集合 CHI 与 IG 的特征选择方法*

王 光, 邱云飞, 史庆伟
(辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105)

摘 要: 通过分析特征词与类别间的相关性,在原有卡方特征选择和信息增益特征选择的基础上提出了两个参数,使得选出的特征词集中分布在某一特定类,并且使特征词在这一类中出现的次数尽可能地多;最后集合 CHI 与 IG 两种算法得到一种集合特征选择方法(CCIF)。通过实验对比传统的卡方特征选择、信息增益和 CCIF 方法,CCIF 方法使得算法的微平均查准率得到了明显的提高。

关键词: 文本分类; 特征选择; 卡方统计; 信息增益
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)07-2454-03
doi:10.3969/j.issn.1001-3695.2012.07.014

Collective CHI and IG feature selection method

WANG Guang, QIU Yun-fei, SHI Qing-wei
(*School of Software Engineering, Liaoning Technical University, Huludao Liaoning* 125105, *China*)

Abstract: In order to make the selected features distribute intensively in a certain class and make features appear in that certain class as many as possible, this paper added the two adjusted parameters to the originally traditional CHI-square feature selection and IG feature selection method through analyzing the relevance between features and classes. Then it proposed a collective feature selection method(CCIF) by combining with CHI and IG feature selection. Experiments show that CCIF improves the Micro-Pmore apparently by comparing with the traditional CHI and IG feature selection method.

Key words: text classification; feature weight; Chi-square statistic; information gain

文本分类主要涉及分词、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]。高维的特征向量是基于向量空间模型(VSM)的一个重要问题。一般大小的数据集就有上万个特征,而这些特征向量大多数都是无用的;通过降维,可以减少分类算法的运行时间和准确率。因此,近些年来,许多工作者倾向于特征选择问题[3,4]。比较常用的特征选择方法主要有文档频数(document frequency,DF)、互信息(mutual information,MI)、期望交叉熵(expected cross entropy,ECE)、文本证据权(weight of evidence for text,WET)、卡方统计量(Chi-square statistic,CHI)、信息增益(information gain,IG)等。

近年来许多学者倾向于研究特征选择问题,在特征选择方面,美国卡内基梅隆大学的 Yang 教授针对文本分类问题,在分析和比较了 IG、DF、MI 和 CHI 等方法后,得出 IG 和 CHI 方法分类效果相对较好的结论[5]。并且,CHI 和 IG 在多次的实验中表现出了良好的准确性。一些国内的学者也倾向于研究特征选择,赵军阳等人[6]提出了一种基于最大互信息最大相关熵的特征选择方法;尚文倩等人[5]提出了一种基于基尼指数的特征选择算法。为此,本文分析了传统的 CHI 和 IG 算法优缺点,通过两个参数并综合两种算法提出了一种新的集合特征选择算法 CCIF。

1 传统的特征选择方法

1.1 卡方统计方法

卡方统计量可以用来度量词条 t 和文档类别 c_i 之间的相关程度;并假设 t 和 c_i 之间符合具有一阶自由度的 CHI 分布。t 对于 c_i 的 CHI 值由下式计算:

$$\chi^2(t,c_i)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

其中:N 表示语料库中文档的总个数;A 表示包含 t 且属于 c_i 类的文档数;B 表示包含 t 但是不属于 c_i 类的文档数;C 表示属于 c_i 类但是不包含 t 的文档数;D 表示既不属于 c_i 类也不包含 t 的文档数。

可以看出 N 固定不变,$A+C$ 为属于类 c_i 的文档数,$B+D$ 为不属于类 c_i 的文档数,所以式(1)可以简化为

$$\chi^2(\hat{t},c_i)=\frac{N(AD-BC)^2}{(A+B)(C+D)} \tag{2}$$

当特征 t 与类别 c_i 相互独立时,$\mathrm{CHI}(t,c_i)=0$。$\mathrm{CHI}(t,c_i)$ 的值越大,特征 t 与类别 c_i 越相关。

