一种词频与方差相结合的特征加权方法*

邱云飞[1,2], 王 威[1], 刘大有[2], 邵良杉[1]
(1. 辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105; 2. 吉林大学 计算机科学与技术学院, 长春 130012)

摘 要: 通过分析特征词与类别间的相关性,提出了一种新的特征加权方法,依据特征词在特定类中出现的次数、特征词在某一类中的集中程度、特征词在特定类中的均匀分布程度来计算特征权值。通过与 TF-IDF 进行实验对比,新提出的 TF-Var 特征权重方法使得分类的微平均准确率得到了明显的提高。

关键词: 文本分类; 特征权重; 词频; 方差
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)06-2132-03
doi:10.3969/j.issn.1001-3695.2012.06.034

Feature weighting method combined with word frequency and variance

QIU Yun-fei[1,2], WANG Wei[1], LIU Da-you[2], SHAO Liang-shan[1]
(1. *School of Software Engineering, Liaoning Technical University, Huludao Liaoning* 125105, *China*; 2. *School of Computer Science & Technology, Jilin University, Changchun* 130012, *China*)

Abstract: Through analyzing the defects of original feature weighting method, this paper proposed a new weighted method. This method calculated the feature weights based on the frequency of feature words appeared in the particular class, the concentration of feature words in one class, and the well distribution of feature words in the particulor class. Through the contrast of the TF-IDF feature weighting method, the new TF-Var feature weighting method has improved the micro-averaging precision of the classification obviously.

Key words: text classification; feature weight; word frequency; variance

文本分类主要涉及文本分词、文本预处理、特征选取、特征权重计算、分类算法、分类性能测评等多个过程[1,2]。其中,文本分类中特征权重的计算方法是基于向量空间模型(VSM)的一个重要问题,特征词的权重综合反映了该特征词对文本内容的贡献度和区分文本类别的能力大小,特征权重算法的选择将会对整个分类过程产生很大的影响。

近年来,许多学者倾向于研究特征权重问题,焦庆争等人[3]提出了一种可靠信任推荐文本分类特征权重算法;文立等人[4]提出了一种新的动态权重更新相关反馈方法;李艳玲等人[5]提出了一种基于反馈信息的特征权重调整方法。这些方法从不同方面对算法进行了改进,但没有综合考虑特征词的类别信息。为此,本文提出了一种综合考虑特征频率、特征在某一类内的分布情况以及特征在各个类别的分布情况的特征权值计算方法,即一种词频与方差相结合(TF-Var)的特征加权方法。

1 传统的特征加权算法

特征项在不同文本中出现的频率符合一定的统计规律,因此常常通过特征项的频率特性来计算其权重。基于 VSM 的特征权重计算方法主要有布尔权重、词频权重 TF 和 TFIDF(term frequency inverse document frequency weighting)权重[6,7]。

特征权重计算的准则就是要最大限度地区分不同类别文档。因此特征词频率 TF 与反比文档频率 IDF 通常是联合使用的,也就是 TF-IDF 权重。TF 是某个特征词在特定文档中出现的次数,IDF 是特征项在文档集分布情况的量化。

目前最常用的特征加权方法是 TF-IDF 方法。由于 TF-IDF 是将训练文档集作为整体来考虑的,特别是其中 IDF 的计算,IDF 函数结构过于简单,认为文本频率少的单词重要,文本频率高的单词没有用,这显然不符合实际规律,不可能很好地反映特征词的分类作用,而且它并没有考虑特征词在各个类之间和特定类内的分布情况。如果某一特征词在某个特定的类别出现较多而在其他类别出现很少,这样的特征词显然是很有价值的特征词,但这在 TF-IDF 算法中是无法体现的。另一方面,在某一类别出现次数相同的不同特征项,分布相对均匀的特征项的权重应该比分布不均匀的要高,因为如果某一特征项只在某个类别的少量文档中大量出现,而在这个类中的其他文档中出现得很少,这样的特征词显然不能很好地代表这个类别,因此这样的特征项不具备代表性,权重相对较低。同样,传统的 TF-IDF 算法也不能很好地处理。

