一种基于权重预处理的中文文本分类算法

何 铠,管有庆,龚 锐
(南京邮电大学 物联网学院,江苏 南京 210003)

摘 要:文本分类是 NLP(natural language processing,自然语言处理)处理技术的重要分支。信息检索、文本挖掘作为自然语言处理领域的关键技术,给人们的生活带来了许多便利,而文本分类正是这些关键技术开展的重要基础。文本分类作为自然语言处理研究的一个热点,其主要原理是将文本数据按照一定的分类规则实现自动化分类。目前常见的文本分类方式主要分为基于机器学习和基于深度学习两种,它们的本质是通过计算机自主学习从而提取文本信息中的规则来进行分类。针对数据量较小、硬件运算能力较低的应用场景,往往使用基于机器学习算法而衍生的文本分类模型。该文以期刊论文作为实验数据,研究中文文本分类问题,在改进传统词频算法的基础上提出了一种基于权重预处理的中文文本分类算法 PRE-TF-IDF(pre-processing term frequency inverse document frequency)。传统词频算法在对词加权时仅考虑词的出现频率而不考虑词在文本中的位置;PRE-TF-IDF 算法在 TF-IDF(term frequency inverse document frequency)算法的基础上增加权重预处理和词密度权重两个环节。实验结果显示 PRE-TF-IDF 算法能够有效提高文本分类的准确性。

关键词:自然语言处理;词频算法;中文文本分类;权重预处理;词密度权重
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)03-0040-06
doi:10.3969/j.issn.1673-629X.2022.03.007

A Chinese Text Classification Algorithm Based on Weight Preprocessing

HE Kai,GUAN You-qing,GONG Rui
(School of Internet of Things,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:Text classification is an important branch of NLP (natural language processing). Information retrieval and text mining,as key technologies in the field of natural language processing,have brought a lot of convenience to people's lives,and text classification is an important basis for the development of those key technologies. Text classification is a hot topic in natural language processing. The main principle of text classification is to automatically classify text data according to certain classification rules. At present,common text classification methods are mainly divided into two types:machine learning and deep learning. Their essence is to extract rules from text information through computer autonomous learning for classification. The text classification model derived from a machine learning algorithm is often used for application scenarios with a small amount of data and low hardware computing power. We take journal papers as experimental data to study the classification of Chinese text. Based on improving the traditional word frequency algorithm,a Chinese text classification algorithm based on weight preprocessing, PRE-TF-IDF (pre-processing term frequency inverse document frequency),is proposed. The traditional word frequency algorithm only considers the occurrence frequency of words but does not consider the position of words in the text when weighing words. Based on the TF-IDF (term frequency inverse document frequency) algorithm,the PRE-TF-IDF algorithm has two additional steps:weight preprocessing and word density weight. Experiment shows that the PRE-TF-IDF algorithm can effectively improve the accuracy of text classification.

Key words:natural language processing;word frequency algorithm;Chinese text classification;weight pretreatment;word density weight

0 引 言

信息检索[1]、文本挖掘[2]作为自然语言处理[3]领域的关键技术,给人们的生活带来了许多便利,而文本分类[4-6]正是这些关键技术开展的重要基础。文本分类作为自然语言处理研究的一个热点,其主要原理是将文本数据按照一定的分类规则实现自动化分类。目前常见的文本分类方式主要分为基于机器学习和基于深度学习两种,它们的本质是通过计算机自主学习从而提取文本信息中的规则来进行分类。针对数据量较小、硬件运算能力较低的应用场景,往往使用基于机器

收稿日期:2021-03-19 **修回日期**:2021-07-21
基金项目:江苏省高校自然科学研究计划项目(05KJD520146)
作者简介:何 铠(1995-),男,硕士研究生,研究方向为自然语言处理;管有庆,副研究员,硕士,硕导,研究方向为深度学习、通信网络等;通信作者:龚 锐(1995-),男,硕士研究生,研究方向为点云深度学习。