对于多个类别问题,首先计算 t 对每个 c_i 的 CHI 值,再分别检验特征 t 对于整个语料的 CHI 值。

$$\chi^2_{\mathrm{avg}}(t)=\sum_{i=1}^{m}P(c_i)\chi^2(t,c_i) \tag{3}$$

$$\chi^2_{\max}(t)=\max_{1\le i\le m}\{\chi^2(t,c_i)\} \tag{4}$$

其中,m 为类别数。式(3)表示特征与各类别的平均 CHI 值,式(4)表示选取特征与各类别 CHI 值中的最大值。根据 CHI 值的大小进行排序选取特定数目的特征词。

1.2 传统的信息增益方法

信息增益被广泛应用在机器学习领域,是信息论中的一个

收稿日期: 2011-12-15; **修回日期**: 2012-01-26 **基金项目**: 国家自然科学基金资助项目(70971059)、辽宁省教育厅基金资助项目(L2010168)

作者简介:王光(1979-),男,讲师,硕士,主要研究方向为数据挖掘(wangguanghero@163.com);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;史庆伟(1973-),男, 副教授,博士,主要研究方向为数据挖掘.

重要概念。对文本分类系统来说,信息增益通过统计某一个特征项 t 在类别 c_i 中是否出现的文档频数来计算特征词 t 对类别 c_i 的信息增益,它考虑了特征词 t 出现前后的信息熵之差,定义为[7]

$$IG(t) = H(C) - H(C|t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\log P(c_i|\bar{t}) \quad (5)$$

其中:$P(c_i)$ 表示 c_i 类文本在语料库中出现的概率,即用 c_i 中的文本数 $A+C$ 除以总的文本数 N;$P(t)$ 表示语料中包含特征项 t 的文本的概率,即用包含特征 t 的文本数 $A+B$ 除以总的文本数 N;$P(c_i|t)$ 表示文本包含特征项 t 时属于 c_i 类的条件概率,即用包含特征 t 且属于类 c_i 的文本数 A 除以包含特征 t 的文本数 $A+B$;同理 $P(\bar{t})$ 是用特征 t 的文本数 $C+D$ 除以总的文本数 N;$P(c_i|\bar{t})$ 是用不包含特征 t 且属于类 c_i 的文本数 C 除以不包含特征 t 的文本数 $C+D$。

$$IG(\hat{t}) = H(C) - H(C|t) \approx -\sum_{i=1}^{m}\frac{A+C}{N}\log\frac{A+C}{N} + \frac{A+B}{N}\sum_{i=1}^{m}\frac{A}{A+B}\log\frac{A}{A+B} + \frac{C+D}{N}\sum_{i=1}^{m}\frac{C}{C+D}\log\frac{C}{C+D} \quad (6)$$

显然,特征 t 的 IG 值越大,表示其越有价值,对分类也就越重要。因此在进行特征选择时,通常选取信息增益值大于特定阈值的若干个单词作为文本的特征向量。

2　集合特征选择方法

2.1　传统的 CHI 与 IG 特征选择方法的不足与改进

传统的 CHI 和 IG 特征选择方法只考虑了特征在所有文档出现的次数,没有考虑特征在某一文档中出现的次数,也就是说它只考虑了特征词的文档频,并没有考虑特征的词频,因此夸大了低频词的作用。这导致了特征预测能力被削弱,无法选出最有效的特征。

一个有分类价值的特征词,应该在指定类文档中出现的次数较多。如果特征词 t_1 在 c_i 类的 100 篇文档中每篇文档出现一次,而另一个特征词 t_2 在 77 篇文档中每篇文档出现 100 次,本应是 t_2 具有较强的分类能力,但根据公式对特征项 t_1、t_2 进行计算时,得到的 CHI 和 IG 值都是 t_1 大,这是一个不合理的结果。由此可见,在没有考虑特征项在各类中的词频分布时,单纯使用文档数的统计方式得到的结果是片面的。

