

文本分类中特征权重算法的改进

沈志斌, 白清源

(福州大学 数学与计算机科学学院, 福州 350002)

[摘要] TFIDF 是文档特征权重表示常用方法. 该方法简单易行, 但忽略了特征词在各个类别中的分布情况, 不能真正地反映特征词对区分每个类的贡献. 针对这个不足, 本文提出了 BOR-TFIDF, 来重新调整每个特征词对各个类别的区分度, 即修正各个特征词的权重, 并用分类器来验证其有效性. 该方法优于原来的 TFIDF 算法, 实验表明了改进的策略是可行的.

[关键词] 文本分类, 特征权重, TFIDF, 类别区分, BOR-TFIDF

[中图分类号] TP 18 [文献标识码] A [文章编号] 1672-1292(2008) 04-0095-04

Improvement of Feature Weighting Algorithm in Text Classification

Shen Zhibin, Bai Qingyuan

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002, China)

Abstract TFIDF is a kind of common methods used to measure the terms in a document. The method is easy but ignores the distribution of the feature in each class. So it can not really reflect each feature's contribution to each class. Aiming at this shortage, we put forward the BOR-TFIDF and use it to readjust each feature's differentiation to each class, i.e., modifies each feature's weight. Then the classifier is used to check its validity. The method is better than traditional TFIDF and proves that the BOR-TFIDF method is feasible.

Key words text classification, feature weight, TFIDF, class difference, BOR-TFIDF

文本自动分类的任务^[1]是: 对未知类别的文档进行自动处理, 判断它所属预定义类别集中了一个或多个类别. 随着各种电子形式的文本文档以指数级的速度增长, 有效的信息检索、内容管理及信息过滤等应用变得越来越重要和困难. 文本自动分类是一个有效的解决办法, 已成为一项具有实用价值的关键技术. 近年来, 多种统计理论和机器学习方法被用来进行文本的自动分类, 掀起了文本自动分类研究和应用的热潮. 现有的分类方法主要是基于统计理论和机器学习方法^[2], 比较著名的文档分类方法有 Bayes^[3], KNN^[4], Centroid^[5], LLSF^[6], Boosting^[7]及 SVM^[8]等.

本文通过研究发现传统的文本特征权重表示方法 TFIDF 的不足: 忽略了特征词在各个类别中的分布情况. 其实特征词在各个类的分布情况会反映特征词对区分每个类的贡献. 本文对此进行了改进, 并通过实验证明了改进的 TFIDF 方法是高效可行的.

1 传统的 TFIDF 及相关工作

传统的 TFIDF 算法是由 Gerahl Salton 和 McGill 针对向量空间信息检索范例提出的文档特征表示方法. 在此方法中, 出现在文档中的文字称为术语 (Term), 每个术语都有对应的权重, 此权重代表术语在文档识别时的重要程度, 术语的权重与术语在文档中出现的频率成正比, 而与术语在所有文档中出现的频率成反比. TFIDF 实际上是: $TF \times IDF$. 其中 TF (Term Frequency) 称为词频, 指术语在给定文档中的出现次数; IDF (Inverse Document Frequency) 称为倒排频度, 是反映一个术语在一个文档集中按文档统计出现的频繁程度的指标. $IDF = \log(N/n)$. 其中, N 为全部文档数, n 表示包含词条 t 的文档数.

收稿日期: 2008-06-18

基金项目: 教育部留学回国人员启动基金、中科院软件所开放课题基金 (SYSKF0701)、福州大学科技发展基金 (2005-XQ-13) 和福建省教育厅基金 (JB06023) 资助项目.

通讯联系人: 白清源, 教授, 研究方向: 数据库技术和数据挖掘. E-mail: baiqy@fzu.edu.cn

由于传统 TFIDF 算法的不足, 后来提出了各种改进算法. 由于 TFIDF 并没有考虑到特征项在类间和类内的分布情况, 文献^[9]提出了 TFIDF-DI 算法; 文献^[10]考虑所有特征词中占主导地位的核心词并提出了 IMPROVED-TFIDF 算法; 文献^[11]和文献^[11]也针对 TFIDF 的不足提出了相应的改进算法. 本文的工作与前面提到的工作所不同的地方是: (1) 本文用了一种新的易实现的方法克服了 TFIDF 的不足; (2) 改进后的 BOR-TFIDF 提高了分类的准确率, 更重要的是 BOR-TFIDF 对分类效果稳定性的支持.

