

Web 文档分类中 TFIDF 特征选择算法的改进

段国仑¹, 谢 钧¹, 郭蕾蕾², 王晓莹¹

(1. 陆军工程大学 指挥控制工程学院, 江苏 南京 210007;

2. 陆军工程大学 通信工程学院, 江苏 南京 210007)

摘 要:随着海量数据资源在网络中的出现, Web 文档分类技术越来越受到重视。在 Web 文档分类的研究中, 特征选择算法有着重要的研究意义。特征选择能有效降低文本向量空间模型的维度, 从而构造出更快, 消耗更低的预测模型。传统的 TFIDF 算法仅仅依靠文档中所包含特征词的词频和逆文档频率来判断该特征词对于文档分类的重要性, 忽略了特征项在类内和类间的分布以及数据集不均衡现象, 从而效果受到制约。针对存在的不足进行改进, 提出了类内分布因子以及类间分布因子。基于类内以及类间因子, 替代逆文档频率, 可以使得改进的表达式能够选择出更加高效的特征词。通过使用 SVM 分类器进行文本分类对比实验, 与改进前的方法相比, 该方法能使 F_1 值得到一定程度的提高, 在不均衡数据集上同样具有较好的分类效果。

关键词: Web 文档分类; 特征选择; TFIDF 算法; SVM

中图分类号: TP391.1

文献标识码: A

文章编号: 1673-629X(2019)05-0049-05

doi: 10.3969/j.issn.1673-629X.2019.05.010

Improvement of TFIDF Feature Selection Algorithm in Web Document Classification

DUAN Guo-lun¹, XIE Jun¹, GUO Lei-lei², WANG Xiao-ying¹

(1. School of Command Control Engineering, Army Engineering University of PLA, Nanjing 210007, China;

2. School of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: With the emergence of massive data resources in the network, Web document classification technology has received more and more attention. In the research of Web document classification, feature selection algorithm has important research significance. Feature selection can effectively reduce the dimensions of the text vector space model, so as to construct a prediction model that is faster and costs less. The traditional TFIDF algorithm only depends on the word frequency and inverse document frequency of the feature words contained in the document to judge the importance of the feature word for document classification, ignoring the distribution of feature items within and between classes and the imbalance of data sets. The effect is limited. In order to improve the existing deficiencies, intra-class distribution factors and inter-class distribution factors were proposed. Based on intra- and inter-class factors, instead of inverse document frequency, improved expressions can be selected for more efficient feature words. By using the SVM classifier for text classification and comparison experiments, this method can increase the F_1 value to a certain extent, and also has better classification effect on the unbalanced data set.

Key words: Web document classification; feature selection; TFIDF algorithm; SVM

0 引 言

随着互联网技术的不断发展以及大规模应用, 网络上的信息资源正以指数级爆炸式增长。Web 已经成为一个巨大的资源海洋。如何管理这些海量信息是一个值得研究的问题。通过 Web 文档分类^[1]手段能

够有效解决这一问题, 因此很多学者在不断努力寻求较好的分类技术。在 Web 文档分类中, 文本信息是最重要的分类信息。分类之前通常将文本表示为向量空间模型(vector space model), 但这种表示方法会使得文本向量维数很高。这带来了较大的特征空间和冗余

收稿日期: 2018-06-19

修回日期: 2018-10-24

网络出版时间: 2018-12-21

基金项目: 国家自然科学基金(61101202)

作者简介: 段国仑(1994-), 男, 硕士研究生, 研究方向为 Web 文档处理、计算机网络; 谢 钧, 博士, 教授, 通讯作者, 研究方向为智能信息处理、计算机网络等。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20181221.1554.052.html>

信息,从而影响文本分类的效率和准确率。

特征选择是通过某种方式或算法选择出有益于分类的特征,去除那些无关或者关联性不强的特征,从而构造出更快,消耗更低的预测模型。

特征选择是文本分类中的关键环节^[2-3]。目前已有很多种特征选择算法用于文本分类,能够较好地解决特征高维性带来的问题,但是针对这些方法仍然有很多需要改进的地方。

1 相关工作

1.1 向量空间模型

文本向量空间模型是为了将每一个文本表示为向量空间中的一个向量,将每一个不同的特征对应为向量空间中的一个维度,而每一维的值就是对应的特征项在文本中的特征值^[4]。向量空间模型中,文档 d 表示为: $V(d) = ((t_1, a_1), (t_2, a_2), \cdots, (t_n, a_n))$ 。其中 t_i 为文档 d 中的特征项, a_i 为 t_i 的特征值,一般取为词频的函数。有了这样的表示以后,就可以用分类器对样本分类。