因此,在传统的 CHI 特征选择和 IG 特征选择的基础上引入特征在某类中出现的总词频作为衡量指标,其参数公式为

$$M = tf_i(t) \quad (7)$$

其中,设训练集中类别 c_i 中有文本 d_{i1}、d_{i2}、d_{in},特征词 t 在文本 d_{ik} 中出现的次数为 $tf_{ik}(t)$,特征词 t 在类 c_i 中出现的次数为 $tf_i(t)$,$tf_i(t) = \sum tf_{ik}(t)$。

传统的 IG 特征选择方法的另一个不足在于它考虑了特征项不出现的情况。虽然特征词不在文本中出现也可能会对判断文本的类别有贡献,但这种贡献往往小于它所带来的干扰。特别是在特征分布不均匀时,IG 较大主要是代表特征不出现的值较大,而不是特征出现的值较大。这样会使在一个类别中出现次数不多而在其他类中经常出现的特征项被选出来,而不倾向于选取在一个类别中出现次数较多而在其他类中出现次数较少的更具代表性的特征项,从而导致信息增益的效果大大降低。

对于 CHI 特征选择来说也有此种困扰,由卡方的计算公式看出,B 和 C 都比较大而 A 和 D 都比较小,并且 $BC > AD$,也就是特征词在其他类别出现次数较多而在特定类中出现很少时,计算出来的卡方统计值比较高,因此在特征选择的时候这种特证词很有可能被保留下来了。卡方统计值很可能把对特定类别贡献低而对其他类贡献高的特征词选择出来,这就是常说的负相关现象。

因此,引入参数 N 作为调节参数,此参数是为了解决 IG 特征选择特征不出现和 CHI 特征选择负相关的困扰,其计算式如下:

$$N = df_i(t) - \overline{df_i(t)} \quad (8)$$

其中:$\overline{df_i(t)} = \frac{1}{m}\sum_{i=1}^{m} df_i(t)$;$m$ 表示类别的个数;类 c_i 中包含特征词 t 的文档数为 $df_i(t)$;$\overline{df_i(t)}$ 表示平均每个类别含有特征词 t 的文档数;$df_i(t) - \overline{df_i(t)}$ 用来保证选择在特定类出现次数较多而在其他类中出现次数较少的特征词。

2.2　集合特征选择方法

通过分析 CHI 特征选择和 IG 特征选择的优缺点,综合两个调节参数 M 和 N,得到一个集合 CHI 特征选择和 IG 特征选择的 CCIF 特征选择方法。此方法的目的是在 CHI 和 IG 特征选择的基础上选择出集中出现在某一特定类、并且在该类中每一篇文档中出现的次数多的特征词。其计算式如下:

$$CCIF = IG(\hat{t}) \times \chi^2(\hat{t}, c) \times M \times N \quad (9)$$

3　实验结果及分析

实验是在 Pentium® Dual-Core CPU E5300 @ 2.60 GHz CPU,1.99 GB 内存,120 GB 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 CHI 和 IG 特征选择方法在 ICTCLAS 发布的中文数据集上与提出的 CCIF 特征选择方法进行比较实验。选用 K-近邻法(KNN)和支持向量机(SVM)两种分类算法。选择 SVM 是因为它是最有效的学习算法之一,选择 KNN 是因为它是通用且性能较好的分类器[8]。

3.1　数据集

复旦大学中文语料库共计 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,采用其中 1 882 篇文本作为训练集,934 篇为文本测试集[9],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召回率 r(recall)、准确率 p(precision)。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:宏平均和微平均。宏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],微平均准确率 Micro_P(micro-averaging precision)能较好地反映分类的效果。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,包括过滤标点符号、去停用词、进行词频和文档频率统计等。采用全局选取的特征选取方式,特征加权方法为 TF * IDF,分类方法分别为 KNN 和 SVM 方法,经反复实验,KNN 算法的 K-近邻值设为 35 时效果最好,对传统的 CHI 方法、IG 方法与新提出的 CCIF 方法进行评估。