2 BOR-TFIDF (Based On Ratio TFIDF) 算法

2.1 传统 TFIDF 的不足

传统的 IDF 并没有考虑特征项在文本集合中的分布比例, 只是简单地认为文本频数少的特征项就重要, 文本频数多的特征项就没什么用. 接下来我们就通过图例来分析 IDF 的不足之处.

在图 1 中, “+”号代表正实例, “-”号代表负实例; item 1, item 2, item 3, item 4 是 4 个不同的特征词; “A”表示包含特征词 item *k* 而且是正实例的文档数; “B”表示包含特征词 item *k* 而且是负实例的文档数; “C”表示不包含特征词 item *k* 而且是正实例的文档数; “D”表示不包含特征词 item *k* 而且是负实例的文档数; 文档总数 $N = A + B + C + D$; 一般情况下 $D \gg A, B, C$. 其中, 对于 item 1, item 2, item 3 这 3 个特征项, 它们的 $A + B$ 值都相等; 而对于 item 4, 它的 $A + B$ 和 A 的值都大于前 3 个特征项对应的值, 它的 B 的值等于 item 1 中 B 的值.

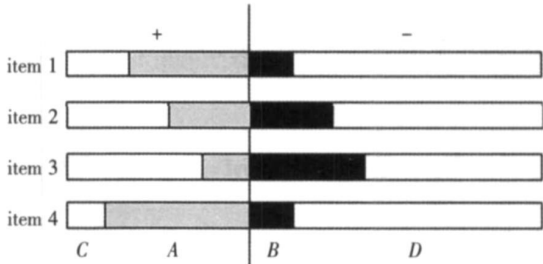


图 1 特征项权重值的比较
Fig.1 The comparison of feature's weight

首先, 我们从 IDF 的角度对这 4 个特征项进行分析, 根据 IDF 的定义, 此时 $IDF1 = IDF2 = IDF3 > IDF4$, 即 item 1, item 2, item 3 对区分某个类的贡献是相等的, 并且大于第 4 个特征项的贡献. 但是从图 1 中我们可以很清楚地看到, 特征项 item 1 的贡献要大于 item 2, item 3, 而 item 4 的贡献要大于 item 1, 因为在包含某个特征项 item *k* 的文档数一样的前提下, 属于正实例的文档数越多的对这个类的贡献就越大. 因此, 在这种情况下, 传统的 TFIDF 计算方法就失去了区分特征项对某个类的贡献的能力. 那如果包含某个特征项 item *k* 的文档数不等的情况下呢, 如 item 1, item 4 有没有什么可行的方法来区分 item 1, item 4 对各个类别的区分度呢?

2.2 BOR-TFIDF 算法

根据前面的分析, 并针对 TFIDF 的缺点, 我们提出了 BOR-TFIDF. 根据 3.1 中的分析, 使用公式: $TFIDF * \lg(1 + A/B)$, 我们就可以算出某个特征项 *T* 对某个类别 *C* 的特征权重. 采用此计算公式后, 对于图 1 中的 4 个特征项, item 1 权重增加, item 2 权重不变, item 3 权重变小, item 4 权重增加且是最大的. 经过这样的调整后, 各个特征项的权重排序如下 item 4 > item 1 > item 2 > item 3 那么前面我们提到的问题就能很好地解决了. 但是这个只是特征项 *T* 针对某个类别 *C* 的权重计算方法. 同样的, 特征项 *T* 对于其它类别也有它们各自的权重值.

那怎么计算出某个特征项 *T* 针对所有类别的权重值呢? 本文是这样处理的. 特征项 *T* 的权重计算公式: $ItemWeight = TFIDF \times \prod_{j=1}^M RATIO_j$, 其中 *M* 表示共有 *M* 个类别, 而 $RATIO_j = \lg(1 + A_j/B_j)$. 通过上述的计算公式我们就可以很简单的计算出特征项 *T* 针对所有类别的权重值. BOR-TFIDF 算法步骤:

- 1) 先按照传统的 TFIDF 方法计算出每个特征词的权重 W_i .
- 2) 计算出每个特征词针对各个类别的 $RATIO_j$, 其中 $A_j = 0$ 或 $B_j = 0$ 的类不用计算

$$MValue = \prod_{j=1}^M RATIO_j$$

- 3) 计算出修正值 $MValue$, 计算公式如下:
- 4) 计算特征词的最终权重值 $Weight = W_i * MValue$.