学习算法而衍生的文本分类模型。这类模型运行速度快、硬件资源占用量小,并且可以取得不错的分类准确性。机器学习算法是当前文本分类领域研究的一个重点。

目前,几种重要的机器学习算法在文本分类领域都有所应用,如 KNN(K-nearest neighbor,K 临近算法)、SVM(support vector machine,支持向量机)和 LR[7](logistic regressive,逻辑回归)等。将基于词频的 TF-IDF(term frequency - inverse document frequency,词频逆文本频率)算法和 NBC[8](naive Bayes classifier,朴素贝叶斯分类器)进行结合,是基于机器学习原理衍生出的一种被广泛应用的文本分类模型[9]。NBC 分类器原理简明易懂,并且由于其所需要估算的参数较少,对于缺失的数据不敏感,所以在进行小规模文本分类时,有着不错的表现。但该算法也存着一些问题,传统 TF-IDF[10-12]算法仅通过词语在文本中出现的频率来判断词语的重要性,无法根据词语所在的位置信息来进行评估,从而导致文本分类的准确性受到限制[13]。

该文提出一种基于权重预处理的文本分类算法,即 PRE-TF-IDF(pre-processing term frequency inverse document frequency,文本预处理的文本词频和逆文本词频)算法。该算法在传统 TF-IDF 模型的基础上,增加了关键信息权重处理和词密度权重处理两个新的处理环节,增加分类模型对词语位置信息的评估,最终提升了文本分类的准确性。

1　TF-IDF 算法

TF-IDF 算法是一种统计方法,该算法在文本分类中的作用是评估某一个词语对其所在文本的重要性,结合 NBC 最终实现对文本的分类。TF-IDF 主要包含两个部分,TF(term frequency,词频)和 IDF(inverse document frequency,逆文本频率)。具体定义如式(1)所示:

$$\mathrm{TF-IDF} = \mathrm{TF} \times \mathrm{IDF} \tag{1}$$

TF-IDF 算法从定义上看是将 TF 和 IDF 两个数值相乘,其中 TF 的定义式为:

$$\mathrm{TF} = \frac{N(w_i, d)}{S} \tag{2}$$

式中,$N(w_i, d)$ 表示词语 w_i 在文本 d 中出现的次数,S 表示文本 d 中所有词语的总数。用词语 w_i 在文本 d 中出现的次数除以文本 d 中所有词语的总数,当词语 w_i 出现的次数越多,TF 值越大,词语 w_i 对文本 d 越重要;当词语 w_i 出现的次数越少,TF 值越小,词语 w_i 对文本 d 越不重要。但仅凭 TF 值来衡量一个词语区分文本类别的能力会出现一些问题,诸如“的”和“是”,这类词语在每个文本中几乎都具有非常高的出现次数。因此,在评价某个词语对于整个文本集的区分能力时,需要依据 IDF 值来判断。IDF 的定义式为:

$$\mathrm{IDF} = \log\left(\frac{N}{N(w_i)}\right) \tag{3}$$

式中,N 表示文本集中所有文本的总量,$N(w_i)$ 表示文本集中出现过词语 w_i 的文本总数。当 $N(w_i)$ 的数值越小,IDF 值就会越大,表示某个词语在整个文本集中出现的次数越少,则该词将具有很强的区分类别的能力。

TF-IDF 算法的含义是:如果某一个词语在一篇文本中出现的概率很高(即 TF 的数值高),但在其他文本中出现的概率很低(即 IDF 的数值高),则可以认为该词语具有很好的区分类别的能力,可以作为特征词语进行分类。

TF-IDF 算法单纯地认为文本频率越小的单词越具有区别文本类别的能力,而文本频率越大的单词就越无用,这样的思想运用于文本集中的文本是同一类型的文本时就显得不正确了;并且 TF-IDF 算法没有根据词语出现的位置赋予不同的权值。这两方面的不足导致 TF-IDF 算法的精度并不是很高。PRE-TF-IDF 算法在传统 TF-IDF 算法的基础上,增加了关键信息权重处理和词密度权重处理两个新的处理环节,以解决上述两点不足,最终提升文本分类的准确性。

2　基于权重预处理的优化算法(PRE-TF-IDF)

传统 TF-IDF 算法在进行文本分类时,主要存在两个问题。首先,算法仅凭某一个词语在文本和整个文本集中的出现频率来判定这个词语的重要性,IDF 值计算式结构简单,不能有效地反映词语的重要程度,导致算法精度不高。其次,不考虑词语在文本中出现的位置,在词频相同的情况下,关键词语和非关键词语的权重相同,从而导致分类的准确性降低。为解决这两个问题,提出了基于权重预处理的改进 TF-IDF 算法,在文本预处理阶段增加了关键信息权重处理环节,对文本中不同位置出现的词语赋予不同的权重,以解决传统算法无法反映词语位置信息的问题。在特征词语的选取阶段,增加了词密度权重处理环节并改进了 IDF 值的计算方法,以便选取出更具有类别区分能力的特征词语。结合上述两方面的改进,最终提出一种基于权重预处理的优化算法,PRE-TF-IDF 算法。

