文章编号:1671-9352(2016)05-0087-07 DOI:10.6040/j.issn.1671-9352.1.2015.E17

结合全局和局部信息的特征选择算法

万中英,王明文,左家莉,万剑怡
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:特征选择方法的优劣直接影响到文本分类的效果。传统的特征选择算法是以全局的方式来选取特征,这种方式忽视了局部特征对分类效果的影响,有时候甚至会导致很多训练文档没有特征。因此,在传统的特征选择方法主要考虑文档集全局特征的基础上,增加词对单篇文档的贡献率的考虑,并结合ALOFT方法,提出了一个结合全局和局部信息的特征选择算法(GLFS)。在路透社文档集及复旦文档集上的实验结果表明,本文提出的算法在保证每个文档都有特征词的同时提高了分类效果。最后讨论了对特征权重的确定方法,经过重新计算特征权重后分类效果有了较大的提高。

关键词:全局和局部信息;特征选择;ALOFT;文本分类;特征权重
中图分类号:TP391 **文献标志码**:A
引用格式:万中英,王明文,左家莉,等. 结合全局和局部信息的特征选择算法[J]. 山东大学学报(理学版),2016,51(5):87-93.

Feature selection combined with the global and local information(GLFS)

WAN Zhong-ying, WANG Ming-wen, ZUO Jia-li, WAN Jian-yi
(School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, Jiangxi, China)

Abstract: Feature selection methods directly affect the effect of text categorization. Traditional feature selection algorithm is based on global approach, ignoring the influence of local features, and even makes a lot of training document has no features. Therefore, the paper proposed a feature selection algorithm combined with the ALOFT method, which unify the traditional globe features and contribution rate of a word to individual document to unify the global and local information(GLFS). Experimental results in the Reuters data set and Fudan data set show that the method can ensure that each document has a characteristic word and improve classification performance. Furthermore, the paper discussed the influence of the new method of feature weights to classification.

Key words: the global and local information; feature selection; ALOFT; text classification; feature weight

0 引言

在文本分类[1-2]过程中,需要利用向量空间模型(vector space model,VSM)将文本表示成由一定数量特征词构成的空间向量。在VSM中,文本集合的空间被视为由一组特征词条所构成的向量空间,每个文档 d 可以由一个空间向量表示。目前的文本分类算法还存在一些不足,其主要原因之一就是特征空间的维数过高。高维的特征向量的处理具有极高的计算复杂度,若用常用的分类算法进行处理,高维带来的噪音会淹没真正的对分类有用的信息,尤其是会产生所谓的"维数灾难问题"。因此要进行维数约简,从约简方式来

收稿日期:2015-09-25; **网络出版时间**: 2016-04-18 11:25:12
网络出版地址: http://www.cnki.net/kcms/detail/37.1389.n.20160418.1125.014.html
基金项目:国家自然科学基金资助项目(61462045,61272212,61462043,61163006);江西省自然科学基金资助项目(20151BAB217014);江西省教育厅科学技术研究项目(GJJ150354)
作者简介:万中英(1977—),女,硕士,副教授,研究方向为信息检索、文本挖掘. E-mail:libby@jxnu.edu.cn

看,可分为两种约简方法:特征选择和特征提取。在原始特征空间中,特征具有语义意义,经过特征选择后,特征仍具有语义意义,而经过特征提取后,就很难再给特征赋予语义意义。

目前传统的特征选择方法[3-5]包括文档频数、互信息、信息增益、期望交叉熵、卡方统计量、文本证据权和各种组合算法[6]等,也有一些对传统算法的改进算法[7-9]。但不管是传统算法还是改进算法,都是对原始项集中的每个项进行独立评估,计算每个项的权值,并且把它们按权值大小进行降序排序,然后根据给定阈值或给定数目选取权值最高的若干项形成新的特征集。这些算法都是以全局的方式来选取特征,而没有考虑局部特征的影响,甚至使得很多训练文档没有特征。ALOFT(at least one feature)方法[10]保证了每篇文档都对最后选择的特征集有贡献,并且所选的特征覆盖了训练集中的每个文档,因为每篇文档至少有一个特征被选择。同时包括 ALOFT 在内的这些方法都没有考虑局部特征的影响,也没有考虑词在单篇文档中的重要性。因此本文在传统的特征选择方法基础上,考虑到词对单篇文档的贡献率(即局部信息),并同时结合 ALOFT 方法,提出了一个结合全局和局部信息的特征选择算法(GLFS)。本文还进一步讨论了对特征权重的确定方法。

