Computer Science and Application 计算机科学与应用, 2015, 5(9), 322-330
Published Online September 2015 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2015.59040

Hans汉斯

An Improved CHI Text Feature Selection Method Based on the Location and Word Frequency Information

Aling Song, Haifeng Liu, Shousheng Liu

Institute of Sciences, PLA University of Science and Technology, Nanjing Jiangsu
Email: hfliu1962@sina.com

Received: Oct. 2nd, 2015; accepted: Oct. 16th, 2015; published: Oct. 21st, 2015

Abstract

Text feature selection is the core technology of text automatic categorization. Aiming at the shortcomings of classical CHI model, we have screened the feature set which is based on the point of view of the positive and negative correlation between the feature and categories firstly. According to the type of deflection classification conditions, we adjust the feature weighting secondly. Thirdly, basing on characteristics of word frequency, we gradually improve the model based on the characteristics of a specific location in the text and the characteristics of distribution of information between classes. Finally, we propose an optimized CHI feature selection method. Text classification experiments demonstrate the effectiveness of the optimized CHI model.

Keywords

Feature Selection, Chi-Square, Relevance, Location Distribution, Class Deflection

基于位置及词频信息的优化CHI文本特征选择方法

宋阿羚，刘海峰，刘守生

解放军理工大学理学院，江苏 南京
Email: hfliu1962@sina.com

文章引用: 宋阿羚, 刘海峰, 刘守生. 基于位置及词频信息的优化 CHI 文本特征选择方法[J]. 计算机科学与应用, 2015, 5(9): 322-330. http://dx.doi.org/10.12677/csa.2015.59040

收稿日期：2015年10月2日；录用日期：2015年10月16日；发布日期：2015年10月21日

摘　要

特征选择是文本自动分类的核心技术。针对经典的CHI模型不足之处，本文首先从特征项与类别之间的正负相关性角度对特征项进行删减；然后针对类偏斜分类环境下的特征项权重进行调整；进而以特征项的词频数为依据，从特征项在文本中的具体位置、特征项的类内及类间分布等层面再对模型逐步改进，提出了一种优化的CHI特征选择方法。随后的文本分类试验验证了该方法的有效性。

关键词

特征选择，χ^2 统计，相关性，位置分布，类偏斜

1. 引言

文本信息处理技术是信息检索研究特别是基于 Web 的信息挖掘研究领域的重要研究方向，而文本自动分类技术是文本信息处理的重要研究内容。文本自动分类(Text Categorization，TC)是指根据待分类文本的具体内容将其自动划分到相应的一个或几个文本类别里。通常使用向量模型用于文本表示，使用文本中的单词或词组作为文本特征项。高维的文本向量不仅使得计算开销加大，而且相应的数据稀疏性问题严重影响着分类算法性能，降低了分类的效率。因此有效的文本特征降维方法成为提高文本分类效率的重要途径。

在文本特征降维方面，特征抽取及特征选择是两种主要途径。相比较于特征抽取模式来说，特征选择模式以其目标函数构造相对简单、具有语义解释、易于理解以及应用方便等优点而得到更加广泛的应用。目前常用的文本特征选择方法有互信息(MI)、信息增益(IG)、特征－倒文本频率(TF-IDF)、χ^2 (CHI)统计以及相关系数(CC)等经典方法。研究表明[1] CHI 和 IG 方法的分类效果相对较好。但是两者相比较来说 IG 的计算量较大，因此 CHI 方法成为最常用的文本特征选择方法之一。

对 CHI 方法的设计原理进行理论研究，探讨该模型的特点、存在的主要问题以及寻找有效的改进途径具有重要的现实需求和应用前景。目前针对 CHI 方法的改进研究得到了越来越多的关注。本文在对相关研究进行梳理的前提下，针对目前的研究途径在文本与特征项之间的正负相关性、特征项的类内基于文本的频数分布、类内词频数目分布、类间词频数目分布等方面信息挖掘的不足之处，对经典 CHI 模型进行优化，提出了一种改进的 CHI 方法。随后的文本分类实验证明了本文提出的模型的有效性。