2.1　关键信息权重处理

(1)算法原理。

针对传统 TF-IDF 算法无法根据特征词在文本中的分布情况而赋予不同权重的问题,基于权重预处理

的 PRE-TF-IDF 优化算法在预处理阶段,对于不同位置出现的词语赋予不同的权重,以突出关键位置词语的重要性,提升区分文本类别的能力。PRE-TF-IDF 算法模型主要针对的应用场景是论文、期刊等文本的分类。这类文本往往包含着标题、发表单位、摘要、关键词等特殊信息,这些段落文字量较少,但对全文起到了概括和提炼的作用。针对这些段落中的词语,赋予更高的权重,有利于更好地选取出具有类别区分能力的特征词语。

文章标题字数一般在 20 字左右,字数较少并且能够简明扼要地概述全文的内容,对标题内的词语赋予高于正文词语的权重。

摘要可以使读者在最短的时间内准确地了解文章的内容,摘要对区分文本类别也起到了十分重要的作用,因此对于摘要段落内出现的词语赋予高于正文词语的权重。

关键词段落常常位于摘要后一段,使用几个词语来概括文章涉及的专业领域,字数较少但概括能力极强,因此需要对关键词赋予高于正文词语的权重。针对不包含摘要和关键词的期刊文本,则不作额外赋值,统一按正文中出现词语赋值。

发表单位常常会出现学校的名称、企业名称或期刊名称等。根据文本所属的出版单位信息,可以大致对文本可能涉及的领域进行一定的评估。例如,一篇发表自理工类学校的文章,该文章属于计算机、电子或能源等领域的可能性要比艺术、教育或法律等领域的可能性高。通过中国大学信息查询系统,获取国内所有高校的名称及其所对应的专业类别,类别包含“综合类”、“理工类”、“师范类”、“财经类”和“农林类”。表1中这五种高校类别与表2中八类文本专业领域分别具有不同的权重配比。

(2)权重处理具体过程。

如图 1 所示,虚线框内的步骤为权重处理的流程。经过预处理后,文本去除了停用词,并以词语的形式保存,词与词之间用空格分隔,段落之间使用换行符分隔。使用预处理后的文本数据作为输入,对文本进行位置权重赋值,赋值规则如下:

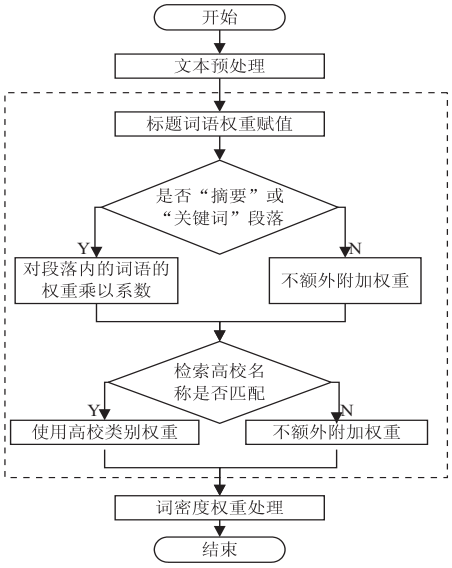

图 1 权重处理流程

对于标题段落内的词语,权重值乘以 2。通过中国大学信息查询系统,获取国内所有高校的名称及其所对应的专业类别。检索“摘要”和“关键词”段落,对“摘要”段落内的词语,权重值乘以 1.5;“关键词”段落内的词语,权重值乘以 2。检索文本中前 300 个词语,与高校名称库进行匹配,若匹配成功,按高校所属类别乘以类别权重,具体类别权重见表 1。若匹配失败则不做额外赋值处理。

