基于方差的 CHI 特征选择方法*

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，在原有的卡方特征选择的方法上增加三个调节参数，使选出的特征词集中分布在某一类，且在某一类中尽可能地均匀分布，并使特征词在某一类中出现的次数尽可能地多。通过实验对比改进前后的卡方特征选择方法，基于方差的卡方统计（Var-CHI）方法使得查全率和查准率都得到了明显的提高。

关键词：文本分类；特征选择；卡方统计量；方差
中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2012)04-1304-03
doi:10.3969/j.issn.1001-3695.2012.04.028

CHI feature selection method based on variance

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

Abstract: In order to make the features selected are distributed intensively in a certain class, evenly in that certain class as much as possible, and make features appear in that certain class as many as possible, this paper added three adjusted parameters to the originally traditional CHI-square feature selection method through analyzing the relevance between features and classes. Var-CHI statistic method based on variance makes the precision and recall improved apparently by comparing the experiments of the traditional CHI-square feature selection method and the improved one.

Key words: text classification; feature selection; CHI-square statistic; variance

0 引言

高维的特征空间是文本分类需要解决的一个问题[1,2]。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，但降低了准确率。因此，近些年来，许多工作者倾向于特征选择问题的研究[3~5]。

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）、卡方统计量（CHI-square statistic，CHI）、信息增益（information gain，IG）、文本证据权（weight of evidence for text，WET）等。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[6]，并且，CHI 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 CHI 特征选择。清华大学李粤等人[7]提出结合传统的互信息方法和 CHI 统计方法，使得查全率和查准率都得到了明显的提高。

通过分析特征词与类别之间的关系发现，一个对某类有分类价值的特征词，应该在该类各个文档中均匀分布且出现的次数较多，并且只在该类中出现的次数较多。对此，在原有的卡方统计方法的基础上增加三个调节参数，提出一种基于方差的特征选择方法，并通过对改进前后的 CHI 特征选择进行实验。实验结果证明了该方法的可行性和有效性。

1 传统的卡方统计方法

卡方统计量可以用来度量词条 t 和文档类别 c 之间的相关程度[8]。假设 t 和 c 之间符合具有一阶自由度的 CHI 分布。t 对于 c 的 CHI 值由式(1)[9]计算：

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

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 类的文档数；B 为包含 t 但是不属于 c 类的文档数；C 表示属于 c 类但是不包含 t 的文档数；D 表示既不属于 c 也不包含 t 的文档频数。可以看出，N 固定不变，$A+C$ 为属于类 c 的文档数，$B+D$ 为不属于类 c 的文档数，所以式(1)可以简化为

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

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

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

收稿日期：2011-09-22；**修回日期**：2011-10-24 **基金项目**：国家自然科学基金资助项目(70971059)；辽宁省创新团队资助项目(2009T045)

作者简介：邱云飞(1976-)，男，辽宁阜新人，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、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.

$$\chi_{\text{avg}}^2(t)=\sum_{i=1}^{m}p(c_i)\chi^2(t,c_i) \tag{3}$$

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

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。然后根据 CHI 值的大小进行排序,选取特定数目的特征词。

2　改进的卡方统计方法

Yang[10] 的研究表明,CHI 统计方法是目前最好的特征选择方法之一。与其他方法相比, CHI 大约减少了 50% 的词汇,分类效果好,在文本数量逐渐增多的过程中,稳定性很好。大多数中文分类系统都采用这种方法,但其依然存在着以下几个方面的缺陷:

a)CHI 统计只考虑了特征在所有文档中出现的次数,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。如果某一特征词在某一类文档的少量文本中出现的次数很多,而在其他文档中没有出现,那么通过传统公式计算得出的 CHI 值可能会很低,但是这种特征词很可能对分类的贡献较大。