2. 相关问题的研究进展及问题

χ^2 统计(CHI)方法以其易于理解、计算量小以及算法复杂度低等优点成为文本特征选择最常用的方法之一。一些学者针对 CHI 方法的不足进行了改进研究[2]-[6]。文[2]通过引入类内频数以及类间频数差值两个调节因子，将卡方统计和信息增益两种特征选择方法相结合优化 CHI 性能，改善了 CHI 方法对低频词效率较差问题，但是对于 CHI 模型的不足之处没有进行优化[2]；文[3]提出一种组合型特征选择方法[3]，借助信息增益及文本频率两种方法的长处改善 CHI 缺少特征类内分布信息的欠缺，但是特征项的类间分布信息没有利用；文[4]通过分析特征项的类内词频数、类间词频数的差异性信息引入三个参数调节模型中特征项的 CHI 值，提出一种基于方差的 CHI 优化模型，使得新的特加权模型的特征项体现了特征

项的词频分布信息[4]。但是三个参数的设计没有反映出不同类别之间特征项词频信息分布上的差异性，而经典的 CHI 模型设计涉及的特征项频数信息包括类内和类间两个层面；文[5]从特征的分散度、频度以及集中度等三个角度对模型进行改进，改善了 CHI 方法倾向于选择那些在特定类内频数较小、而在其他类内普遍存在的特征项的不足[5]。但是模型是以三个参数的算术平均值为权重对 CHI 进行加权，在类别分布相对均匀时性能较好，而在类偏斜条件下小类别特征项被淹没问题没有涉及，使得分类性能在出现类别分布不均时性能降低。文[6]从特征与类别相关性角度出发对模型优化，通过引入正相关因子以及强相关因子对模型加权，降低了负相关特征项对 CHI 算法的干扰程度[6]，但是文本的类别之间词频分布信息没有利用。

本文着眼于特征项与文本的正负相关性、特征项的类内基于文本的频数以及词频数、类间词频数等方面的信息对特征项的类别表示能力进行分析，逐步修正 CHI 在理论设计方面的不足之处，提出一种基于词频信息的优化 CHI 模型，随后的文本分类试验验证了该模型的有效性。

3. 基于特征项位置及类内、类间分布信息的 CHI 优化

3.1. CHI 方法特点以及存在问题

χ^2 统计方法描述如下：记 c_j 为第 j 类文本集，n_{c_j} 为类 c_j 内文本数，$j=1,2,\cdots,r$；$S=\bigcup_{1\le j\le r} c_j$ 为文本集，$n=\sum_{j=1}^{r} n_{c_j}$ 为文本总数，$T=\{t_1,t_2,\cdots,t_m\}$ 为特征项集合。

在 χ^2 统计(CHI)方法中，特征项 t_k 和类别 c_j 之间的相关性度量如(1)式所示：

$$\chi^2(t_k,c_j)=\frac{n(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

其中，A 表示 S 里包含 t_k 并且属于 c_j 的文本数，B 表示 S 里包含 t_k 但是不属于 c_j 的文本数；C 表示 c_j 中不含 t_k 的文本数；D 表示不在 c_j 中且不含 t_k 的文本数。

数值 $\chi^2(t_k,c_j)$ 是量化特征项 t_k 和类别 c_j 之间的相互关系程度的指标，反映的是特征项在文本集里的分布状况信息，体现的是文本与特征之间的相互依存程度。

该模型的意义是 $\chi^2(t_k,c_j)$ 值越大，表示特征项 t_k 和类别 c_j 之间的相关性越高。当 t_k 与 c_j 相互独立时有 $\chi^2(t_k,c_j)=0$，这意味着 t_k 不含有 c_j 的任何类别信息。而特征项 t_k 在整个文本集里的度量值如下式所示：

$$\chi^2(t_k)=\max\left\{\chi^2(t_k,c_j), j=1,2,\cdots,r\right\} \tag{2}$$

该方法的主要优点是计算量较小并且算法复杂度较低，从语义角度看模型的设计易于理解、使用方便，在类别分布相对均匀的情况下分类效果较好。

对(1)式的结构进行理论分析可以发现模型的不足之处：一是特征项与类别之间的正负相关性没有区分；二是模型的设计没有考虑类别之间文本数目大小对特征项权值的影响；三是该方法是基于含有 t_k 的文本数为计量单位，没有考虑 t_k 在类内的词频因素对模型的影响；四是没有考虑 t_k 在文本的不同位置上对其类别表现能力的影响，对这些方面模型设计的不足之处逐步优化是可以提高模型的特征选择性能。