1 相关算法

1.1 卡方统计量特征选择算法

通过对各种特征选择方法的比较分析发现,卡方统计量效果较好[11],并且在文献[10]中的实验表明 FEF(是某种传统特征选择算法得到的值)如果采用卡方统计量的实验结果最好,因此本文选用了衡量词与类之间独立性的 x^2(CHI)统计量的特征选择方法。此方法源于统计学的 chi-square 分布,也是从(类,词项)相关性出发,考虑的是每一个类和每一个词项的相关情况 chi-square(t,c),其中 t 表示项(term),c 表示类。也就是 chi-square 值越大,说明 t 和 c 的相关性越高,说明这个项更能反映这个类的特征。其评价函数如下:

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)},$$

其中,$\chi^2(t,c)$表示项 t 与类 c 的关联程度,A 为训练文本集中项 t 和类 c 同时出现的次数,B 为项 t 出现而类 c 不出现的次数,C 为项 t 不出现而类 c 出现的次数,D 为项 t 和类 c 都没有出现的次数,N 为训练文本集中的样本总数。

1.2 ALOFT 算法

Pinheiro 等人在文献[10]中提出了 ALOFT(at least one feature)的特征选择算法。该算法保证了每篇文档都对最后选择的特征集有贡献,并且所选的特征覆盖了训练集中的每个文档,每篇文档至少有一个特征被选择。由于有很多文档会选择相同的特征,所以所选的特征集大小和传统的特征选择算法相比要少。算法描述如下。

算法 ALOFT

步骤 1:计算文档集中每个词的 FEF 值,其中 FEF 是某种传统特征选择算法得到的值。

步骤 2:设置一个空特征集 F_S,针对每一篇文档,选取其中 FEF 值最大的词作为候选词。如果这个词不在 F_S 中,则将其加入 F_S。

步骤 3:用新得到的特征集 F_S 形成新的文档集合。

其中,FEF 是某种传统特征选择算法得到的值,如卡方值。该算法在文本分类实验中取得了和某些传统方法相近或更好的结果,同时保证了每个文档都有特征词,而不会出现空文档的情况。

2 结合全局和局部信息的特征选择算法(GLFS)

2.1 GLFS 算法描述

传统的特征选择算法是对原始项集中的每个项进行独立的评估,计算每个项的权值,并且把它们按权值大小降序排序,然后根据给定阈值或给定数目选取权值最高的若干项形成新的特征集。但这种选择方式会使得很多训练文档没有特征,因此本文考虑结合 ALOFT 算法解决该问题,而且这些方法都是以全局的方式来选取特征,没有考虑局部特征的影响,因此按类别选取的方式能更好地选出反映该类别的特征,同时也考

虑了词在单篇文档中的重要性,即结合了局部信息,最终提出了结合全局和局部信息的特征选择算法(GLFS)。该算法按类别根据 TF * IDF * CHI 值的大小来选择特征,其中,TF 是词频,代表词在文档中的重要性,即是局部信息;IDF 是反文档频率,代表整个数据集的全局信息;而 CHI 是词在类中的贡献率,考虑的是类内信息的重要性。算法的描述如下。

算法 GLFS

步骤 1:按类别计算出每个词的卡方值 $CHI_{h,c}$。

步骤 2:选取一个类 C 的前 N 个卡方值最大特征的加入特征集 F_c。

步骤 3:对于类 C 中的每篇文档进行特征词的选取。选取方法:先算出文档中每个词的卡方值与该词的权值的乘积 CT_h;再选取 CT_h 值最大的词,如果该词不在 F_c 集中,则将该词加入 F_c 中。