表 1 高校类别对应专业领域权重

	电子通信	工程制造	医学	艺术人文	经济管理	政治法律	教育	农业
综合类	0.16	0.14	0.15	0.12	0.13	0.11	0.10	0.09
理工类	0.22	0.25	0.13	0.08	0.10	0.05	0.06	0.11
师范类	0.09	0.09	0.06	0.21	0.12	0.15	0.23	0.05
财经类	0.09	0.05	0.03	0.27	0.27	0.17	0.06	0.06
农林类	0.13	0.14	0.12	0.05	0.21	0.05	0.05	0.25

通过中国大学信息查询系统,收集“综合类”、“理工类”、“师范类”、“财经类”和“农林类”这五类大学,每类 10 所院校。通过统计不同专业研究生数量进行加权平均的方式,求得每个专业领域的权重,绘制成表 1。

在求得待分类文本中所有特征词语出现在不同类别的联合概率分布后,可以得到该文本分别属于各个类别的概率值,再将各个类别的概率值与表 1 的专业领域权重进行相乘,最终取概率值较大的类别,即为待分类文本的类别。

关键信息权重处理中标题段、摘要段和关键词处的权重系数为通过多次实验后,经过分类效果对比,最终确定的具体数值。

2.2 词密度权重处理

传统 TF-IDF 算法单纯地认为文本频率越小的词语越具有区别文本类别的能力,而文本频率越大的词

语就越无用,这样的思想并不是完全正确的。造成这一问题的主要原因是 IDF 值的计算方式较为简单,只考虑了某个词语与其出现的文本数量之间的关系。为解决这一问题,在 PRE-TF-IDF 算法中增加了词密度权重处理环节,该环节的主要原理是通过类别内词密度和类别外词密度两个指标对特征词语的类别区分能力进行衡量。

在传统 TF-IDF 算法中,IDF 的计算式为 $IDF = \log(\frac{N}{N(w_i)})$,其中 N 表示文本集中所有文本的总数,$N(w_i)$ 表示文本集中出现过词语 w_i 的文本总数。仅凭出现过词语 w_i 的文本数和文本集的总数来衡量一个词语区分类别的能力,忽略了词语在类别中和其他类别文本集的出现情况。例如,某一词语在某个类别中出现频率较高,在其他类别中出现频率较低,这样的词语可以被认为是具有较强区分能力的特征词语;若某一词语在某类别中的个别文本中出现频率较高,但在该类的其他文本中出现频率较低,那么可以认为这样的词语存在个例情况,对于整个类别而言并不具有很强的代表作用。出于这两方面的考虑,结合类别内词密度和类别外词密度,在特征词语的选取过程中增加词密度权重处理。

通过 ICD(intra class density,类别内词密度)来表示特征词语在类别内文本中的出现密度权重;用 OCD(outer class density,类别外词密度)来表示特征词语在其他类别文本中出现的密度权重。同时引入 WF(word frequency,词语出现频数),即 $WF(w_i)$、$WF(w_i, C_j)$ 和 $WF(w_i, C_{jk})$ 这三个参数进行计算。

类别内词密度权重 ICD 表示为:

$$ICD = \sqrt{\frac{\sum_{k=1}^{n}\left(WF(w_i, C_{jk}) - \frac{WF(w_i, C_j)}{m}\right)^2}{n-1}} \quad (4)$$

类别外词密度权重 OCD 表示为:

$$OCD = \sqrt{\frac{\sum_{j=1}^{m}\left(WF(w_i, C_j) - \frac{WF(w_i)}{m}\right)^2}{m-1}} \quad (5)$$

式(4)和式(5)中,$WF(w_i)$ 表示特征词语 w_i 在所有类别文本中出现的频数总数,$WF(w_i, C_j)$ 表示特征词语 w_i 在第 j 类中的频数,$WF(w_i, C_{jk})$ 表示特征词语 w_i 在第 j 类中第 k 篇文本中出现的频数,n 表示第 j 类中文本的总数,m 表示文本的类别总数。

类别内词密度权重 ICD 的取值范围为 [0,1]。当 ICD 值趋向于 0 时,表明在类别内特征词语 w_i 的出现密度较为平均,能够很好地体现该类文本的共性;当取值趋向于 1 时,表明特征词语 w_i 在该类文本中出现密度不平均,存在某些文本频数过高的情况,不具有代表性。