3.2. 基于特征项、类别之间正负相关性的模型优化

注意到(1)式中 A 是包含 t_k 的属于 c_j 的文本数，C 是类 c_j 中不含 t_k 的文本数，因此 $A+C=n_{c_j}$ 为类 c_j 中文本数；此时由 B 与 D 的定义易知有 $B+D=n-n_{c_j}$ 表示文本集 S 里除去类 c_j 的剩余文本数，对于确定

的文本集 S 来说这两个值为常数。因而(1)式可以表示为：

$$\chi^2\left(t_k,c_j\right)=\frac{n}{n_{c_j}\left(n-n_{c_j}\right)}\frac{\left(AD-CB\right)^2}{\left(A+B\right)\left(C+D\right)}=\frac{n}{n_{c_j}\left(n-n_{c_j}\right)}M\left(t_k,c_j\right) \tag{3}$$

对上式中因子 $M\left(t_k,c_j\right)=\frac{\left(AD-CB\right)^2}{\left(A+B\right)\left(C+D\right)}$ 的结构进行分析是模型优化的可能途径。

类别属性表现能力强的特征项的共同特点是仅在一个类别或少数几个类别内的文本里频繁出现、同时在其他类别内却很少出现[7]。这个思想反映在(3)式中，是指对于类别属性较强的特征项 t_k 及其所属类别 c_j 之间的关系为 c_j 内含有 t_k 的文本数应该很多，即 A 值应该较大；相应的根据 $C=n_{c_j}-A$ 可以推得 C 值应该较小，所以 A/C 值就变得较大。另一方面，由于其它类内含有 t_k 的文本数应该较少，从而 B 值较小，而根据关系 $B+D=n-n_{c_j}$ 可知 $D=n-n_{c_j}-B$ 应该较大，从而 B/D 的值应该较小。这样应有不等式 $A/C>B/D$，由此推得满足：

$$AD-CB>0 \tag{4}$$

也就是说符合(4)式的特征项 t_k 的类别表现能力强。

反之，那些在所属类内文本中出现的次数很少、同时在其他类内的文本里出现的次数较多的特征项，A/C 值就变得较小，而 B/D 的值应该较大，从而这样的特征项相应的基于文本的特征频数信息应满足 $A/C<B/D$，由此得到

$$AD-CB<0 \tag{5}$$

即满足(5)式的特征项的类别表现能力很差，应该避免选择这样的特征项。

但是分析 CHI 模型(3)式可以发现：其因子 $M\left(t_k,c_j\right)=\frac{\left(AD-CB\right)^2}{\left(A+B\right)\left(C+D\right)}$ 是以 $\left(AD-CB\right)^2$ 形式进行计算，这就掩盖了(4)式与(5)式的正负性问题。

事实上符合(4)式的特征项与类别之间是呈正相关性，而符合(5)式的特征项与类别之间呈负相关性，特征选择只能选择符合(4)式的特征项，而呈负相关性的特征经常是类内低频词，应该被剔除。为此考虑对(3)式进行优化。

记：

$$M_1\left(t_k,c_j\right)=\begin{cases}\frac{\left(AD-CB\right)^2}{\left(A+B\right)\left(C+D\right)}, & AD-CB>0\\ 0, & AD-CB<0\end{cases} \tag{6}$$

将(3)式改进为：

$$\chi_1^2\left(t_k,c_j\right)=\frac{n}{n_{c_j}\left(n-n_{c_j}\right)}M_1\left(t_k,c_j\right) \tag{7}$$

(7)式将与类别之间存在负相关的特征项删除，排除低频词对特征选择的干扰。

3.3. 基于特征分布信息的模型优化

3.3.1. 类偏斜条件下基于类内文本频数分布的模型优化

由于 CHI 模型是基于类别的特征项分类性能评估，所以类别之间的文本数目差异经常导致小类别文本特征无法被选择到，这是因为小类别内类别表现能力较强的特征即使其类内的频数较大，但是由于类

内文本数量偏低使得其 CHI 值很难超过阈值而被选出。这种小类别样本被淹没现象会影响 CHI 模型的特征选择性能。但是类偏斜现象却是文本分类中的常见现象，特别是在基于 web 的文本自动分类过程中。因此对类偏斜条件下模型的优化有助于提高模型的适应性。