关于算法的正确性和稳定性, 我们后面的实验有详细的数据支持.

3 实验结果与分析

3.1 实验环境与实验数据集

我们用 VisualC++ 6.0实现本文的算法, 在 C4 2.0 1G, Windows XP的环境下进行实验. 实验数据集①是从中文自然语言处理开放平台网站获取李荣陆收集的新华社的新闻样本^[12]. 其中训练样本 1882个, 测试样本 934个, 共 2816个样本. 样本有 10个类别, 分别为环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治; 采用信息增益和基于文档统计并取 1 000个特征词(即特征维数为 1 000). 我们把实验数据集①记为 Chs2816. 实验数据集②来源于 reuters21578^[13]. 我们用工具^[12]把 reuters21578生成文本文档, 并从中抽取含文档数最多的 10个类, 7 053个训练样本, 2 726个测试样本, 共 9 779个样本. 样本有 10个类别, 分别是 acq, com, crude, eam, grain, interest, money-fx, ship, trade, wheat. 利用信息增益和基于文档统计取 1 000个特征词(即特征维数为 1 000). 我们把实验数据集②记为 reuters9779.

3.2 实验度量标准

使用目前常用的度量标准 Precision(P)和 Recall, F1以及 Macro-avg (宏平均)和 Micro-avg (微平均).

3.3 算法实验结果与分析

本文的实验统一使用 KNN 分类器, 其中 K 取 35, 这个是为了保证分类的公平性. 表 1列出了采用 TFIDF和 BOR-TFIDF对 Chs2816的分类测试结果. 由表 1可知, 采用新的词权重计算方法 BOR-TFIDF后, 除了“计算机”和“体育”这两个类别的准确率略有下降外, 其它类别的分类准确率都有不同程度的提高. 其中, “环境”类别的提升幅度最大, 达到 9个百分点; 其它的类别也至少提高了 1个百分点. 从总体的宏平均 F1和微平均 F1来说, 也有了明显的提高, 分别提高了 2.8和 2.4个百分点. 表 2列出了采用 TFIDF和 BOR-TFIDF对 reuters9779的分类测试结果. 如表 2所示, 除了“ship”和“com”这两个类别的准确率略有下降外, 其它类别的分类准确率也都有不同程度的提高. 其中, “wheat”类别的提升幅度最大, 达到 4.3个百分点. 从总体的宏平均 F1和微平均 F1来说, 也有了明显的提高, 都提高了 1.1个百分点.

表 1 Chs2816开放测试分类结果比较

Table 1 Comparison of open test in Chs2816

F1	TFIDF	BOR-TFIDF
计算机	91.056%	90.769%
交通	91.044%	97.841%
教育	90.277%	92.307%
经济	85.355%	87.337%
军事	72.340%	76.389%
体育	96.104%	95.454%
医药	89.600%	95.384%
艺术	93.827%	94.340%
政治	86.956%	88.023%
环境	80.645%	89.764%
微平均	88.330%	90.685%
宏平均	88.269%	91.066%

表 2 reuters9779开放测试分类结果比较

Table 2 Comparison of open test in reuters9779

F1	TFIDF	BOR-TFIDF
Ship(航运)	60.377%	58.108%
Trade(贸易)	82.456%	84.259%
Wheat(小麦)	12.987%	17.284%
Acq(协定)	93.802%	93.872%
Com(玉米)	17.241%	16.950%
Crude(原料)	79.551%	81.794%
Eam(工资)	95.632%	97.283%
Grain(谷物)	59.130%	60.463%
Interest(利率)	63.551%	65.438%
money-fx(金融)	71.428%	74.586%
微平均	85.400%	86.574%
宏平均	66.639%	67.714%

表 1和表 2对我们改进后的方法 BOR-TFIDF做了很好的支持, 证明了我们改进后的特征词权重计算方法可行的.