类别外词密度权重 OCD 的取值范围也为 [0,1]。当取值趋向于 0 时,表明特征词语 w_i 在不同类别的文本中都有较为平均的出现密度,不能很好地代表某一类文本;当取值趋向于 1 时,表明特征词语在不同类别中的出现密度分布不均,类别区分能力较强。

综上所述,当某个特征词语的 ICD 值趋向于 0,OCD 值趋向于 1 时,代表该词语针对某一类文本具有较强的代表能力。基于传统 TF-IDF 算法,结合 ICD 和 OCD 两种词密度权重,最终形成 PRE-TF-IDF 权重计算函数:

$$PRE\text{-}TF\text{-}IDF = TF \times IDF \times OCD \times (1 - ICD) \quad (6)$$

式(6)中,TF 表示词频,由式(2)定义;IDF 表示逆文本频率,由式(3)定义;OCD 表示类别外词密度权重,由式(5)定义;ICD 表示类别内词密度权重,由式(4)定义。

3　实验结果与分析

采用三个性能评估指标来对基于权重预处理的 PRE-TF-IDF 分类算法进行实验分析,分别是精确率(Precision)、召回率(Recall)和 F1 值(F1 Score)。通过对相同数据集使用传统选取方式和该文优化后的选取方式,进行对照实验并评估。实验运行设备是在安装了 Windows10 专业版操作系统,内存为 16 GB,CPU(central processing unit,中央处理器)主频为 2.8 GHz 的 PC 机上进行的。主要使用的软件环境是基于 Python3.6.7 内核和 Pycharm 2018.12.5 版本,采用的数据集源于复旦大学中文文本分类语料库。在实验过程中,将获取的数据集分为训练集和测试集并且按照 1∶1 的比例进行实验评估。分类类别为 8 种,训练集共 8 800 篇文章,测试集共 8 800 篇文本。文本以“.txt”的格式进行保存,实验共分为 10 组,将这 8 类文本进行等比例缩放,形成 10 组数据集,具体数据集明细如表 2 所示。

表 2　数据集分类明细

类别名	训练集	测试集	所占比例/%
政治法律	740	740	8.4
教育	640	640	7.2
工程制造	1 350	1 350	15.3
医学	1 200	1 200	13.6
艺术人文	1 020	1 020	11.6
电子通信	1 600	1 600	18.3
农业	1 000	1 000	11.4
经济管理	1 250	1 250	14.2

将上述数据按照所占比例的大小,分成 10 组实验

数据集,其中训练集和测试集的比例为 1 ∶ 1,表 3 描述了每组数据的大小。

表 3 数据集分组大小

类别	1	2	3	4	5	6	7	8	9	10
政治法律	74	148	222	296	370	444	518	592	666	740
教育	64	128	192	256	320	384	448	512	576	640
工程制造	135	270	405	540	675	810	945	1 080	1 215	1 350
医学	120	240	360	480	600	720	840	960	1 080	1 200
艺术人文	102	204	306	408	510	612	714	816	918	1 020
电子通信	160	320	480	640	800	960	1 120	1 280	1 440	1 600
农业	100	200	300	400	500	600	700	800	900	1 000
经济管理	125	250	375	500	625	750	875	1 000	1 125	1 250
总计	880	1 760	2 640	3 520	4 400	5 280	6 160	7 040	7 920	8 800

在完成分词后,针对文本中出现的语气助词、人称、标点符号这类对文本特征没有贡献的字词,将其收集、合并,形成了一个停用词列表。通过与停用词列表匹配并将停用词从文本中去除掉,以达到提升程序运行效率、减少干扰因素和提高算法准确性的目的。

3.1 特征词语选取

在实验过程中,特征词语选取的数量对 PRE-TF-IDF 算法的精确率和运行效率都有一定的影响。通过实验计算出兼顾精确率与运行效率的特征词语占比。实验时,将训练集和测试集的数量都定为 8 800,在保持这一条件不变的情况下,通过调整特征词语所占的比重,观察运行效率和精确率的变化,最终选取最佳的特征词语占比。