记 $n_{cj}(t_k)$ 表示在类 c_j 内含有特征项 t_k 的文本数，$j=1,2,\cdots,r$。在类 c_j 内由于含有特征项的文本频数随着类内文本数目的大小而波动，因此特征项的频数大小不能真正反映其类别表现能力，而频率的大小从一个方面体现了 t_k 的类别属性。一般说来类别属性较强的特征项 t_k 在其类内相应的文本频率应较大，而类别属性较差的特征项在类内出现的文本频率应该较小，这一指标基本不受类别分布倾斜程度的影响。故考虑引入权重因子(8)修正 t_k 与类 c_j 之间的相关程度：

$$\lambda_{c_j}(t_k)=\log_2\left[1+\frac{\max\limits_{1\le p\le m}\left\{n_{cj}(t_p)\right\}-\min\limits_{1\le p\le m}\left\{n_{cj}(t_p)\right\}}{n_{cj}(t_k)-\min\limits_{1\le p\le m}\left\{n_{cj}(t_p)\right\}+1}\right]$$

上式体现出 t_k 随着类 c_j 内文本数量增加其权重 $\lambda_{c_j}(t_k)$ 降低的趋势，从而大类别样本的特征项占优现象将得到相应抑制。

考虑到利用 t_k 的类间分布信息，故对上式作归一化处理：

$$\tilde{\lambda}_{c_j}(t_k)=\frac{\lambda_{c_j}(t_k)}{\sqrt{\sum_{j=1}^{r}\lambda_{cj}^2(t_k)}},\tag{8}$$

进而对(7)式优化得：

$$\chi_2^2(t_k,c_j)=\frac{n}{n_{c_j}\left(n-n_{c_j}\right)}\tilde{\lambda}_{cj}(t_k)M_1(t_k,c_j)\tag{9}$$

3.3.2. 基于特征项的类内频数分布信息模型优化

特征项的基于文本的类内出现频率是量化其类别表示能力的一个因素，而其在类内不同文本之间出现的词频数目的大小是反映其类别表现能力的另一个因素[8]。注意到(3)式中的 A、B、C、D 表示的是与特征项 t_k 对应的文本频数相关信息，体现的是否含有以及不含有特征项 t_k 的文本在各个类别 $c_j, j=1,2,\cdots,r$ 内出现或者不出现的频数，是以文本数为基本计算单位，并没有反映特征项 t_k 的词频数，因而词频信息没有被充分利用。这种度量方式的不足之处是明显的：

假定特征项 t_k、t_p 在 c_j 内的大多数文本 $d_{ji}, i=1,2,\cdots,n_{c_j}$ 中出现同时在其它类别内很少出现，则这两个特征项均可能是 c_j 的类别特征项。由(3)式可以判定特征项 t_k、t_p 与类别 c_j 之间的 χ^2 值相差不大；但是假如 t_k 在 c_j 内的文本里出现的频数远远大于 t_p 在 c_j 内文本里出现的频数，即假设 $tf_{ji}(t_k)$ 表示特征项 t_k 在 c_j 类内的文本 d_i，$i=1,2,\cdots,n_{c_j}$ 内出现的频数，且满足 $\sum_{i=1}^{n_{c_j}}tf_{ji}(t_k)\gg\sum_{i=1}^{n_{c_j}}tf_{ji}(t_p)$,此时 t_k 对 c_j 内文本的表现能力远远超过 t_p，反映出 t_k 的特征项特点更加明显。但是(3)式却无法体现出这种差异性。

显然在一个或少数几个类别内出现的频率越大的特征项对该类别具有更强的类属表现能力。为此记：

$$\mu_{cj}(t_k)=\sum_{i=1}^{n_{cj}}\frac{tf_{ji}(t_k)}{\max\limits_{1\le q\le m}\left(tf_{ji}(t_q)\right)},\quad j=1,2,\cdots,r$$

并将上式进行归一化。为此引入权重因子：

$$\tilde{\mu}_{cj}\left(t_k\right)=\frac{\mu_{cj}\left(t_k\right)}{\sqrt{\sum_{i=1}^{r}\mu_{ci}\left(t_k\right)}} \tag{10}$$

将(9)式进一步改进为：

$$\chi_3^2\left(t_k,c_j\right)=\frac{n}{n_{c_j}\left(n-n_{c_j}\right)}\tilde{\lambda}_{cj}\left(t_k\right)\times\tilde{\mu}_{cj}\left(t_k\right)\times M_1\left(t_k,c_j\right) \tag{11}$$