步骤 4:重复步骤 2 和步骤 3 直到所有类别的词都已选取完成。

步骤 5:将每个类别的特征集 F_c 合并为一个集合 F_S,该集合没有重复的词。

步骤 6:根据新的特征集 F_S 生成新文档集。

假设文档集有 n 个词 m 个类别,那么 CHI 就是一个 nm 的矩阵,其中,$CHI_{h,c}$ 表示第 h 个词在类 c 中的卡方值,CT_h 表示在当前文档中第 h 个词的卡方值与该词的权值(TFIDF 的值)的乘积。在步骤 2 中选取类 C 的前 N 个卡方值最大的特征加入到特征集 F_c,主要是考虑将对整个类贡献较大的词加入特征集,而 N 的取值不能过大;步骤 3 中不仅考虑词在类中的重要性,同时也考虑了词在该篇文档中的重要性,并保证每篇文档中至少有一个特征词出现。最后将所有特征集 F_c 合并为一个集合 F_S。其伪代码描述如下。

```
1: 加载训练集 D_tr
2: for c = 1 to m do {为每个类的每个词计算卡方值,其中 m 为类别数,V 为词数}
3:    for h = 1 to V do
4:       CHI_{h,c} = FEF(w_h)
5:    end for
6: end for
7: for c = 1 to m do          {按不同类别选取特征}
8:  f = 0                     {该类中特征词的个数}
9:  设 F_c 为一个空特征集
10: 先选取该类的前 N 个卡方值最大特征的加入特征集 F_c;
11:    for all d_i in D_c do   {对该类中的每篇文档选取 CT_i 值最大的词}
12:       bestscore = 0.0
13:          for h = 1 to k do {该篇文档中有 k 个词}
14:             CT_h = CHI_{h,c} * TFIDF(w_h) {TFIDF(w_h)表示 w_h 这个词的 TFIDF 值}
15:             if CT_h > bestscore then
16:                bestscore = CT_h
17:                bestfeature = h
18:             end if
19:          end for
20:       if bestfeature not in F_c then
21:             f = f + 1
22:             F_{c,f} = bestfeature
23:       end if
24:    end for
25: end for
26:    设 F_S 为一个空的特征集合
27: for c = 1 to m
28:    将 F_c 集合中的词加入 F_S 中,重复的不加,最后得到 f_c 个特征词
```

29: end for
30: 设 D_{nv}为一个空文档集
31: for all d in D_{tr} do
32: for $h=1$ to f_c do
33: $d'=d_{F_S}$
34: end for
35: 把文档 d'加入 D_{nv}
36: end for

2.2 特征权重的计算

在特征选择算法中希望选取的是能代表类别特征的词,因此在 GLFS 算法中我们采用 TF * IDF * CHI 的值来选择特征词,综合 TF、IDF、CHI 这些信息,就能选出既代表类别特征又能突出在文档中重要性的词。考虑到在分类中更应该突出词在文档中的重要性,如果采用 TF * IDF * CHI 的值作为特征权重,CHI 的值往往远远大于 TFIDF 的值,则 CHI 的值就会起主要的作用而掩盖 TFIDF 的作用。因此将 TF * IDF * CHI 改为 TFIDF + LOG(CHI + 1),由于在一篇文档中单篇文档的信息更加重要,而且 CHI 的取值有的很大反而会使得单篇文档的信息会被忽略,因此对 CHI 值取对数,从而达到既不忽略单篇文档信息又能适当地加入类内信息的目的。加 1 是为了防止 CHI 值太小而出现负值。因此在进行相应的分类实验之前,要对训练集中每个类中的每篇文档采用 TFIDF + LOG(CHI + 1)重新计算权重。而测试类中的每一篇文档根据类别数的不同利用 TFIDF + LOG(CHI + 1)计算出不同的权重,最后进行相应的分类实验。根据本文提出的特征权重的计算进行相应的 KNN 算法描述如下。

算法 KNN

步骤 1:按类别根据公式 TFIDF + LOG(CHI + 1)重新计算训练集中每篇文档的词的权重。其中 CHI 就是 $\text{CHI}_{h,c}$表示词 h 在类别 c 中的卡方值。

步骤 2:每一篇测试文档 f 根据公式 TFIDF + LOG(CHI + 1)计算在每个类中的特征权重。即如果有 m 个类,则可以得到 m 个不同的特征向量。

步骤 3:将得到到 m 个特征向量分别与相应类别中的文档计算距离。

步骤 4:调用 KNN 算法。

3 实验

3.1 相关说明

本文选用了标准文档集英文路透社文档集 Reuters-21578 和中文的复旦文档集。由于我们选用了 KNN 算法进行分类,K 的取值一般在 5 到 35 之间不等,文档数太少会影响分类效果,因此在每个文档集中选取文档数大于 100 个的类别。最后路透社文档集选取了 14 个类别,训练文档数为 7 583,测试文档数为 2 897;复旦文档集选取了 9 个类别,训练文档数为 7 736,测试文档数为 5 695。经过预处理后路透社文档集有 9 596 个特征词,复旦文档集有 43 168 个特征词。