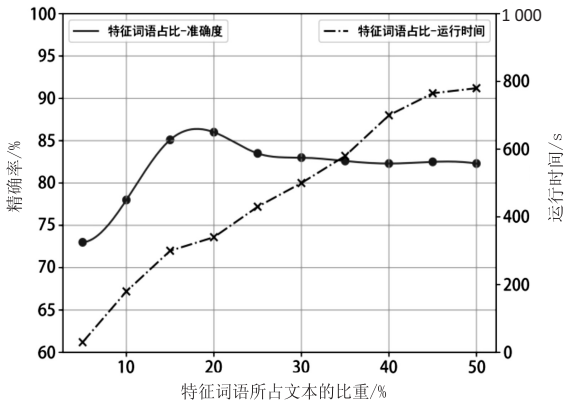

图 2 特征向量占文本比重

根据图 2 可知,在一定范围内,PRE-TF-IDF 算法的分类准确性随着特征词语在文本中的比重增加而增加。但当特征词语占文本比重超过一定数值后,反而使得算法分类的效果下降,对分类的精确率产生负面影响。所以,特征词语在文本中的比重存在一个峰值。随着特征词语在文本中的比重不断增加,算法进行文本分类时所需要的时间也随之变长。最终得出精确率峰值时的平均值为特征词语所占文本的比重 17.57%。此时,能使得 PRE-TF-IDF 算法兼顾分类精确率和运行效率。

3.2 精确率

精确率定义为测试集文本经过算法所分类出的类别与其正确类别之间的百分比,也就是正确分类的文本占所有文本的百分比,其对应的公式如下:

$$P = \frac{TP}{TP + FP} \tag{7}$$

其中,TP 表示被正确分类的文本,FP 表示被错误分类的文本数量,(TP + FP) 即文本的总数量[14]。

这里将上述 8 类文本按照文本数量的大小进行从小到大的排序,随着训练集数量的增加,观察不同算法对于文本分类精确率的表现。实验中将 KNN[15]、LR[7]、TF-IDF[12] 算法和所提出的 PRE-TF-IDF 算法进行对比,结果如图 3 所示。

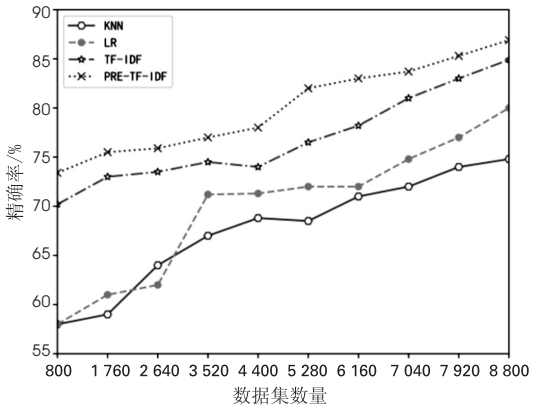

图 3 四种分类算法在不同数据集下的分类精确性对比

由图 3 可知,PRE-TF-IDF 权重预处理优化算法进行分类的准确率比 KNN、LR 和传统 TF-IDF 算法都要高。随着训练集文本量的增加,各个分类模型的精确率也在不断增加。当数据集数量达到最大时,KNN 算法、LR、TF-IDF 和 PRE-TF-IDF 算法对应的精确率分别为 74.8%、80.0%、84.9% 和 86.9%。LR 算法当遭遇特征空间较大时,进行 LR 分类时的性能不是很

好,容易出现欠拟合,精确性不高的情况。传统 TF-IDF 算法结合朴素贝叶斯分类器在进行分类时,虽然精确性相比于 KNN 和 LR 算法有所提升,但是由于传统 TF-IDF 算法存在无法根据词语位置信息分别赋予权重和仅凭文本词频进行 IDF 值计算的问题,所以精确性存在一定的限制。PRE-TF-IDF 算法由于增加了权重预处理和词密度处理两个环节,相比于传统的算法,精确率提升了 2% ~5.5% 。

3.3　召回率

召回率作为一项评估文本分类系统从数据集中分类成功度的指标,用来体现分类算法的完备性,数值越高代表算法的成功度越高。具体公式如下:

$$R = \frac{TP}{TP + FN} \tag{8}$$

其中,TP 表示被正确分类的文本数量,FN 表示应当被分到错误类别中的文本的数量。为了评估 PRE-TF-IDF 算法的召回率指标,同样进行十组不同数据量的对照实验。分别采用 KNN、LR、TF-IDF 算法和 PRE-TF-IDF 算法进行实验。实验结果如图 4 所示。