3.3.3. 基于特征项的类间频数分布信息模型再优化

特征项的类别表现能力也体现在类间的词频差异上。类别属性强的特征项 t_k 应该集中出现在少数类别文本里、同时在其它类别内较少或者不出现。这一特点在词频数据分布方面表现为特征项在不同类别的词频数之间的离散性较大，从统计学角度考虑体现为样本之间的方差较大时特征项的类别属性较强，并且样本方差是总体方差的无偏估计，因此使用样本方差对 t_k 在不同类别的词频数分布信息进行量化是合理的。记 $tf_{cj}\left(t_k\right)=\sum_{i=1}^{n_{c_j}}tf_{ji}\left(t_k\right)$ 表示特征项 t_k 在类 c_j 内出现的频数， $j=1,2,\cdots,r$ ，则 t_k 在各个类别之间出现频数的样本均方差为

$$\theta\left(t_k\right)=\sqrt{\frac{1}{r-1}\sum_{j=1}^{r}\left[tf_{cj}\left(t_k\right)-\overline{tf}_{cj}\left(t_k\right)\right]^2} \tag{12}$$

其中 $\overline{tf}_{cj}\left(t_k\right)=\frac{1}{r}\sum_{j=1}^{r}tf_{cj}\left(t_k\right)$ 为样本均值。

将(12)式作归一化处理，记

$$\tilde{\theta}\left(t_k\right)=\frac{\theta\left(t_k\right)}{\sqrt{\sum_{k=1}^{m}\theta^2\left(t_k\right)}}, \tag{13}$$

(11)式则相应的改进为：

$$\chi_4^2\left(t_k,c_j\right)=\frac{n}{n_{c_j}\left(n-n_{c_j}\right)}\tilde{\lambda}_{cj}\left(t_k\right)\times\tilde{\mu}_{cj}\left(t_k\right)\times\tilde{\theta}\left(t_k\right)\times M_1\left(t_k,c_j\right) \tag{14}$$

3.3.4. 基于特征在文本内位置的 CHI 模型再优化

特征项在文本里出现的位置不同，其类别标引能力的差异会很大。一般说来，文章标题是该文主旨内容的概括与总结，文章摘要精炼了文章的核心内容，关键词及章节标题则反映了文章的主体布局，这些位置上出现的特征项的类别属性要远远大于正文里出现的特征项。

研究表明[9]：特征项对文本表达能力从大到小顺序为：标题 > 摘要 > 关键词 > 副标题 > 第一段首句 > 第一段尾句 > 尾段 > 其它，实验表明[10]：文章标题、摘要、首段；章节标题、第一段首句、第一段尾句、文本其他部分对主题的表现能力权重为 5:5:5:4:4:4:2。因此将训练集作相应处理以体现不同位置的特征项在文本类别表示方面的差异。方法如下：

1) 将训练文本集分解为三部分，第一部分由文章标题、关键词、摘要及首段组成；第二部分包含章节标题、第一段首句、第一段尾句；其他内容属于第三部分，分别归入 3 个集合 $S_i\left(i=1,2,3\right)$ 里，形成 3 个“伪训练集”， $S_i\left(i=1,2,3\right)$ 里文本类别属性与原始训练集一致；

2) 对于特征项 t_k ，使用(14)式在 $S_i\left(i=1,2,3\right)$ 内计算其与类别 c_j 之间的 CHI 值，分别记为

$\chi^2_{4i}(t_k,c_j), i=1,2,3$；

3) 特征项 t_k 与类别 c_j 之间对应的最终 χ^2 值为：

$$\chi^2_{new}(t_k,c_j)=2.5\chi^2_{41}(t_k,c_j)+2\chi^2_{42}(t_k,c_j)+\chi^2_{43}(t_k,c_j) \tag{15}$$

特征项 t_k 在整个文本集里的 χ^2 值如下式所示：

$$\chi^2_{new}(t_k)=\max\left\{\chi^2_{new}(t_k,c_j), j=1,2,\cdots,r\right\} \tag{16}$$

4. 一种基于位置及词频信息的优化 CHI 文本特征选择

将本文提出的一种基于位置及词频分布信息的优化 CHI 特征选择方法描述如下：

1) 对训练集根据 3.3.4 节方法分成三个“伪训练集” $S_i(i=1,2,3)$；

2) 在 $S_i(i=1,2,3)$ 内分别进行分词、去除特高频词及特低频词及过滤掉禁用词等预处理，将剩余特征项合并得到文本特征集 $T=\{t_1,t_2,\cdots,t_m\}$；