图 2和图 3展示了权重计算方法 TFIDF和 BOR-TFIDF在不同的特征维数下对 Chs2816的分类效果. 在这两幅图中, 我们可以很直观的看出 BOR-TFIDF的优势. 优势①在同一特征维数下, 采用 BOR-TFIDF的分类效果比传统的 TFIDF的有了很显著的提高; 优势②随着特征维数的增加, 采用 BOR-TFIDF的分类效果先是增加然后趋于稳定, 而采用传统的 TFIDF的分类效果虽然也有不同程度的增加, 但分类效果的起伏比较大.

图 4和图 5展示了权重计算方法 TFIDF和 BOR-TFIDF在不同的特征维数下对 reuters9779的分类效果. 从图中我们可以清楚地看出 BOR-TFIDF的两个优势——分类效果的准确性和稳定性, 这点与 Chs2816

分类效果中的优势是一致的. 特别是稳定性. 这点在图 5 有明显的体现.

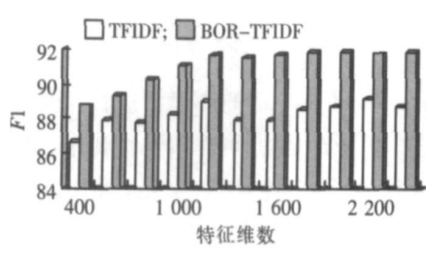


图 2 Chs2816 宏平均 F1
Fig.2 Macro-avg F1 in Chs2 816

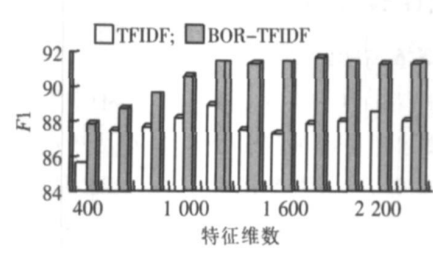


图 3 Chs2816 微平均 F1
Fig.3 Micro-avg F1 in Chs2 816

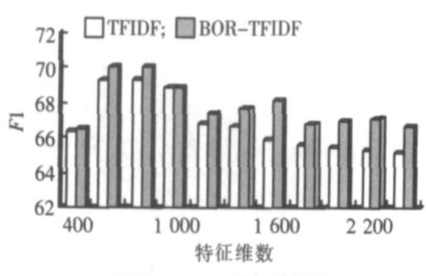


图 4 Reuters9779 宏平均 F1
Fig.4 Macro-avg F1 in Reuters9 779

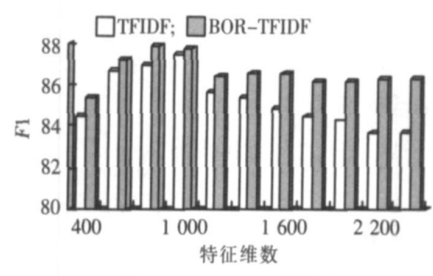


图 5 Reuters9779 微平均 F1
Fig.5 Micro-avg F1 in Reuters9 779

综上所述, 从这些实验的图表我们可以看出, 与 TFIDF 相比, BOR-TFIDF 的优势有以下几点: ①继承了传统的 TFIDF 的优势; ②克服了 TFIDF 忽略特征词在各个类别中的分布情况的缺点, 提高了分类的准确性; ③提高了对类别区分度贡献大的特征词的权重, 降低了对类别区分度贡献小的特征词的权重, 从而削弱了低权重特征词对分类结果的影响, 确保了分类效果的稳定性.

4 结语

特征权重算法的选择对文本分类的精确度有很大影响. 本文研究了传统的 TFIDF 算法, 并在分析其不足的基础上, 提出了新的权重计算方法 BOR-TFIDF. 实验证明, 与原来的 TFIDF 相比, BOR-TFIDF 在分类的精确度上有更好的表现, 并为分类效果的稳定性做了很好的支撑. 因此, BOR-TFIDF 是一种高效可行的特征项权重计算方法.

[参考文献] (References)

[1] 张玉芳, 彭时名, 吕佳. 基于文本分类 TFIDF 方法的改进与应用 [J]. 计算机工程, 2006, 32(19): 76-78
Zhang Yufang Peng Shining Lü Jia Improvement and application of TFIDF method based on text classification [J]. Computer Engineering 2006, 32(19): 76-78 (in Chinese)

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

[3] Lewis D D. Naïve Bayes. The independence assumption in information retrieval [C] // The 10th European Conf on Machine Learning New York: Springer-Verlag 1998.