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 类的 100 篇文档中的每篇文档出现一次,而另一个特征词 t_2 在 77 篇文档中的每篇文档出现 100 次,那么明显 CHI(t_1,C)比 CHI(t_2,C)大,这是 CHI 的不足之一。所以采用特征在某类内出现的总词频作为调节参数。因此,引入参数 α 来进行调节,其主要是为了解决 CHI 统计方法对文档频率低的特征词不可靠的问题。设训练集中类别 c_j 中有文本 $d_{j1},d_{j2},\cdots,d_{jn}$,特征词 t 在文本 d_{jk} 中出现的次数为 tf_{jk},特征词 t 在类 c_j 中出现的次数为 $tf_j(t)$。其中 $tf_j(t)=\sum_{k=1}^{n}tf_{jk}$,那么

$$\alpha=tf_j(t) \tag{5}$$

可以看出,特征词出现的次数越多,α 值越大。

b)特征词的卡方统计值比较了特征词对一个类别的卡方值和对其他类别的卡方值,最后以由高到低的顺序选出合适的特征词。由卡方的计算公式可看出, 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很少出现时,计算出来的卡方统计值比较高,因此在特征选择时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。

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。比如说生物类别里“中国”这种特征词,它在生物类别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,很显然这种特征词对分类的贡献不大,在特征选择的时候应该被排除。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且数量较多而不是均匀地分布在各个类中。

因此引入参数 β 来进行调节:

$$\beta=(df_j(t)-\overline{df_j(t)})\cdot(df_j(t)-\overline{df_j(t)})^2 \tag{6}$$

其中:$\overline{df_j(t)}=\frac{1}{m}\sum_{j=1}^{m}df_j(t)$,$m$ 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,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,若只集中在该类的个别文本中出现,而在其他文本中很少出现,则该词的价值就要小很多。比如说生物类别里“比赛”和“飞机”这两个特征词,“比赛”会在体育类别的多数文档中出现,而“飞机”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},而 $\overline{tf_j(t)}=\frac{1}{n}\sum_{k=1}^{n}tf_{jk}$,那么

$$\gamma=(tf_{jk}-\overline{tf_j(t)})^{-2} \tag{7}$$

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特征选择方法——基于方差的 CHI(Var-CHI)特征选择方法,公式如下:

$$\chi^2(t,c)=\frac{(AD-BC)^2}{(A+B)(C+D)}\times\alpha\times\beta\times\gamma \tag{8}$$

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 & 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,选用传统的卡方(CHI)和信息增益(IG)两个特征选择方法在 ICTCLAS 发布的中文数据集上与本文提出的 Var-CHI 方法进行比较实验。选用两种分类算法,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[11],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三种特征选择方法。

3.1　数据集

数据来源于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。

表 1　数据集

文档集	计算机	艺术	经济	环境	教育	政治	医药	军事	体育	交通
训练集	134	166	217	134	147	338	136	166	301	143
测试集	66	82	108	67	73	167	68	83	149	71
文档个数	200	248	325	201	220	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 个不相交的自己,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,特征加权方法为 TF × IDF,分类方法分别为 KNN 和 SVM。KNN 算法的 K-近邻值设为 35,对传统 CHI、IG 和新提出的 Var-CHI 方法进行评估。

图 1 显示的是 KNN 分类器分别采用 IG、CHI 和 Var-CHI 三种特征选择方法在 ICTCLAS 发布的中文数据集上的 micro_P 曲线。从图中可以看出,Var-CHI 方法在特征数目为 1 000 时,将 KNN 分类器的 micro_P 值从 86% 提高到 90%;Var-CHI 方法在选取 400 个特征时就可以使该分类器的 micro_P 值达到 89.61%;CHI 方法在选取 1 500 个特征时使该分类器的 micro_P 值达到 87.90%;IG 方法能够使该分类器的 micro_P 值达到最高值 87.69%。

图 2 显示的是 SVM 分类器分别采用 IG、CHI 和 Var-CHI 三种特征选择方法在 ICTCLAS 发布的中文数据集上的 micro_P 曲线。从图中可以看出,Var-CHI 方法在特征数目为 1 000 时将 SVM 分类器的 micro_P 值从 92% 提高到 96%;Var-CHI 方法在选取 400 个特征时就可以使该分类器的 micro_P 值达到 93.15%,并且在选取 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;CHI 方法在选取 1 300 个特征时使该分类器的 micro_P 值达到 93.15%;IG 方法能够使该分类器的 micro_P 值达到最高值 92.82%。

