，贵阳 550022;3. 提升政府治理能力大数据应用技术国家工程实验室，贵阳 550022;4. 南京供电公司，南京 210000;5. 中国电子科技网络信息安全有限公司，成都 610041)

摘　要：为了解决文本分类任务中未标注数据无法即时标注及成本过高的问题，提出一种面向文本分类的不确定性主动学习方法。提出MSDL(Measure sample density by LDA)算法对未标注样本密集度进行计算，引入新的度量样本聚集情况的密集度计算方式，在密集度高的样本区域选取初始训练集样本，从而使初始训练集更具代表性；从未标注样本中选取更具不确定性的样本加入到训练集中，并基于信息熵对样本进行加权训练，迭代更新分类器模型，直至达到预期终止条件。实验结果表明，在文本分类任务中，该方法相较于其他传统主动学习算法性能更优。

关键词：文本分类；主动学习；隐含狄利克雷分布；不确定性；密集度

中图分类号：TP391　　**文献标志码：**A

MSDL-IEW: Active Learning Algorithm for Text Classification Based on Density Perception

TRAN Baphan[1,2,3], MA Feifei[4], MING Jingjing[1], YU Qinyong[2,3], YANG Hui[2,3], LI Quanbing[5], WANG Yongli[1]

(1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; 2. CETC Big Data Research Institute Co Ltd, Guiyang 550022, China; 3. Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, Guiyang 550022, China; 4. Nanjing Power Supply Company, Nanjing 210000, China; 5. China Electronics Technology Cyber Security Co Ltd, Chengdu 610041, China)

Abstract: To solve the problem that the unlabeled data in the text classification task cannot be immediately marked and the cost is too high, this paper proposes an active learning method for uncertainty based on text classification. The MSDL (Measure sample density by LDA) algorithm is proposed to calculate the unlabeled sample density, and the new metric sample aggregation situation is introduced. The initial training set sample is selected in the densely sampled region, thus making the initial The training set is more representative. The more uncertain samples from the unlabeled samples are added to the training set, the samples are weighted based on the information entropy, and the classifier model is iteratively updated until the expected termination condition is reached. Experimental results show that this method is better than other traditional active learning algorithms in text classification tasks.

Key words: text classification; active learning; Latent Dirichlet allocation (LDA); uncertainty; density

---

基金项目：国家自然科学基金(61941113)资助项目；中央高校基本科研业务费专项(30916011328, 30918015103)资助项目；南京市科技计划(201805036)资助项目；提升政府治理能力大数据应用技术国家工程实验室开放基金资助项目。
收稿日期：2020-06-04；**修订日期：**2020-11-29

引　言

随着互联网技术的迅速发展,信息数据呈指数趋势增长,更多的人可通过互联网进行消息的接收和传播,产生了越来越多的新闻标题、微博评论和产品评价等即时性文本信息,很难在短时间内得到大量的已标注数据,因此文本信息的自动分类成为当前的研究热点[1]。传统的文本分类算法依赖于已标注数据作为模型的训练集,其分类性能也会随着训练集规模增大而逐步提升,然而上述的即时性文本分类任务面临的重大问题是极少的已标注数据和大量的未标注数据[2],因此传统的文本分类方法不适于解决即时性文本分类任务。

为了解决无法即时标注及成本过高的问题,主动学习得到众多的国内外学者的广泛关注。主动学习算法可以从未标注样本集中选择最有价值的样本交由专家进行标注,从而在不损失训练精度的情况下减少标注样本的代价[3]。主动学习中常用的采样策略可以分为以下几种:基于不确定性的采样策略、基于版本空间缩减的采样策略、基于模型改变期望的采样策略以及基于误差缩减的采样策略[4]。聂嘉贺[5]基于主动学习算法采用RCNN模型作为分类器,提出了一种基于主动学习的文本分类框架,同时结合数据挖掘技术和深度文档向量模型,改进了初始样本选择算法,能从未标注数据集中选取出更能代表样本空间的样本。黄永毅等[6]以支持向量机文本分类方法为基础,提出基于主动学习的交互式支持向量机文本分类方法,即依据已知样本设计分类规则,然后通过主动学习提取不确定样本,二次构建交互式分类器。但现有工作未考虑离群点对分类模型造成的影响,泛化能力不强。因此,本文在构建初始训练集时选择更具代表性的样本加入,引入密集度更好地覆盖整个样本空间。