因为路透社文档集是一个不均衡的多标签的文档集,因此有些类别中的文档数相差较大,如:所选最大类 earn 类的训练文档数为 2 877,而最小类 gnp 类的训练文档数为 101;且有很多文档同时属于多个类别,如 corn 类别中的文档几乎同时属于 grain 类。这些都将影响分类效果,因此在后续的 KNN 分类算法中路透社文档集选取 $K=30$ 分类效果较好,而复旦文档集选取 $k=5$ 的分类效果较好。

文中采用宏平均 F1 值和微平均 F1 值来评价分类效果,分别用 MacF1 和 MicF1 来表示。下述实验中路透社文档集用 LT 表示,复旦文档集用 FD 表示。

3.2 采用卡方统计量进行特征选择

采用卡方统计量进行特征选择,路透社文档集分别选取了 1 000,500,400,300,200 和 100 个特征,再采用 KNN 进行分类,得到的分类结果如表 1 所示。在复旦文档集中分别选取了 1 600,1 400,1 200,1 000,800,500,400,300,200 和 100 个特征,再采用 KNN 进行分类,得到如表 2 所示的分类结果。

表 1　在 LT 上分类的 F1 值
Table 1　The value of F1 in LT

维数	1 000	500	400	300	200	100
MacF1	0. 706 45	0. 719 81	**0. 727 95**	0. 727 59	0. 708 36	0. 694 24
MicF1	0. 849 30	0. 846 98	**0. 863 21**	0. 857 29	0. 841 34	0. 836 32

表 2　在 FD 上分类的 F1 值
Table 2　The value of F1 in FD

维数	1 600	1 400	1 200	1 000	800	500	400	300	200	100
MacF1	0. 845 5	**0. 846 1**	0. 844 7	0. 832 5	0. 844 1	0. 826 7	0. 779 2	0. 803 3	0. 796 7	0. 709 6
MicF1	0. 869 8	**0. 870 9**	0. 869 9	0. 857 1	0. 868 1	0. 852 7	0. 799 7	0. 824 0	0. 821 1	0. 728 4

从表 1 和表 2 可以看出,在路透社文档集上当维数降到 400 维的时候分类性能是最好的,在复旦文档集中当维数降到 1 400 维时性能最好,且两个文档集都随着维数的减少性能有所下降;而造成在两个文档集上随着维数下降而性能有所下降的一个主要原因是大量空文档的出现,如下面图 1 和图 2 所示,随着维数越来越少,训练集的空文档数越来越多。在路透数据集中,当维数降到 100 时空文档数达到 114 篇,在复旦数据集中,当维数降到 100 时空文档数达到 405 篇,这将极大地影响分类效果。

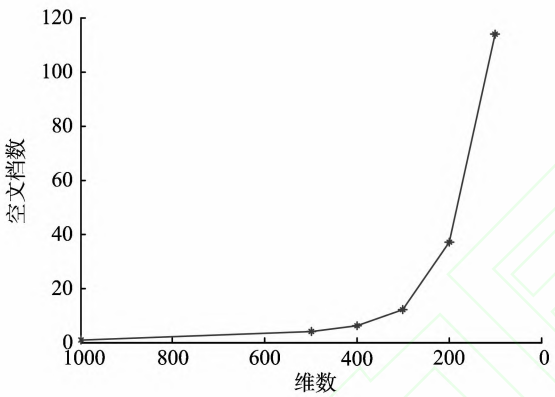

图 1　LT 中出现的空文档数
Fig. 1　The number of empty document in LT

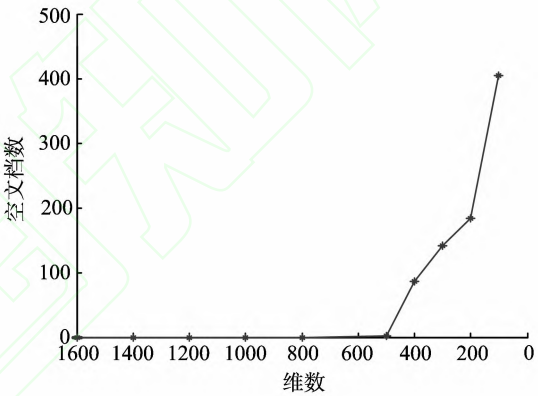

图 2　FD 中出现的空文档数
Fig. 2　The number of empty document in FD

3. 3　GLFS 算法

在同等条件下采用 GLFS 算法进行实验。GLFS 算法中的 N 表示卡方值按由大到小排序时的前 N 个特征词,由图 3 和图 4 可以看出随着 N 值增加,最后选取的特征维数也随之增加。