[4] Ying Yang X in Liu. A re-examination of text categorization methods [C] // SIGIR' 99. New York: ACM Press 1999: 42-49.

[5] Yang Y, Chute C G. An example-based mapping method for text categorization and retrieval [J]. ACM Trans on Information Systems 1994, 12(3): 252-277.

[6] Han E H, Karypis G. Centroid-based document classification analysis and experimental results [C] // Proc of PKDD' 00 London: Springer-Verlag 2000: 424-431.

[7] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions [C] // Proc of the 11th Annual Conf on Computational Learning Theory Madison: ACM Press 1998: 80-91.

[8] Joachims T. Text categorization with support vector machines: learning with many relevant features [C] // The 10th European Conf on Machine Learning Berlin: Springer 1998: 137-142.

(下转第 149 页)

- [2] Linux 与 Windows 的系统安全性比较 [J/OL]. [2005-01-06]. <http://www.hc360.com>
Security comparison between Linux and Windows [J/OL]. [2005-01-06]. <http://www.hc360.com> (in Chinese)
- [3] 唐续, 刘心松, 杨峰. Linux 网络协议栈分析及协议添加的实现 [J]. 计算机科学, 2003 30(2): 130-132
Tang Xu, Liu Xinsong, Yang Feng. Linux network protocol analysis and protocol addition implementation [J]. Computer Science, 2003 30(2): 130-132 (in Chinese)
- [4] 李善平. Linux 内核 2.4 版源代码分析大全 [M]. 北京: 机械工业出版社, 2004
Li Shanping. Source Codes Analysis Guide of Linux 2.4 [M]. Beijing: China Machine Press, 2004 (in Chinese)
- [5] 李长河, 杜辉天, 吕林涛. 一种小型嵌入式 TCP_IP 协议栈的设计与实现 [J]. 微电子学与计算机, 2003(6): 40-43
Li Changhe, Du Huitian, Lv Lintao. Design and implementation of new mini embedded TCP_IP protocols [J]. Microelectronics & Computer, 2003(6): 40-43 (in Chinese)
- [6] 毛新宇. Linux 内核防火墙 netfilter 的原理和应用 [J]. 微型机与应用, 2004(4): 35-37.
Mao Xinyu. Principle and application of Linux kernel firewall——netfilter [J]. Microcomputer & Its Applications, 2004(4): 35-37. (in Chinese)

[责任编辑: 丁蓉]

(上接第 98 页)

- [9] 徐凤亚, 罗振声. 文本自动分类中特征权重算法的改进研究 [J]. 计算机工程与应用, 2005(1): 181-184
Xu Fengya, Luo Zhensheng. An improved approach to term weighting in automated text classification [J]. Computer Engineering and Applications, 2005(1): 181-184 (in Chinese)
- [10] 张云涛, 龚玲, 王永成. 文本分类中 TFIDF 方法的改进 [J]. 浙江大学学报, 2005 6A(1): 49-55
Zhang Yuntao, Gong Ling, Wang Yongcheng. An improved TF-IDF approach for text classification [J]. Journal of Zhejiang University, 2005 6A(1): 49-55 (in Chinese)
- [11] 寇莎莎, 魏振军. 自动文本分类中权值公式的改进 [J]. 计算机工程与设计, 2005, 26(6): 1616-1618
Kou Shasha, Wei Zhenjun. Improved weighting formula in auto text classification [J]. Computer Engineering and Design, 2005, 26(6): 1616-1618 (in Chinese)
- [12] 李荣陆. 文本分类系统 [DB/OL]. http://www.nlp.org.cn/docs/download.php?doc_id=102 2004-08-19
Li Ronglu. Text classification system [DB/OL]. Data Set http://www.nlp.org.cn/docs/download.php?doc_id=102 2004-08-19. (in Chinese)
- [13] David D. Lewis. Reuters-21578 Test Collections [R/OL]. <http://www.daviddleeis.com/resources/testcollections/reuters21578/>. 1996

[责任编辑: 顾晓天]