2 词频与方差相结合的特征加权方法

通过分析传统的特征加权方法的不足,一个有价值的特征词应该同时考虑以下三个因素:

a)词频 TF。根据前面的分析,一个特征在某类文档中经常出现,说明这个特征对该类文档具有代表性,那么它对分类的作用是比较大的。TF 较大的特征项在该类文档中具有较高

收稿日期: 2011-11-15; **修回日期**: 2011-12-26 **基金项目**: 国家自然科学基金资助项目(70971059);辽宁省创新团队项目(2009T045)

作者简介: 邱云飞(1976-),男,副教授,博士,主要研究方向为数据挖掘理论及其应用(qyf321@sohu.com);王威(1987-),女,硕士研究生,主要研究方向为数据挖掘及社会网络分析;刘大有(1942-),男,教授,博导,主要研究方向为知识工程与专家系统、数据挖掘等;邵良杉(1961-),教授,博导,主要研究方向为数据挖掘.

的权重,一个对分类有区分价值的特征词,应该在指定类文档中出现的次数较多。

因此,引入参数 α,表示特征词在某个特定类 c_j 中出现的总次数。

$$\alpha = tf_j(t) \tag{1}$$

训练集中类别 c_j 中有文本 $d_{j1}, d_{j2}, \cdots, d_{jn}$,特征词 t 在文本 d_{jk} 中出现的次数 tf_{jk},特征词 t 在类 c_j 中出现的次数为 $tf_j(t)$。其中,$tf_j(t) = \sum_{k=1}^{n} tf_{jk}$。

可以看出,特征词出现的次数越多,α 值越大。α 值大说明该特征词对该类有代表价值。

b)一个有分类价值的特征词,应该在指定类中分布较多,而在其他类中分布较少。比如,生物类别里面“中国”这种特征词,它在生物类别里面的文档中出现比较少,而在政治、经济、体育等类别的文档中普遍存在,很显然这种特证词对分类的贡献不大,在特征选择时应该被排除。所以,根据方差的思想,样本分布越均匀,方差越小,样本分布越是集中,方差越大。为此,希望找到集中分布于某一类而不是均匀地分布在各个类中的特征词。

引入参数 β,表示某一特定类中包含特征词 t 的文档数与其他类中包含特征词 t 的所有文档数的平均值差值的一个度量。

$$\beta = (df_j(t) - \overline{df_j(t)}) \cdot (df_j(t) - \overline{df_j(t)})^2$$

$$\overline{df_j(t)} = \frac{1}{m}\sum_{j=1}^{m} df_j(t) \tag{2}$$

其中:m 表示类别的个数;类 c_j 中包含特征词 t 的文档数为 $df_j(t)$;$\overline{df_j(t)}$ 表示平均每个类别含有特征词 t 的文档数。β 表示 $df_j(t)$ 与中心 $\overline{df_j(t)}$ 的偏离程度;第一个 $df_j(t) - \overline{df_j(t)}$ 保证特征词出现在特定类中的文本数大于平均值。

β 值用来度量某一类中包含特征词的文档频与所有类文档频的平均值间的偏离程度。β 值大,说明类 c_j 中包含特征词 t 的文档数比所有类中含特征词次数的平均值大,并且大得比较多。

c)一个有分类价值的特征词应该在特定类中均匀地出现,如果某一特征项只在某个类别的少量文档中大量出现,而在这个类中的其他文档中出现得很少,这样的特征词显然不能很好地代表这个类别,因此这样的特征项不具备代表性,权重相对较低。比如,生物类别里面“比赛”和“飞机”这两个特征词,“比赛”这个特征词会在体育类别的多数文档中出现,“飞机”这个特征词只会在体育类别的极少数文档中出现,很显然“比赛”这个特征词对体育类别的标引价值比“飞机”高。而卡方统计方法并没有考虑到这一点。同样,根据方差的思想引入参数 γ 来进行调节,df_{jk} 与 $\overline{df}$ 的差值越大,偏离程度越大,N 越小。