本文提出了一种面向文本分类的不确定性主动学习算法,为了增加已标注样本的代表性,本文利用隐含狄利克雷分布(Latent Dirichlet allocation,LDA)算法的思想对未标注样本集进行密集度计算,从样本集中密集度高的样本群进行取样,相较于随机选取初始训练集更能体现选取样本的代表性;在后续选取待标注样本时,考虑到在传统主动学习算法中,无法量化样本的不确定程度,本文在对已标注样本进行训练时引入加权项,加快分类模型训练,重复上述步骤,直至分类器达到预期准确度。

1　基于MSDL-IEW的主动学习文本分类模型

采用主动学习算法解决文本分类问题总体上来说分为3个阶段:(1)初始训练集的构建;(2)待标注文本采样;(3)设定终止条件[7]。针对已有方法的不足,本文对基于不确定性的主动学习文本分类方法进行了改进,提出基于MSDL-IEW (Measure sample density by LDA information entropy weighting)的主动学习文本分类算法,该算法的整体模型结构如图1所示。首先,通过LDA算法得到样本的主题分布,并计算密集度,选取样本密集度高的样本加入到初始训练集中,训练一个初始分类器;然后采用最大熵策略采样,选择出"价值"最大的多个文本;然后,读取其对应样本类别标签标注,并计算其权重,加入训练集重新学习一个新的分类器;重复上述过程,直到分类器性能不再提升。

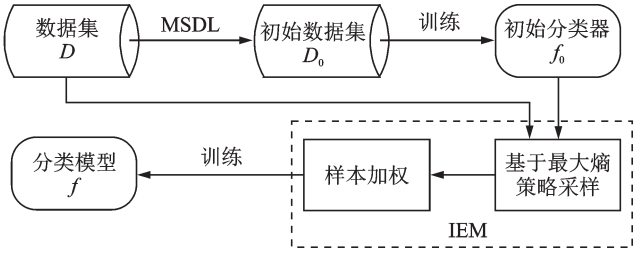

图1　MSDL-IEW主动学习算法框架

Fig.1　MSDL-IEW active learning algorithm model

1.1　基于MSDL的初始训练集选择策略

对于文本分类任务,体现文本类别的特征不一定会显示地出现在文档中,因此提取出文档的主题对文本分类十分重要,本文采用LDA模型计算主题分布[8],并进一步计算数据集的样本密集度,从而可

以将具有代表性的样本加入到初始训练集中。LDA是一种对离散数据集建模的概率主题模型，是一种对文本数据的主题信息进行建模的方法，通过对文本进行一个简短的描述，保留本质的统计信息，有助于高效地处理大规模的文档集。本文中LDA模型可将语料中的每个文档中的信息转换为主题向量，根据每个文档的主题向量进一步计算文档间的语义距离，模型的各符号含义如表1所示。

表1 MSDL算法中各符号的含义

Table 1 Symbolic meaning in MSDL algorithm

符号	含义	符号	含义
α	θ的超参数	m	语料的文档数
β	φ的超参数	n	次数
θ	文档-主题概率分布	T	设置的主题数
φ	主题-词概率分布	N	词数
z	词的主题分配	D_0	初始数据集
w	词向量	D	全部数据集

针对传统主动学习算法可能选出离群点的问题，本文将样本的代表性加入样本价值的度量中，提出MSDL算法(Measure sample density by LDA)进行初始样本集的选取，即结合LDA算法对样本进行密集度计算，并在密集度较高的样本群范围内进行采样，MSDL算法描述见算法1。

算法1 Measure sample density by LDA

输入：主题个数T，超参数α、β，数据集D，初始训练集的文档个数m_0

输出：初始训练集D_0

1. 随机初始化文档中每个词$\boldsymbol{w}$的topic编号z
2. while not Gibbs采样收敛
3. for each $x_i \in D$
4. for each $w \in x_i$
5. 重新采样w的topic
6. 根据Dirichlet分布用式(1,2)得到该文档一个主题分布概率θ
7. for each $p \in \theta$
8. for each $q \in \theta_{q \neq p}$
9. 利用式(3,4)计算文档间距离$d_{\mathrm{js}}(p,q)$
10. 利用式(5)计算文档的密集度$\rho_{\mathrm{text}}(p)$
11. for each $x_i \in D$
12. if len$(D_0) < m_0$
13. $D_0 \leftarrow x_i$
14. else
15. if $\rho_{\mathrm{text}}(x_i) > \min\{\rho_{\mathrm{text}}(x_j) | x_j \in D_0\}$
16. $D_0 \leftarrow x_i$
17. D_0.del(x_j)