1.2 特征选择

特征选择^[5]的方式有多种,如过滤式、包裹式、嵌入式等。过滤式独立于后续要使用的模型,只针对数据本身而进行选择特征,从而将特征子集用于分类。由于 Web 文档分类维数较高,所以使用过滤式方法可以有效减少特征选择所用时间。文本特征选择,主要是根据某种准则从原始特征中选择部分最有区分类别能力的特征词。

目前,国内外常用的文本特征选择方法主要有以下几种:词频、TFIDF^[6]、卡方统计量^[7-9]、互信息^[10-11]、信息增益^[12-13]等,也可以通过遗传算法以及特征相关性等进行选择^[14]。

1.3 TFIDF 特征选择算法

如果某个特征词在文档中出现的频率高,并且集中出现在少量文档之中,则认为该特征词具有很好的类别区分能力,适合用来分类。这就是 TFIDF 特征选择算法的主要思想^[15]。TFIDF 是比较常用的特征选择算法,因为它易于将那些词频较高又不在大量文章中出现的词项选出,而且计算复杂度不高,易于实现。

计算出特征项 t 的 TFIDF 的值,然后根据其值来进行特征选择。其计算公式如下:

TFIDF(t) = TF(t) * IDF(t)

(1)

其中, TF(t) 表示特征项 t 在文档中出现的词频数; IDF(t) 表示特征项 t 的逆文档频率数。

下面以一个较为简单的例子对 TFIDF 算法进行分析。假设有三个类 C_1 、 C_2 、 C_3 ,每个类包含 3 篇文档,特征词集合有 t_1 、 t_2 、 t_3 ,分布情况如表 1 所示。

表 1 文档词条频度分布

类别	文档	特征 t_1	特征 t_2	特征 t_3
C_1	D_{11}	3	3	1
C_1	D_{12}	3	3	1
C_1	D_{13}	3	0	1
C_2	D_{21}	0	3	1
C_2	D_{22}	0	0	1
C_2	D_{23}	0	0	1
C_3	D_{31}	0	0	1
C_3	D_{32}	0	0	1
C_3	D_{33}	0	0	1

对于 TFIDF 的计算公式中词频计算方式有很多种,列举如表 2 所示。

表 2 TF 的计算方式

计算方式	表达式
二值	0,1
原始频数	f_t
对数标准化	$1 + \lg f_t$
0.5 标准化	$0.5 + 0.5 \frac{f_t}{\max f_{t'}}$
K 标准化	$K + (1 - K) \frac{f_t}{\max f_{t'}}$

其中, f_t 为特征 t 在文档中出现的频数; $\max f_{t'}$ 为在文档中出现次数最多的特征 t' 对应的频数; K 取值范围为(0,1)。

对于 TFIDF 的计算公式中逆文档频率的计算方式也有很多种,列举如表 3 所示。

表 3 IDF 的计算方式

计算方式	表达式
一元	1
逆文档频率	$\lg \frac{N}{n_t}$
平滑逆文档频率	$\lg \left(\frac{N}{n_t + 1} \right)$
最大值逆文档频率	$\lg \left(\frac{\max n_{t'}}{n_t + 1} \right)$
概率逆文档频率	$\lg \frac{N - n_t}{n_t}$

其中, n_t 为出现特征 t 的文档频数; N 为文档总个数; $\max n_{t'}$ 为文档频数最多的特征 t' 对应的文档频数值。

通过表 2 和表 3 中的表达式,TFIDF 的计算方法可以通过多种组合方式得到,如:

TFIDF(t) = $f_t \cdot \lg \left(\frac{N}{n_t + 1} \right)$

(2)

其中, f_i 表示特征 t 在文档中出现的词频数; n_i 表示包含特征 t 的文档个数; N 表示文档总数。

从表 1 的分布中可以分析得到: t_1 有益于 C_1 的分类, t_2 的作用明显没有 t_1 大, t_3 对于分类毫无意义。三个特征的权重大小应该为:

$$w(t_1) > w(t_2) \gg w(t_3)$$

而通过式 2 计算 t_1 、 t_2 和 t_3 的权重值为:

$$w(t_1) = 3.17 = w(t_2) = 3.17 > w(t_3) = 1.31$$

可以看出算法并没有十分突出分类能力的差异性,而且对于 t_3 这种无助于分类的特征依然获得较大权重。所以,这样的结果并不能选出较好的特征。因此需要对权重值的计算进行调整。