类别 c_j 中有文本 $d_{j1}, d_{j2}, \cdots, d_{jn}$,特征词 t 在文本 d_{jk} 中出现的次数为 tf_{jk},那么

$$\gamma = (tf_{jk} - \overline{tf_j(t)})^{-2} \tag{3}$$

其中:$\overline{tf_{jt}} = \frac{1}{n}\sum_{k=1}^{n} tf_{jk}$。

γ 值用来度量某一文档的词频与这个类中所有文档词频的平均值间的偏离程度。γ 值大,说明文本 d_{jk} 中包含特征词 t 与类别 c_j 中的所有文本中特征词次数的平均值差距比较小,也就是说明这个类别的每一个文本中含有 t 的次数相差不大。

综合 α、β 和 γ,目的是选出集中分布在某一类(β),且在这类均匀分布的特征词(γ),并且在每类的每一篇文档中出现的次数尽可能地多(α)。本文提出了一种新的结合 α、β 和 γ 的特征加权方法——词频与方差相结合的特征加权方法(TF-Var 方法),其公式为

$$W_i(t) = tf_j(t) \cdot (df_j(t) - \overline{df_j(t)})^3 \cdot (tf_{jk} - \overline{tf_j(t)})^2 \tag{4}$$

3　实验结果与分析

实验是在 Pentium® Dual-Core CPU E5300 @ 2.60 GHz CPU,1.99 GB 内存,120 GB & 7200 rpm 硬盘和 Microsoft Windows XP 操作系统下进行的,使用 VC ++ 6.0 开发环境。

为了验证该算法的有效性,选用传统的 TF-IDF 特征加权方法在 ICTCLAS 发布的中文数据集上与提出的 TF-Var 方法进行实验。使用卡方特征选择(CHI)方法并分别选用两种分类算法,即 K 近邻法(KNN)和支持向量机(SVM)方法。选择 KNN 是因为它是通用且性能较好的分类器[8,9],选择 SVM 方法是因为它是最有效的学习算法之一。利用 SVM 和 KNN 两种不同的文本分类器评估上述两种特征加权方法。

3.1　数据集

复旦大学的中文语料库共计 2 816 篇,分为计算机、经济、环境、交通、教育、军事、体育、医药、艺术、政治十类,其中采用 1 882 篇文本作为训练集,934 篇作为文本测试集,如表 1 所示。

表 1　数据集

文档集	计算机	艺术	经济	环境	教育
训练集	134	166	217	134	147
测试集	66	82	108	67	73
文档个数	200	248	325	201	220
文档集	政治	医药	军事	体育	交通
训练集	338	136	166	301	143
测试集	167	68	83	149	71
文档个数	505	204	249	450	214

3.2　分类性能评估

文档分类中普遍使用的性能评估指标有查全率(recall,简记为 r)、查准率(precision,简记为 p)。如果用列联表对单类赋值分类器的性能进行统计计算,则有两种方法可进行,即宏观平均和微观平均。宏观平均是先对每一个类统计 r、p 值,然后对所有的类求 r、p 的平均值。微观平均是先建立一个全局列联表,然后根据这个全局列联表计算。最终分类结果的好坏在一定程度上反映了特征选择的效果,因此可以通过评价分类结果来测试特征选择方法的效果。

根据文献[10],微平均精确率(micro-averaging precision)被广泛用于交叉验证比较,这里用它来比较不同的特征选择算法的效果。

3.3　实验步骤与结果

对数据集中的文本进行预处理,包括去停用词、过滤标点符号、进行词频和文档频率统计等;然后对数据集采用 10 折的交叉验证方法(10 fold cross validation),即将数据集随机划分成 10 个不相交的子集,进行 10 次训练和测试,依次将其中的一个子集作为测试集,其他子集作为训练集,取 10 次训练的平均值作为最终的分类结果。采用全局选取的特征选取方式,选取的特征数目设定为 1 000,特征选择方法为信息增益的特征选择方法与卡方特征选择方法,分类方法分别为 KNN 和 SVM。KNN 算法的 K 近邻值设为 35,对传统的 TF-IDF 方法和新提出的 TF-Var 方法进行评估。