4 结束语

通过分析特征词与类别之间的相关性,在原有的卡方特征选择上增加三个调节参数:第一个参数通过计算特征词在某一文档中出现的频率解决 CHI 对低频词不可靠(夸大低频词的作用)的问题;第二个参数度量某一类特征词的文档频与所有类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。分别利用 KNN 和 SVM 分类算法进行对比实验,结果表明,基于方差的卡方特征选择方法使得查全率和查准率都得到了明显的提高。

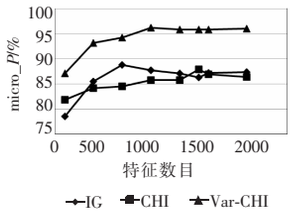

图 1 采用三种不同特征选择方法的 KNN 分类器性能比较

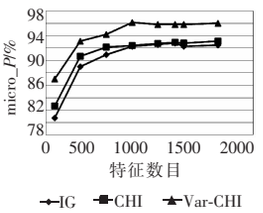

图 2 采用三种不同特征选择方法的 SVM 分类器性能比较

参考文献:

[1] 唐焕玲,孙建涛,陆玉昌. 文本分类中结合评估函数的 TEF-WA 权值调整技术[J]. 计算机研究与发展,2005,42(1):47-53.

[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] 张莉,孙钢,郭军. 基于 K-均值聚类的无监督的特征选择方法[J]. 计算机应用研究, 2005,22(3):23-24,42.

[4] 周宇,覃征. 聚类分析中特征选择的研究[J]. 计算机应用研究, 2006,23(5):55-57,62.

[5] 杨恢先,杨心力,曾金芳,等. 在线特征选择的目标跟踪[J]. 计算机应用研究, 2010,27(3):1180-1182.

[6] YANG Yi-ming, LIU Xin. A re-examination of text categorization methods[C]//Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,1999:42-49.

[7] 李粤,李星,刘辉,等. 一种改进的文本网页分类特征选择方法[J]. 计算机应用,2004,24(7):119-121.

[8] 张俊丽. 文本分类中的关键技术研究[D]. 武汉:华中师范大学, 2008.

[9] 余俊英. 文本分类中特征选择方法的研究[D]. 南昌:江西师范大学,2007.

[10] YANG Yi-ming. An evaluation of statistical approaches to text categorization[J]. Information Retrieval,1999,1(1-2):69-90.

[11] 李荣陆. 文本分类及其相关技术研究[D]. 上海:复旦大学,2005.

(上接第 1303 页)

[19] SHIN K S,LEE T S,KIM H J. An application of support vector machines in bankruptcy prediction model[J]. Expert Systems with Applications,2005,28(1):127-135.

[20] KOHONEN T. Self-organizing maps[M]. New York:Springer,1995.

[21] EKLUND T,BACK B,VANHARANTA H,*et al*. Assessing the feasibility of self-organizing maps for data mining financial information[C]// Proc of the 10th European Conference on Information Systems. 2002:528-540.

[22] SILVA B L, MARQUES N. Feature clustering with self-organizing maps and an application to financial time-series portfolio selection[C]//Proc of International Conference on Neural Computation. 2010:301-309.

[23] CARLOS S C. Self organizing neural networks for financial diagnosis[J]. Decision Support Systems,1996,17(3):227-238.

[24] 荆新,王化成,刘俊彦. 财务管理学[M]. 4 版. 北京:中国人民大学出版社,2006.

[25] MIZUNO H,KOSAKA M,YAJIMA H,*et al*. Application of neural network to technical analysis of stock market prediction[J]. Studies in Informatic and Control,1998,7(2):111-120.

[26] 孙宝珩,藏洪阁,张以宽. 现代商业企业财务分析与评价[M]. 北京:中国国际广播出版社,1996.