2 TFIDF 特征选择算法的改进

TFIDF 用于特征选择具有一定的区分能力,但同时也存在不足之处。因此,很多研究人员对其进行了改进,如文献[16-17]使用了信息熵进行改进,文献[18-19]使用了类间散度和类内信息熵,Chen 等将其改为 TF-IGF^[20]等。根据以上分析,以及参考前人研究成果,文中分析总结了 TFIDF 特征选择算法的两个缺点:

(1) 没有考虑特征项在类间及类内的分布情况。集中出现在某个类别,而且在这一类中能够大量出现,在其他类中出现相对较少,这样的特征具有较好的分类能力。这些特征本应被选出,但是其 IDF 值可能并不高,通过 TFIDF 难以被选择出来。

(2) 有的类别文档数较多,有的类别却很少,数据集不均衡是十分常见的问题。TFIDF 并没有将类别考虑在内,这就使得选择出的特征词偏向于大类,而针对数据集较小的类的分类特征较少。

针对 TFIDF 的以上缺点,文中对该算法进行改进,不再使用 IDF,而是使用了类内分布因子(α)以及类间分布因子(β)。

$$W(t) = TF(t) \times \alpha(t) \times \beta(t) \tag{3}$$

式 3 考虑了词频、类内分布、类间分布以及数据集不均衡。特征权重值不依赖于类别,因此这是一个全局特征选择算法。

词频(term frequency, TF)反映的是特征项 t 在文档中出现的次数,词频高的有价值。这里选用的词频计算公式如下:

$$TF(t) = 0.5 + 0.5 \frac{f_i}{\max f_i} \tag{4}$$

其中, $\max f_i$ 为在文档中出现次数最多的特征 t 对应的频数值。通过计算式 4,词频较高的值较大,同时也将原来的词频选择作用弱化,并将值归一化。

类内分布因子(α)反映的是特征项 t 在所有类中

分布最广的那一类的分布情况,其值越大表示该特征项在某类中的分布越广。计算公式如下:

$$\alpha(t) = \max(X_1, \dots, X_i, \dots, X_k) \tag{5}$$

其中, $X_i = \frac{n_i}{N_i}$, N_i 表示第 i 类的文档数量, n_i 表示第 i 类中包含特征项 t 的文档数量, k 为类别数目。

类间分布因子(β)反映的是特征项 t 在各类之间的分散程度,值越大说明特征越集中在某一类或几类当中,特征项越有价值。计算公式如下:

$$\beta(t) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \tag{6}$$

其中

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \tag{7}$$

希望选择的特征在各类之间分布差异越大越好。因为存在差异性的特征才是有益于分类的特征, β 就是据此来定义的。这里考虑了数据集不均衡问题:在计算各类分布时,使用的是在各类中所占比重的因式。文献[18]提出过类似作用的因子,但没有考虑各类包含文档数的差异性。对于不均衡的数据集,只考虑各类之间包含特征 t 文档数的差异性显然没有意义。

上文已经对表 1 进行了分析,通过 TFIDF 算法计算得到的权重值并不能很好地代表各个特征对于分类的重要性。而通过式 3 计算 t_1 、 t_2 和 t_3 的权重值为:

$$w(t_1) = 0.22 > w(t_2) = 0.11 > w(t_3) = 0$$

可以看出,改进方法算出的权重值可以将特征的类别区分能力较好地表现出来。文中将通过实验来验证改进方法的有效性。

3 实验

实验在 Anaconda 环境下调用 sklearn、matplotlib、numpy、math、re 等模块实现,所有的实验结果均是在一台 2.50 GHz Intel Core(TM) i7-4710MQ 处理器,8 Gbytes 内存的笔记本电脑上测试获得。

3.1 实验数据集

为了验证该方法的有效性,分别针对两个语料库进行分类实验,包括搜狗中文语料库以及从凤凰网和新浪网上爬取数据自建的语料库。

搜狗中文语料库是由搜狗实验室提供,这里选择了 C000008 财经(300 篇)、C000010 IT(200 篇)、C000013 健康(100 篇)、C000014 体育(47 篇)、C000016 旅游(340 篇)、C000020 教育(350 篇)、C000022 招聘(350 篇)、C000023 文化(1 000 篇)、C000024 军事(500 篇)。

自建语料库主要是通过 python 编程从凤凰网和

新浪网爬取得到,其中包括文化(487 篇)、娱乐(1 182 篇)、财经(934 篇)、健康(1 097 篇)、历史(269 篇)、军事(797 篇)、体育(943 篇)、科技(905 篇)以及社会(897 篇)。