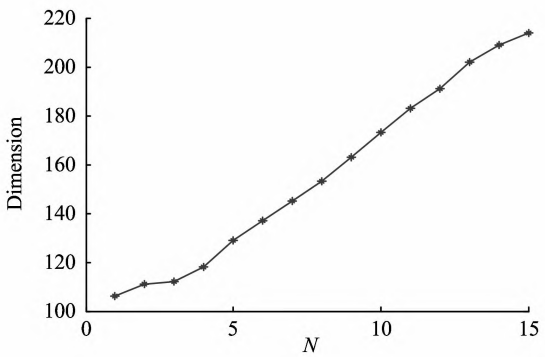

图 3　LT 的 N-Dimension 图
Fig. 3　Figure N-Dimension of LT

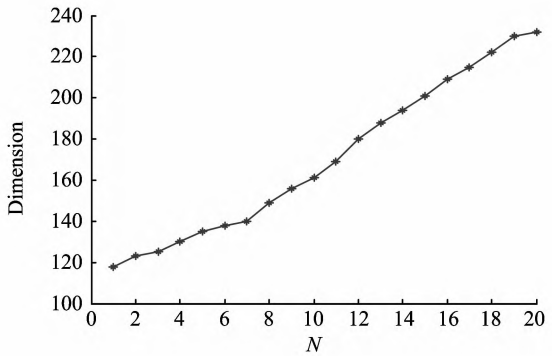

图 4　FD 的 N-Dimension 图
Fig. 4　Figure N-Dimension of FD

特征选择后进行 KNN 分类,其分类效果如图 5 和图 6 所示。从图 5 可看出,在路透社文档集中,当 N = 9 时性能达到最好,此时维数降到了 163 维;从图 6 可看出,在复旦文档集中,当 N = 14 时性能达到最好,此时降到了 194 维;而在这两种文档集上,训练集中都没有出现空文档,且当性能达到最好时随着 N 值的增大,性能基本保持稳定。而从表 1 和表 2 可以看出采用 CHI 方法选取特征后进行分类,当性能达到最好时,性能会随着维数的减少而下降。

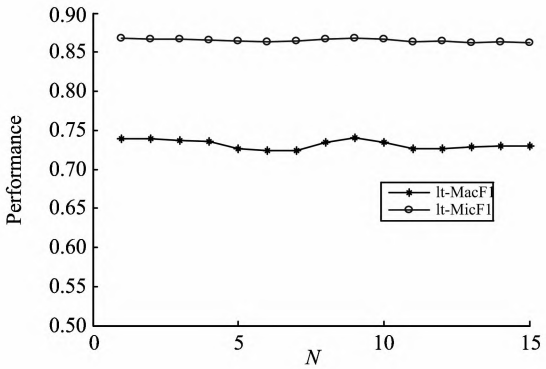

图 5 LT 的实验结果

Fig. 5 The experimental result of LT

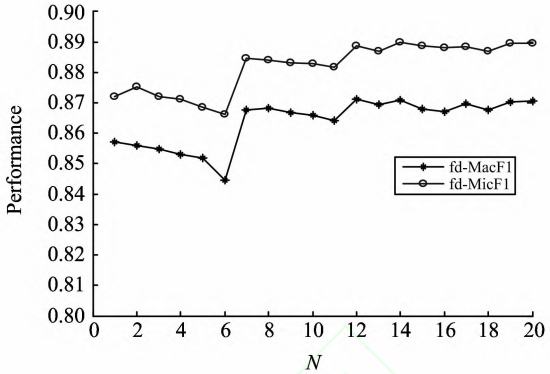

图 6 FD 的实验结果

Fig. 6 The experimental result of FD

3.4 特征权重的计算

针对 GLFS 算法,本文对特征的权重作了相应的修改,即 TFIDF 改为 TFIDF + LOG(CHI + 1) ,改后的权重方法用 TFIDF-CHI 表示。本文在两大文档集上进行了实验,对比实验结果如图 7 和图 8 所示。从图中可以看出采用新的权重值得到的结果比采用 TFIDF 取的结果有明显的提高,且性能提高的速度也明显加快。