3) 对于特征集 T 中的特征项 t_k，在三个分别“伪训练集” $S_i(i=1,2,3)$ 内分别基于(8)式、(10)式及(13)式计算权重因子 $\tilde{\lambda}_{cj}(t_k)$、$\tilde{\mu}_{cj}(t_k)$、$\tilde{\theta}(t_k)$，$k=1,2,\cdots,m$；

4) 使用(14)式在 $S_i(i=1,2,3)$ 内分别计算特征项 t_k 与类别 c_j 之间的 CHI 值 $\chi^2_{4i}(t_k,c_j), i=1,2,3$；

5) 根据(15)式计算 t_k 与 c_j 之间的 χ^2 值 $\chi^2_{new}(t_k,c_j)$；

6) 根据(16)式计算 t_k 对应的 χ^2 值 $\chi^2_{new}(t_k)$，$k=1,2,\cdots,m$。按照 $\chi^2_{new}(t_k)$ 的大小顺序选择前 $q \ll m$ 个特征项构成文本特征集 $T_{new}=\{t_k,k=1,2,\cdots,q\}$ 用于文本表示；

7) 对于选定的特征集 T_{new}，以文本分类中常用的 $tf-idf$ 公式[11]对特征项赋权，构造文本向量；

8) 以类内文本向量算术平均值构造类别向量，使用向量夹角余弦公式计算测试文本与类别文本相似度；

9) 将测试文本归入与其相似度最大的类别文本向量所属的类别内。

本文实验中相应参数取值 $q=150, r=6$，这是由于一篇 2000 字左右的文本用约 150 个左右的特征项即可较为有效的表示。

5. 实验结果及分析

试验数据使用复旦大学提供的中文文本分类语料库[12]，从中选取 6 类文本共计 2530 篇，包括计算机(230 篇)、艺术(740 篇)、经济(180 篇)、环境(620 篇)、教育(310 篇)以及军事(450 篇)进行模型实验。使用东北大学的自然语言处理实验室分词软件分词，剔除停用词、虚词、代词等预处理后得到特征集共含 3726 个特征项。试验采用 4 分交叉试验法循环测试 4 次取平均值为最终结果。使用文本分类中常用的查全率(recall)、查准率(precision)及 F_1 测试值作为效率评估指标。其中：查全率 = 分类正确文本数/测试文本总数；查准率 = 分类正确文本数/分类器识别为该类文本数；F_1 值 = 2 × 查准率 × 查全率/(查准率 + 查全率)。

为了考查本文提出的改进模型分类效果，将其与常用的 χ^2 统计方法(即公式(3))以及文[6]提出的方法的分类效率进行了比较，将上述两个方法的实验数据分别以后缀 1、2 标注，分别记为 CHI-1 及 CHI-1，将本文提出的优化模型数据以后缀 new 标注。实验结果相关数据统计如表 1 所示。

将本文提出的优化模型与前两种方法在三个指标方面的增长程度分别进行了比较，相应的增长率以后缀 1-new 及 2-new 表示，数据统计结果如表 2 所示。

分析表 1 发现本文的优化 CHI 模型的特征选择性能较高。综合指标 F1 值介于 0.8947 至 0.9635 之间，

Table 1. The statistics results of the three indexes of three methods in text classification experiment
表 1. 三种方法在文本分类实验的三个指标结果统计

类别	查准率			查全率			F1 值		
	CHI-1	CHI-2	CHInew	CHI-1	CHI-2	CHInew	CHI-1	CHI-2	CHInew
计算机	0.8865	0.9127	0.9542	0.9137	0.9316	0.9721	0.8999	0.9221	0.9635
艺术	0.9121	0.9146	0.9219	0.9011	0.9224	0.9418	0.9066	0.9185	0.9317
经济	0.8227	0.8761	0.8879	0.8347	0.8764	0.9016	0.8287	0.8762	0.8947
环境	0.8625	0.8867	0.9138	0.8713	0.9003	0.9172	0.8669	0.8934	0.9155
教育	0.9218	0.9121	0.9348	0.8792	0.9027	0.9416	0.8999	0.9074	0.9382
军事	0.9135	0.9128	0.9427	0.9043	0.9019	0.9220	0.9089	0.9073	0.9322
平均值	0.8896	0.9036	0.9307	0.8914	0.9110	0.9375	0.8904	0.9072	0.9346