选用的数据集类别较多但样本不多,同时也存在着数据集不均衡现象。

3.2 评价指标

分类模型评估指标有准确率 p 、召回率 r 和 F 度量值^[21]。

$$p = \frac{a}{a+b} \quad (8)$$

$$r = \frac{a}{a+c} \quad (9)$$

$$F_{\beta}(p, r) = \frac{(\beta^2 + 1)pr}{\beta^2 p + r} \quad (10)$$

其中, a 表示实际属于该类并预测正确的个数; b 表示实际不属于该类并预测正确的个数; c 表示实际属于该类但预测错误的个数。

通常希望分类结果中准确率和召回率都要高,而且同等重要。文中使用的是 $\beta = 1$ 的 F_1 度量值,即:

$$F_1(p, r) = \frac{2pr}{p+r}$$

3.3 分类器

选用的分类器是支持向量机 (support vector machine, SVM)。SVM 是一种在缺乏先验知识的条件下,以最小化结构风险为目标,对有限样本进行学习的机器学习方法。支持向量机的基本思想是寻找一个最优超平面或最优超曲面,使得不同类样本之间的间距达到最大^[22]。

支持向量机是目前文本分类中使用较多的分类器。支持向量机擅长解决小样本、高维度的分类问题,而文本分类就是一个高维度的分类问题,所以支持向量机相对较优。

文中实验选用的是 python 工具包 svm. SVC 的线性分类器,损失函数选用 squared hinge loss,使用 L2 正则化,二类向多类的推广采用的是“一对多”的方式。

3.4 实验结果

对选用的两个语料库,分别使用 TFIDF 以及改进后的 TFIDF 特征选择算法。通过计算得到各个特征词的权重值,将权重值排序建立有序特征字典,然后根据字典选取 topN 个特征。在建立文本的向量空间模型时,特征对应的特征值是由 TFIDF 计算得到。于是,可将一个文档转化为一个 N 维向量。这里需要说明的是,实验中选用的特征是根据不同的特征选择算法计算出的权重值决定,即全部特征的子集,而向量空间模型中的特征值在文中均是使用其 TFIDF 值,当然也可以通过其数据式计算得到。也就是说权重值用于

特征选择,而特征值是由于分类。

实验中,按照 2 : 1 的比例将语料库分为训练集和测试集。针对训练集采用三次三折交叉验证方法确定分类器参数。最后在测试集上进行测试,计算分类的准确率、召回率以及 F_1 度量,并对不同特征选择后的分类结果进行比较。实验流程如图 1 所示。