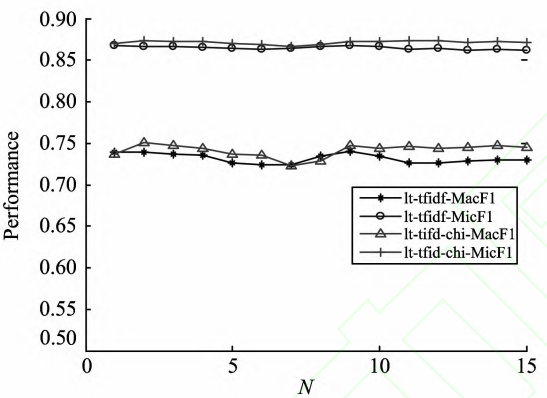

图 7 LT 取两种不同权重值的结果对比

Fig. 7 Comparison of two different weight values result in LT

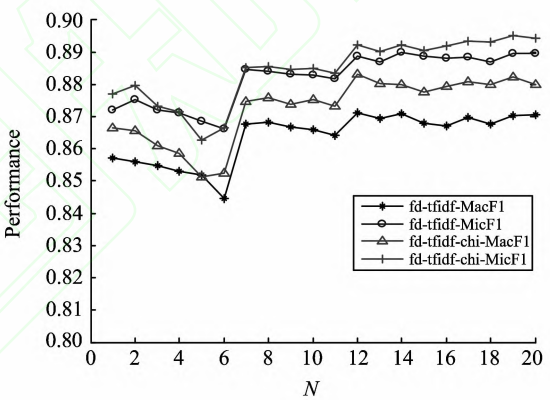

图 8 FD 取两种不同权重值的结果对比

Fig. 8 Comparison of two different weight values result in FD

3.5 实验结果的对比

选取卡方统计量和 GLFS 算法的最好结果与 ALOFT 的实验结果进行对比,结果如图 9 和图 10 所示。图中 lt_tfidf 和 fd_tf_idf 表示文档集中采用 GLFS 算法选取特征后词的权重是 TFIDF 的结果;lt_tfidf_chi 和 fd_tfidf_chi 表示文档集中采用 GLFS 算法选取特征后词的权重是 TFIDF-CHI 的结果;CHI 表示采用卡方法选取特征后分类的最好结果;ALOFT 表示采用 ALOFT 方法选择特征后的分类结果。