算法1的时间复杂度为$O(N_{\mathrm{iter}} m T \bar{l})$，其中$N_{\mathrm{iter}}$为最外层迭代次数，$\bar{l}$为文档的平均长度。

通过对变量z进行Gibbs采样[9]间接估算θ和φ

$$\theta_{mz} = \frac{n_m^{(z)} + \alpha}{\sum_{j=1}^{T} n_m^{(j)} + T\alpha} \tag{1}$$

$$\varphi_{zk}=\frac{n_s^{(k)}+\beta}{\sum_{i=1}^{N}n_z^{(i)}+N\beta} \tag{2}$$

式中:$n_m^{(j)}$表示文本x_m中赋予主题j的词的总数,$n_z^{(i)}$表示词w_i被赋予主题z的总次数。

通过计算文本的相似度进而得到数据的密集度,由于文本的主题分布是文本向量空间的映射,因此在文本的主题表示情况下,计算两个文本的相似度可以通过计算与之对应的主题概率分布来实现。以JS距离[10]公式为标准来度量文本p和q之间的相似度。

$$d_{\mathrm{js}}(p,q)=\frac{1}{2}\left[d_{\mathrm{KL}}\left(p,\frac{p+q}{2}\right)+d_{\mathrm{KL}}\left(q,\frac{p+q}{2}\right)\right] \tag{3}$$

$$d_{\mathrm{KL}}(p,q)=\sum_{j=1}^{T}p_j\ln\frac{p_j}{q_j} \tag{4}$$

根据计算出的文本间相似度,在整个数据集范围内,针对样本p的周围样本的密集度进行计算,根据初始样本集需求,按照密集度递减顺序选取样本。文档p的密集度计算公式为

$$\rho_{\mathrm{text}}(p)=\frac{\left|N_\varepsilon(p)=\{q\in D|d_{\mathrm{js}}(p,q)<\varepsilon\}\right|}{\sum_{i=1}^{N_\varepsilon(p)}d_{\mathrm{js}}(p,q)} \tag{5}$$

式中:ε代表度量样本p和q之间距离是否相近的阈值,$N_\varepsilon(p)$代表与样本p距离小于阈值的样本个数。

1.2 基于信息熵的样本加权

在传统基于不确定的主动学习算法中,假设已有部分已标注样本集$D=\{x_1,x_2,\cdots,x_n\}\subset\mathbf{R}^d$,其中$D_{\mathrm{L}}$为已标注样本集,未标注样本集为$D_{\mathrm{u}}=D-D_{\mathrm{L}}$,系统以已标注样本$D_{\mathrm{L}}$为训练集训练出初始分类器,然后在未标注样本$D_{\mathrm{u}}$中根据不确定性选择部分样本进行标注并将其加入到训练集中,从而训练出新的分类器,整个过程循环多次,直到分类器的某种评价指标达到预设值或循环次数达到预设值为止。

上述基于不确定性的采样(Uncertainty sampling)过程是使用分类器直接估计未标记样本的后验概率值,选择最难被分类器区分的样本[11]。信息熵[12]是基于不确定性采样中最常用的方法,即基于最大熵(Maximum entropy,ME)原则采样,样本的信息熵越大则其不确定性越大,而对当前分类器来说就是最不能确定其所属类别的样本。信息熵的定义为

$$H(x)=-\sum_i p(y_i|x)\log p(y_i|x) \tag{6}$$

式中$p(y_i|x)$表示在给定样本x的情况下其标签属于y_i的可能性。

在上述MSDL算法中,为减少离群点对模型的影响,基于样本密集度选取初始训练集。在后续选择待标注样本时,使用分类器直接估计未标记样本的后验概率值,选择最难被分类器区分的样本,但传统的基于最大熵的样本选择策略无法量化每个样本对分类器的重要性。为此,本文使用加权算法对样本进行重要性标注,形成一种基于信息熵加权(Information entropy weighting,IEW)的主动学习算法。对整个未标注样本集的样本信息熵Z-score规范化[13]后,每个样本的加权系数计算定义为

$$\mathrm{IEW}(x)=\frac{H(x)-\frac{1}{n}\sum_{i=1}^{n}H(x_i)}{\sqrt{\frac{\sum_{i=1}^{n}\left(H(x)-\frac{1}{n}\sum_{i=1}^{n}H(x_i)\right)^2}{n}}} \tag{7}$$