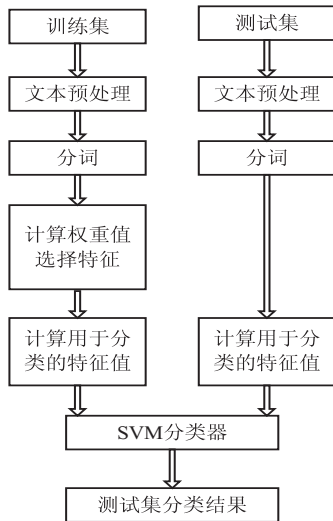


图 1 文本分类实验流程

使用两种特征选择算法在两个语料库中的实验对比结果如图 2 和图 3 所示。图中清晰显示了在选取不同的特征维数时分类 F_1 度量值的对比结果。

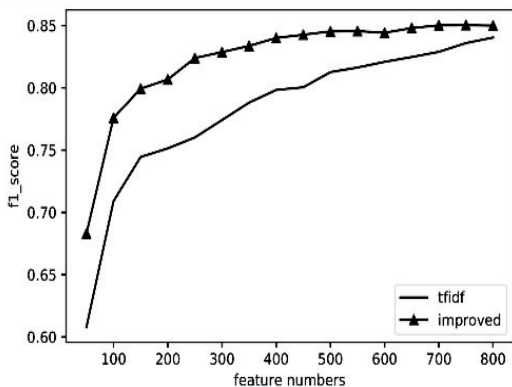


图 2 采用搜狗语料库的 F_1 度量值

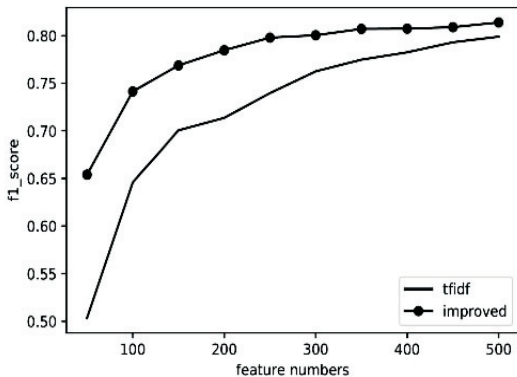


图 3 采用自建语料库的 F_1 度量值

实验结果表明:使用改进后的特征选择算法,在较

少的特征维数下,与传统的 TFIIDF 算法相比能有效提升分类效果。表明提出的算法更加有效,同时也表明特征的重要性与其在类间与类内的分布关系密切。

4 结束语

针对传统的 TFIDF 特征选择算法,主要是针对逆文档频率的计算中没有考虑特征项类内分布、类间分布以及数据集不均衡状况,提出了一种改进方法。通过结合词频、类内分布因子以及类间分布因子,计算各个特征的权重值,并选择数值较高的特征项作为分类特征。通过实验与传统 TFIDF 算法进行了对比。实验中将两种特征选择算法分别应用于搜狗中文语料库以及自建语料库,采用支持向量机作为分类器。最后在测试集上进行测试,得到分类的 F_1 度量值。对比结果表明,改进算法确实对 Web 文档分类效果有了一定的提升。

参考文献:

- [1] 靳小波. 文本分类综述[J]. 自动化博览,2006,23(S1):24-29.
- [2] 奉国和,郑 伟. 文本分类特征降维研究综述[J]. 图书情报工作,2011,55(9):109-113.
- [3] 徐泓洋,杨国为. 中文文本特征选择方法研究综述[J]. 工业控制计算机,2017,30(11):80-81.
- [4] 庞剑锋,卜东波,白 硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,18(9):23-26.
- [5] 计智伟,胡 珉,尹建新. 特征选择算法综述[J]. 电子设计工程,2011,19(9):46-51.
- [6] LAN Man, TAN C L, SU Jian, et al. Supervised and traditional term weighting methods for automatic text categorization[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2009,31(4):721-735.
- [7] 闫健卓,李鹏英,方丽英,等. 基于 X2 统计的改进文本特征选择方法[J]. 计算机工程与设计,2016,37(5):1391-1394.
- [8] 袁 磊. 基于改进 CHI 特征选择的情感文本分类研究[J]. 传感器与微系统,2017,36(5):47-51.
- [9] 王 振,邱晓晖. 混合 CHI 和 MI 的改进文本特征选择方法[J]. 计算机技术与发展,2018,28(4):87-90,94.
- [10] IMAN R L, DAVENPORT J M. Approximations of the critical region of the Friedman statistic[J]. Communications in Statistics,1979,9(6):571-595.
- [11] 成卫青,唐 旋. 一种基于改进互信息和信息熵的文本特征选择方法[J]. 南京邮电大学学报:自然科学版,2013,33(5):63-68.
- [12] 郭亚维,刘晓霞. 文本分类中信息增益特征选择方法的研究[J]. 计算机工程与应用,2012,48(27):119-122.
- [13] HE Haibo, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge & Data Engineering,2009,21(9):1263-1284.
- [14] 庞景安. Web 文本特征提取方法的研究与发展[J]. 情报理论与实践,2006,29(3):338-340.
- [15] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29:167-170.
- [16] 周炎涛,唐剑波,王家琴. 基于信息熵的改进 TFIDF 特征选择算法[J]. 计算机工程与应用,2007,43(35):156-158.
- [17] 陈国松,黄大荣. 基于信息熵的 TFIDF 文本分类特征选择算法研究[J]. 湖北民族学院学报:自然科学版,2008,26(4):401-404.
- [18] 易军凯,田立康. 基于类别区分度的文本特征选择算法研究[J]. 北京化工大学学报:自然科学版,2013,40(s1):72-75.
- [19] YI J, YANG G, WAN J. Category discrimination based feature selection algorithm in Chinese text classification[J]. Journal of Information Science & Engineering,2016,32(5):1145-1159.
- [20] CHEN K, ZHANG Z, LONG J, et al. Turning from TF-IDF to TF-IGM for term weighting in text classification[J]. Expert Systems with Applications,2016,66:245-260.
- [21] 李 航. 统计学习方法[M]. 北京:清华大学出版社,2012.
- [22] CRISTIANINI N, JOHN SHAWE-TAYLOR J. 支持向量机导论[M]. 北京:电子工业出版社,2004.