图 1 显示的是 KNN 分类器分别采用 IG、CHI 和 CCIF 三种特征选择方法在 ICTCLAS 发布的中文数据集上的 Micro_P 曲线。从图 1 可以看出,CCIF 方法在特征数目为 1 000 时将 KNN 分类器的 Micro_P 从 86% 提高到 92%;CHI 方法在选取 1 500 个特征时使该分类器的 Micro_P 达到 87.90%;IG 方法能够使该分类器的 Micro_P 达到最高值 87.69%。CCIF 方法在选取 400 个特征就可以将该分类器的 Micro_P 值达到 89%,提高了分类器的性能。

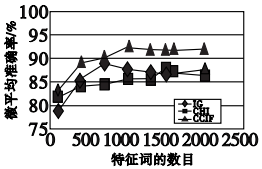

图1 采用三种不同特征选择方法的KNN分类器性能比较

图 2 显示的是 SVM 分类器分别采用 IG、CHI 和 CCIF 三种特征选择方法在 ICTCLAS 发布的中文数据集上的 Micro_P 曲线。从图 2 可以看出,CCIF 方法在特征数目为 1000 时将 SVM 分类器的 Micro_P 从 92% 提高到 94%;CHI 方法在选取 1300 个特征时使该分类器的 Micro_P 达到 93.15%;IG 方法能够使该分类器的 Micro_P 达到最高值 92.82%。CCIF 方法在选取 400 个特征就可将该分类器的 Micro_P 达到 92%,并且在选取 1 000 个特征时使该分类器的 Micro_P 达到最高值后趋于稳定。

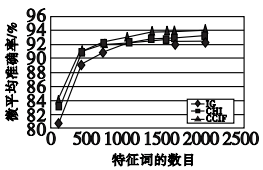

图2 采用三种不同特征选择方法的SVM分类器性能比较

4 结束语

通过分析特征词与类别之间的相关性,使用两个参数 M 和 N 在传统的卡方特征选择和信息增益特征选择的基础上对两种特征选择方法进行改进,用参数 M 表示特征在某类中出现的总词频,它主要是为了解决传统特征选择方法对低频词不可靠的问题。用参数 N 来选择在特定类出现次数较多而在其他类中出现次数较少的特征词,它主要是为了解决 IG 特征选择特征不出现和 CHI 特征选择负相关的困扰。最后,集合卡方特征选择和信息增益特征选择方法,提出了一种集合特征选择方法 CCIF,分别利用 KNN 和 SVM 分类算法进行对比实验数据,结果表明,集合特征选择方法使得算法的微平均准确率得到了明显的提高。

参考文献:

[1] 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.

[2] 苏金树,张博锋. 徐昕. 一种快速文本归类算法的设计与实现[J]. 软件学报,2006,17(9):1848-1859.

[3] 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):136-137.

[4] 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.

[5] 尚文倩,黄厚宽,刘玉玲,等. 文本分类中基于基尼指数的特征选择算法研究[J]. 计算机研究与发展,2006,43(10):1688-1694.

[6] 赵军阳,张志利. 基于最大互信息最大相关熵的特征选择方法[J]. 计算机应用研究, 2009,26(1):232-235.

[7] DONOHO D L. De-noising by soft-thresholding[J]. IEEE Trans on Inform Theory,1995,41(3):613-627.

[8] YANG Yi-ming, PEDERSEN O J. A comparative study on feature selection in text categorization [C]//Proc of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers,1997:412-420.

[9] 李荣陆. 文本分类及其相关技术研究[D]. 复旦大学,2005.

[10] ROGAT M, YANG Yi-ming. High-performing feature selection for text classification[C]//Proc of the 11th International Conference on Information and Knowledge Management. New York: ACM Press,2002: 659-661.