式中:$H(x)$为样本x的信息熵,n为未标注样本集的样本个数,$\mathrm{IEW}(x)$表示基于信息熵计算的每个样

本权重，$x_i \in D_u$代表未标注样本中的第i个样本。

算法2 Information entropy weighting for active learning

输入：初始训练集D_0，预期分类器查准率τ，每次迭代采样文本数c

输出：分类模型f

1. 初始化已标注训练集$D_L = D_0$，待标注样本集$D_{candidates} = \varnothing$

2. $f_0 \leftarrow \text{learn}(D_0)$

3. while f.auc $> \tau$ and D_u is not NULL

4. $\quad D_{candidates} \leftarrow \{\text{argmax}_{x \in D_u} H(x)\}^c$

5. D_u.del($x_i \in D_{candidates}$)

6. for each $x_i \in D_{candidates}$

7. $\quad D_L \leftarrow \text{IEW}(x_i) * x_i$

8. $\quad f \leftarrow \text{learn}(D_L)$

9. return f

算法2的时间复杂度为$O(M_{iter}(m - c*M_{iter}/2))$，其中$M_{iter}$为最外层迭代次数，$m$为文档的个数。

综上所述，本文提出了一种面向文本分类的不确定性主动学习算法。首先，在挑选初始训练集时引入MSDL算法，基于密集度度量选择具有代表性的样本；接着，根据IEW算法按照最大熵策略挑选待标注样本，并计算每个样本的权重，将带有权重的样本加入到模型中训练，在保证目标样本具有不确定性的基础上，又可减少模型训练的迭代次数。

2 实验验证

2.1 实验环境及数据

2.1.1 实验环境

操作平台：Win10；CPU：Intel 双核4.0 GHz；内存：16 GB；硬盘：1 TB；开发语言：Python。

2.1.2 实验数据

数据集采用Sogou实验室提供的分类语料库[14]，该数据集在文本分词、新闻分类等任务中经常作为标准数据集，实验中采用的文本为财经、健康、教育、军事、体育5大类，每个类别中有2 000篇文档，共10 000篇文档。

2.1.3 实验评价

评价分类方法常用查准率(P)和查全率(R)。查准率用于表示分类的正确性，即检测出分类文档中正确分类的文档所占的比率；查全率表示分类的完整性，即所有应分类的文档中被正确分类的文档所占的比率，表达式分别为

$$P_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{8}$$

$$R_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{9}$$

同时为了综合评价分类效果，使用F值

$$F = \frac{P \times R \times 2}{P + R} \tag{10}$$

2.2 实验步骤

为了证明本文所提的MSDL-IEW算法的有效性,本文对比了传统的主动学习算法。对原始样本进行预处理,包括清洗数据、分词和去停用词。选用朴素贝叶斯分类器为基本分类器。MSDL-IEW中使用MSDL算法选择出初始训练集,对分类器进行训练,使用IEW算法对未标注样本进行样本选择并加入权重项,标注后加入到训练集中再次训练,依次迭代。实验运用五折交叉验证方法来选取模型参数,实验中的对比方法包括:

(1) Random + ME。在样本集中随机选择出初始训练集,对分类器进行训练,使用基于最大熵(ME)策略采样,将样本直接进行标注后加入到训练集中再次训练,依次迭代。

(2) MSDL + ME。在样本集中使用MSDL算法选择出初始训练集,对分类器进行训练,使用基于最大熵(ME)策略采样,将样本直接进行标注后加入到训练集中再次训练,依次迭代。

(3) Random + IEW。在样本集中随机选择出初始训练集,对分类器进行训练,使用IEW算法对未标注样本进行样本选择并加入权重项,标注后加入到训练集中再次训练,依次迭代。