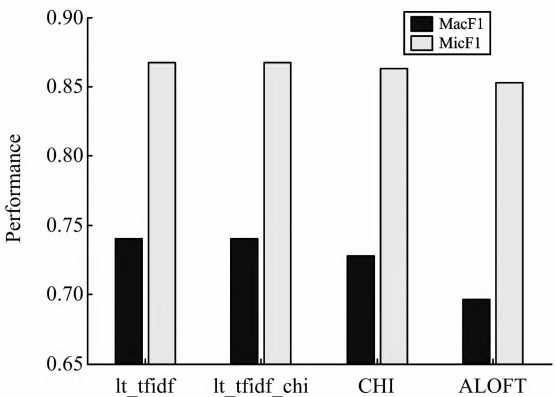

图 9 LT 的 4 种结果对比

Fig. 9 Comparison of the four results of LT

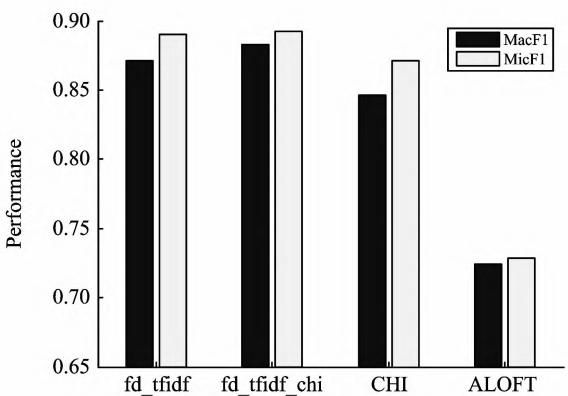

图 10 FD 的 4 种结果对比

Fig. 10 Comparison of the four results of FD

从图中可以看出采用新权重的结果优于采用 TFIDF 权重的结果,而这两种结果都优于 CHI 方法,ALOFT方法的结果最差。在 LT 文档集中采用 ALOFT 方法选择的特征为 80 个,在 FD 文档集上选取的特征是 31 个;很明显采用 ALOFT 方法选取的特征非常少,但其分类效果却明显不理想,其主要原因在于此方法是从全局的角度出发,为了保证每篇文档中至少有一个特征词,而忽略了词对单篇文档的重要性及词对类别的贡献,导致所选的词在一个类中重要,在另一个类中也很重要,从而难以区分。

由此可以看出本文中提出的 GLFS 方法达到的性能是最好的,且无论是在中文文档集还是在英文文档集上都能取得了较好的结果。

4　总结

本文在传统的特征选择方法基础上,考虑到词对单篇文档的贡献率并同时结合 ALOFT 方法,提出了一个 GLFS 算法。在英文路透社文档集和中文的复旦文档集上进行了实验,实验结果表明该算法不仅保证了每个训练文档都不为空,而且取得了较好的分类结果。本文还针对所提出的算法对特征的权重作了相应的改进,且实验结果明显有所提高。最后将该算法与卡方算法和 ALOFT 算法进行了对比实验,结果表明本文提出的算法性能是最好的。在下一步的工作中将考虑重要样本信息,希望能对本方法作进一步的改进。

参考文献:

[1] 谭松波. 高性能文本分类算法研究[D]. 北京:中国科学院计算机研究所,2006.
TAN Songbo. Research on high-performance text categorization[D]. Beijing: Institute of Computing Technology Chinese Academy of Sciences, 2006.

[2] Fabfizio Sebastiani. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1):1-47.

[3] 尚文倩. 文本分类及其相关技术研究[D]. 北京:北京交通大学,2007.
SHANG Wenqian. Research on text categorization and technologies[D]. Beijing: Beijing Jiaotong University, 2007.

[4] 张玉芳,万斌候,熊忠阳. 文本分类中的特征降维方法研究[J]. 计算机应用研究,2012,29(7):2541-2543.
ZHANG Yufang, WAN Binhou, XIONG Zhongyang. Research on feature dimension reduction in text classification[J]. Application Research of Computers, 2012, 29(7):2541-2543.

[5] 郑俊飞. 文本分类特征选择与分类算法的改进[D]. 西安:西安电子科技大学,2012.
ZHENG Junfei. Improvement on feature selection and classification algorithm for text classification[D]. Xi'an: Xidian University, 2012.

[6] SANTANALEA L E A, DE OLIVEIRA D F, CANUTO A M P, et al. A comparative analysis of feature selection methods for ensembles with different combination methods[C]// Proceedings of Internation Joint Conference on Neural Networks. Piscataway: IEEE Press, 2007: 643-648.

[7] 郭颂,马飞. 文本分类中信息增益特征选择算法的改进[J]. 计算机应用与软件, 2013(08):139-142.
GUO Song, MA Fei. Improving the algorithm of information gain feature selection in text classification[J]. Computer Applications and Software, 2013(08):139-142.

[8] 辛竹,周亚建. 文本分类中互信息特征选择方法的研究与算法改进[J]. 计算机应用,2013,33(S2):116-118, 152.
XIN Zhu, ZHOU Yajian. Study and improvement of mutual information for feature selection in text categorization[J]. Journal of Computer Applications, 2013, 33(S2):116-118, 152.

[9] 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.
CHENG Weiqing, TANG Xuan. A text feature selection method using the improved mutual information and information entropy[J]. Journal of Nanjing University of Posts and Telecommunications(Natural Science), 2013, 33(5):63-68.

[10] PINHEIRO R H W, CAVALCANTI G D C, CORREA R F, et al. A global-ranking local feature selection method for text categorization[J]. Original Research Article Expert Systems with Applications, 2012, 39(17):12851-12857.

[11] 胡改蝶. 中文文本分类中特征选择方法的应用与研究[D]. 太原:太原理工大学,2011.
HU Gaidie. Application and research of feature selection method in chinese text categorization[D]. Taiyuan: Taiyuan University of Technology, 2011.

(编辑:许力琴)