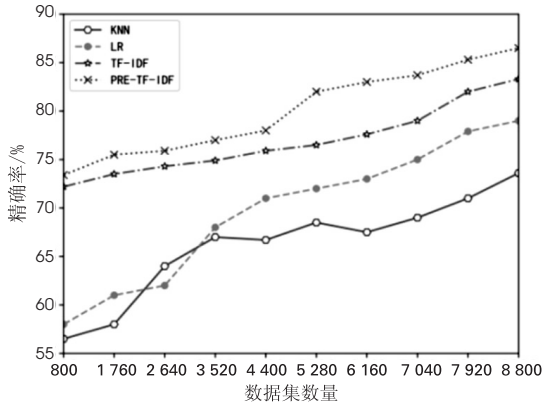

图 4　不同分类算法的召回率对比

由图 4 可以看出,PRE-TF-IDF 的召回率比其他三种文本分类算法的召回率都要高。文本分类的召回率和精确率往往随着数据集的增加而有所提升,召回率与数据集的数量总体上成线性增长。当数据集数量达到最大时,KNN 算法、LR、TF-IDF 和 PRE-TF-IDF 算法对应的召回率分别为 73.6% 、79.0% ,83.3% 和 86.5% 。

3.4　F1 值

F1 值是一个综合考虑精确率和召回率的指标,同时兼顾了分类模型中的精确率和召回率,也可以将这个指标看作是算法精确率和召回率的调和平均。计算公式如下:

$$F1 = \frac{2PR}{R + P} \tag{9}$$

其中,P 表示精确率(Precision),R 表示召回率(Recall),这两个指标反映了分类准确性和成功性两个不同的方面。将精确率和召回率数据进行计算,并绘制成如图 5 所示的折线图。

F1 值通过精确率和召回率计算而得,可以用来评价整个分类器分类效果的优劣。KNN、LR、TF-IDF 和 PRE-TF-IDF 的 F1 值最终分别为 0.742 、0.795 、0.841 和 0.867 。

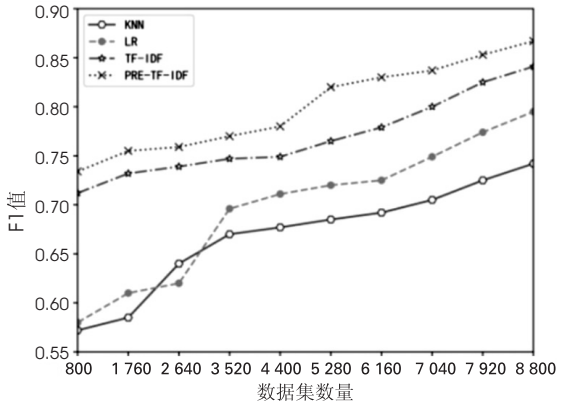

图 5　不同分类算法的 F1 值对比

4　结束语

该文首先介绍了传统 TF-IDF 算法的实现原理,并指出了传统 TF-IDF 算法存在的两个问题,即无法根据词语的位置信息进行权重赋值和仅凭文本词频计算 IDF 值。对此,该文提出了一种基于权重预处理的 PRE-TF-IDF 算法。通过 PRE-TF-IDF 算法中的关键信息权重处理和词密度权重处理两个环节来相应地解决传统 TF-IDF 算法存在的两个问题,并且描述了原理和处理流程。通过实验,将 PRE-TF-IDF 算法与现有的 KNN、LR 和传统 TF-IDF 算法进行对照,在精确率、召回率和 F1 值这三个方面进行对比,对 PRE-TF-IDF 算法进行了评估。

参考文献:

[1]　王继成,萧　嵘,孙正兴,等. Web 信息检索研究进展[J]. 计算机研究与发展,2001,38(2):187-193.

[2]　王继成,潘金贵,张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展,2000,37(5):513-520.

[3]　李　生. 自然语言处理的研究与发展[J]. 燕山大学学报,2013,37(5):377-384.

[4]　武永亮,赵书良,李长镜,等. 基于 TF-IDF 和余弦相似度的文本分类方法[J]. 中文信息学报,2017,31(5):138-145.

[5]　YAO L,MAO C,LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the 33rd AAAI conference on artificial intelligence. Honolulu,Hawaii,USA:AAAI,2019:7370-7377.

[6]　KOWSARI K,JAFARI MEIMANDI K,HEIDARYSAFA M,et

(下转第 53 页)