2.3 实验结果分析

为了验证本文提出的MSDL-IEW算法是否可达到减少人工标注样本的目的,在上述数据集上进行实验,分别从P、R、F值进行分析,结果如图2所示。

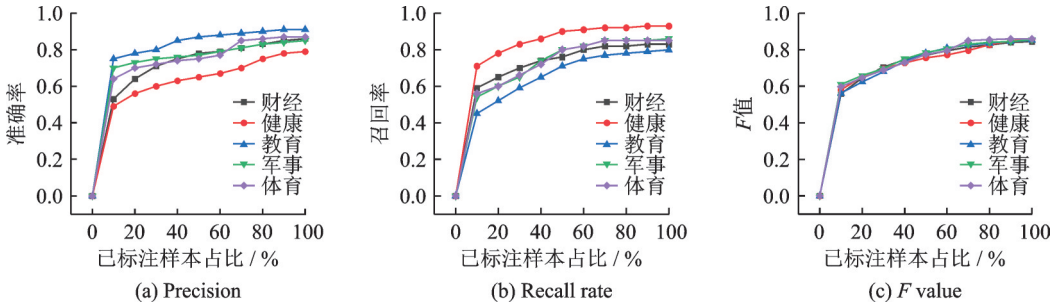

图2 已标注样本数量对查准率、召回率、F值的影响

Fig.2 Influence of the number of labeled samples on the precision, recall rate and F value

从图2可以看出,随着标注样本数量的增加,文本分类的各项评估指标逐渐升高,在标注70%样本之后,各指标不再呈明显上升趋势,说明本文提出的MSDL-IEW主动学习算法可有效地减少人工标注的代价。为了进一步分析本文算法的性能,进行对比实验,在标注样本占比为70%时,各对比算法的评估指标结果如表2所示。

表2 标注样本为70%时各算法的实验结果对比

Table 2 Comparison of experimental results of each algorithm when the labeled sample is 70%

类别	Random+ME			MSDL+ME			Random+IEW			MSDL+IEW		
	P	R	F	P	R	F	P	R	F	P	R	F
财经	0.787	0.797	0.792	0.786	0.796	0.791	0.793	0.805	0.799	0.816	0.827	0.821
健康	0.709	0.889	0.789	0.711	0.895	0.792	0.732	0.917	0.814	0.740	0.923	0.821
教育	0.859	0.750	0.801	0.862	0.754	0.804	0.880	0.763	0.817	0.899	0.778	0.834
军事	0.780	0.814	0.797	0.796	0.813	0.804	0.796	0.833	0.814	0.814	0.853	0.833
体育	0.820	0.828	0.824	0.831	0.830	0.830	0.843	0.845	0.844	0.856	0.856	0.856

由表2可以得出:本文提出的算法在查准率、召回率和F值3个指标上相较于其他算法都表现更好。从Random+ME和MSDL+ME的对比实验中,可看出MSDL相较于传统随机构建初始训练集的方式在各个评价指标上均表现得更为优异,说明MSDL样本选择算法可有效地避免孤立点样本对算法的影响;在Random+IEW和Random+ME的对比实验中,由于传统的Random+ME算法采用基于最大熵的样本选择策略,无法量化每个样本对分类器的重要性,导致模型收敛速度慢,然而本文提出的Random+IEW算法使用加权方式对样本进行重要性标注,这种基于信息熵的样本加权方式可有效地加快模型训练。将MSDL-IEW与传统的主动学习策略Random+ME对比,结果表明MSDL-IEW算法可有效提高算法的查准率和召回率。综上所述,本文提出的MSDL-IEW算法相比于传统主动学习算法更具优势,可以改善主动学习算法中训练集代表性不足的问题,提高分类效果。

3 结束语

本文针对样本标注代价高的问题,提出了一种面向文本分类的不确定性主动学习方法,首先设计MSDL算法选取初始训练集,即通过LDA算法对样本密集度进行计算,从样本集的群体中选择密集度高的部分样本作为初始训练集,建立初始分类器,可保证初始分类器具有一定的代表性;接下来为了加快模型训练速度,本文引入了基于信息熵的样本加权策略,实验表明,IEW策略相较于传统最大熵样本选择策略更优;最后,实验结果表明,本文所提出的算法在查准率、召回率和F值3个指标上均有提升,融合了代表性和不确定性,从而提高了分类器的性能,使用70%的标注样本可接近全部标注样本的训练结果,可有效减少标注样本的成本。

参考文献:

[1] 胡小娟, 刘磊, 邱宁佳. 基于主动学习和否定选择的垃圾邮件分类算法[J]. 电子学报, 2018, 46(1): 203-209.
HU Xiaojuan, LIU Lei, QIU Ningjia. A novel spam categorization algorithm based on active learning method and negative selection algorithm[J]. Acta Electronica Sinica, 2018, 46(1): 203-209.