Table 2. The data statistics of new method's growth in the classification experiment
表 2. 新方法在分类实验指标增长率方面的数据统计

	计算机	艺术	经济	环境	教育	军事	平均值
查准率 1-new	7.64%	1.07%	7.93%	5.95%	1.41%	3.20%	4.53%
查全率 1-new	6.39%	4.52%	8.01%	5.27%	7.08%	1.96%	5.54%
F1 值 1-new	7.02%	2.77%	7.96%	5.61%	4.26%	2.56%	5.03%
查准率 2-new	4.55%	0.80%	1.35%	3.06%	2.49%	3.28%	2.59%
查全率 2-new	4.35%	2.10%	2.88%	1.88%	4.31%	2.23%	2.96%
F1 值 2-new	4.49%	1.44%	2.11%	2.47%	3.39%	2.76%	2.78%

平均 F1 值达到 93.46%。分析表 2 可以看出，相比较于经典的 CHI 方法，本文优化模型的查准率指标平均提高 4.53%，最大提高率为经济类文本的 7.93%，最低提高率为艺术类的 1.07%；在查全率方面效果更好平均增长率 5.54%。相比较于文[6]提出的 CHI 改进方法，本文的优化模型效率仍然较好。查全率增长值接近 3%，综合评估指标 F1 值提高 2.78%。整体上看本文提出的改进 CHI 特征选择方法具有较强的特征选择效果。

6. 结语

特征选择是文本分类中必须面对的主要问题。好的特征选择方法对特征项选择的质量以及随后的文本分类效率影响甚大。对于经典的特征选择方法进行深入研究，改进其方法设计上的不足，对于文本信息挖掘研究具有重要意义及现实需求。本文针对 CHI 模型的欠缺，从四个不同角度逐步对其优化，使之具有更强的特征选择能力。

作为文本特征降维的两种主要模式，特征选择和特征抽取各有其优点，同时也各自存在不足之处。特征抽取方法具有坚实的数学理论支撑，但是其算法复杂度较高，很难适应于大规模文本分类，特别是应用于网络的实时在线分类需求；而特征选择模式以其语义性及计算量较低而广为应用，但是特征项之间隐含的相关性难以量化。探讨这两种主流特征降维方法的结合应该是一个文本特征降维的可行途径，这也是我们下一步要进行的工作。

基金项目

国家自然科学基金(71071161，61273209)；江苏省自然科学基金(BK2012511)。

参考文献 (References)

[1] Yang, Y.M. and Liu, X. (1999) A re-examination of text categorization on methods. *Proceedings of the* 22*nd Annual International ACM SIGIR Conference Research and Development in Information Retrieval*, New York, 15-19 August 1999, 42-49. http://dx.doi.org/10.1145/312624.312647

[2] 王光, 邱云飞, 史庆伟 (2012) 集合 CHI 与 IG 的特征选择方法. *计算机应用研究*, **7**, 2454-2456.

[3] Meesad, P., Boonrawd, P. and Nuipian, V. (2012) A chi-square-test for word importance differentiation in text classification. *Proceedings of* 2011 *International Conference on Information and Electronics Engineering*, Singapore, 110-114.

[4] 邱云飞, 王威, 刘大有, 等 (2012) 基于方差的 CHI 特征选择方法. *计算机应用研究*, **4**, 1304-1306.

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

[6] 林少波, 杨丹, 徐玲 (2012) 基于类别相关的新文本特征提取方法. *计算机应用研究*, **5**, 1680-1683.

[7] 郭颂, 马飞 (2013) 文本分类中信息增益特征选择算法的改进. *计算机应用与软件*, **8**, 139-142.

[8] 黄志艳 (2013) 一种基于信息增益的特征选择方法. *山东农业大学学报*, **2**, 252-256.

[9] 丁璇 (2002) 中文网页标引源主题表达能力的调查. *大学图书馆学报*, **6**, 70-72.

[10] 侯汉清, 张成志, 郑红 (2005) Web 概念挖掘中标引源加权方案初探. *情报学报*, **1**, 87-92.

[11] 刘海峰, 姚泽清, 汪泽焱, 等 (2009) 一种基于位置的文本特征加权方法研究. *微电子学与计算机*, **2**, 188-192.

[12] http://www.nlpir.org/download/tc-corpus-answer.rar