图 1 显示的是 KNN 分类器分别采用传统的 TF-IDF 方法和新提出的 TF-Var 方法在 ICTCLAS 发布的中文数据集上的

Micro_P 曲线。从图中可以看出,TF-Var 方法在特征数目为 1 000 时将 KNN 分类器的 Micro_P 从 85.76% 提高到 95.15%,并且选取 1 500 个特征可以使该分类器的 Micro_P 的最大值达到 87.90%,在选取 1 000 个特征时使该分类器的 Micro_P 达到 95.15%。TF-IDF 方法在选取 1 500 个特征时使该分类器的 Micro_P 最大值达到了 87.90%。

图 2 显示的是 SVM 分类器分别采用传统的 TF-IDF 方法和新提出的 TF-Var 方法在 ICTCLAS 发布的中文数据集上的 Micro_P 曲线。从图中可以看出,TF-Var 方法在特征数目为 1 000 时将 SVM 分类器的 Micro_P 从 92.40% 提高到 95.15%,并且在选取 400 个特征时就可以使该分类器的 Micro_P 值达到 92.15%,在选取 1 000 个特征时使该分类器的 Micro_P 值达到最高值后趋于稳定。TF-IDF 方法在选取 2 000 个特征时使该分类器的 Micro_P 最大值达到 93.15%。

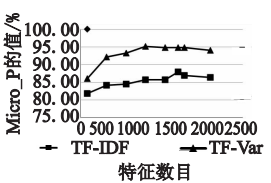

图1 采用两种不同特征加权方法的KNN分类器性能比较

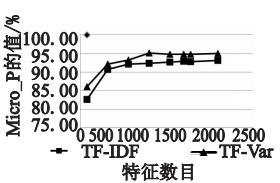

图2 采用两种不同特征加权方法的SVM分类器性能比较

4 结束语

通过分析特征词与类别之间的相关性,提出了一种新的特征权重计算方法——词频与方差相结合的特征加权 TF-Var 方法。此方法运用三个参数来调整特征的权值:第一个参数通过计算特征词在某一类文档中出现的频率来调整其权值;第二个参数度量特定类中某个特征词的文档频与所有类中该特征词的文档频的平均值之间的偏离程度;第三个参数度量在特定类的某一文档中特征词的词频与这个类中所有文档词频的平均值之间的偏离程度。本文分别利用 KNN 和 SVM 分类算法进行对比实验。结果表明,基于词频与方差相结合的特征权重方法使得分类的查全率和查准率都得到了明显的提高。

参考文献:

[1] 熊忠阳,张鹏招,张玉芳. 基于 χ^2 统计的文本分类特征选择方法的研究[J]. 计算机应用, 2008,28(2):513-518.

[2] 肖婷,唐雁. 改进的 χ^2 统计文本特征选择方法[J]. 计算机工程与应用, 2009,45(14):136-140.

[3] 焦庆争, 蔚承建. 一种可靠信任推荐文本分类特征权重算法[J]. 计算机应用研究, 2010,27(2): 472-474.

[4] 文立,石跃祥,莫浩澜. 一种新的动态权重更新相关反馈方法[J]. 计算机应用研究, 2007,24(8): 81-83.

[5] 李艳玲,戴冠中,余梅. 基于反馈信息的特征权重调整方法[J]. 计算机工程, 2009,35(2):206-207.

[6] 初建崇,刘培卫. Web 文档中词语权重计算方法的改进[J]. 计算机工程与应用, 2007,43(19): 192-194.

[7] 罗欣,夏德辨. 基于词频差异的特征选取及改进的 TF-IDF 公式[J]. 计算机应用,2005,25(9): 2031-2033.

[8] 刘赫. 文本分类中若干问题研究[D]. 长春:吉林大学,2009.

[9] ROGAT M, YANG Y M. High-performing feature selection for text classification[C]//Proc of the 11th International Conference on Information and Knowledge Management. 2002:659-661.

[10] YANG Y, PEDERSEN J. A comparative study on feature selection in text categorization[C]//Proc of the 4th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Press,1997:412-420.