[2] 周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013, 39(11): 1871-1878.
ZHOU Zhihua. Disagreement-based semi-supervised learning[J]. Acta Automatica Sinica, 2013, 39(11): 1871-1878.

[3] LI Y C, WANG Y L, YU D J, et al. ASCENT: Active supervision for semi-supervised learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 35(5): 868-882.

[4] 谭侃, 高旻, 李文涛, 等. 基于双层采样主动学习的社交网络虚假用户检测方法[J]. 自动化学报, 2017(3): 448-460.
TAN Kan, GAO Min, LI Wentao, et al. Two-layer sampling active learning algorithm for social spammer detection[J]. Acta Automatica Sinica, 2017(3): 448-460.

[5] 聂嘉贺. 基于主动学习的文本分类系统设计与实现[D]. 北京: 北京邮电大学, 2019.
NIE Jiahe. Design and implementation of text classification system based on active learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.

[6] 黄永毅, 龚垒. 基于主动学习的交互式支持向量机文本分类学习方法[J]. 电子技术与软件工程, 2016(14): 168.
HUANG Yongyi, GONG Lei. Interactive support vector machine text classification learning method based on active learning [J]. Electronics and Software Engineering, 2016(14): 168.

[7] YAZHOU Y, MARCO L. A benchmark and comparison of active learning for logistic regression[J]. Pattern Recognition, 2018, 83: 401-415.

[8] 李文波, 孙乐, 张大鲲. 基于Labeled-LDA模型的文本分类新算法[J]. 计算机学报, 2008, 31(4): 620-627.
LI Wenbo, SUN Lei, ZHANG Dakun. Text classification based on labeled-LDA model[J]. Chinese Journal of Computers, 2008, 31(4): 620-627.

[9] 刘伟峰, 杨爱兰. 基于BIC准则和Gibbs采样的有限混合模型无监督学习算法[J]. 电子学报, 2011, 39(z1): 134-139.
LIU Weifeng, YANG Ailan. Unsupervised learning for finite mixture models based on BIC criterion and Gibbs sampling[J].

Acta Electronica Sinica, 2011, 39(z1): 134-139.

[10] 王振振, 何明, 杜永萍. 基于LDA主题模型的文本相似度计算[J]. 计算机科学, 2013, 40(12): 229-232.
WANG Zhenzhen, HE Ming, DU Yongping. Text similarity computing based on topic model LDA[J]. Computer Science, 2013, 40(12): 229-232.

[11] KRISHNAMURTHY A, AGARWAL A, HUANG T K, et al. Active learning for cost-sensitive classification[J]. Proceedings of the 34th International Conference on Machine Learning. [S.l.]: ACM, 2017: 1915-1924.

[12] 骆俊帆, 陈黎, 于中华,等. 长度分布约束下的摘要文本无监督分割算法[J]. 中文信息学报, 2017, 31(4): 138-144.
LUO Junfan, CHEN Li, YU Zhonghua, et al. A length distribution constrained text segmentation for paper abstracts[J]. Journal of Chinese Information Processing, 2017, 31(4): 138-144.

[13] 蔡维玲, 陈东霞. 数据规范化方法对K近邻分类器的影响[J]. 计算机工程, 2010, 36(22): 175-177.
CAI Weiling, CHEN Dongxia. Influence of data normalization methods on K-nearest neighbor classifier[J]. Computer Engineering, 2010, 36(22): 175-177.

[14] Sogou labs. 全网新闻数据[EB/OL].[2012-08-16]. http://www.sogou.com/labs/sogoulab _new/.

作者简介:

TRAN Baphan(1990-),男,硕士研究生,研究方向:知识计算、机器学习,E-mail:2710943547@qq.com。

马菲菲(1984-),女,硕士,高级工程师,研究方向:输配电管理、输电电缆运行检修。

明晶晶(1994-),女,硕士研究生,研究方向:主动学习。

余秦勇(1972-),男,博士,高级工程师,研究方向:云计算、大数据、软件定义网络、网络空间安全、区块链。

杨辉(1969-),男,硕士,高级工程师,研究方向:网络空间安全和区块链。

李全兵(1978-),男,硕士,工程师,研究方向:云计算、大数据、网络空间安全、机器学习。

王永利(1974-),**通信作者**,男,博士,教授,博士生导师,研究方向:机器学习、知识计算、海量数据分析。

(编辑:夏道家)