DOI: 10.13718/j.cnki.xdzk.2015.06.022

基于最低词频 CHI 的特征选择算法研究①

肖　雪[1]，　卢建云[1,2]，　余　磊[3]，　龚　恒[4]

1. 重庆电子工程职业学院 软件学院，重庆 401331；2. 重庆大学 计算机学院，重庆 400044；
3. 重庆师范大学 计算机与信息科学学院，重庆 400044；4. 重庆电子工程职业学院 应用电子学院，重庆 401331

摘要：CHI 是文本分类中特征选择的重要方法. 本文分析了 CHI 特征选择的特点，针对该方法的不足之处，提出了一种新的基于最低词频 CHI 的特征选择算法. 该方法通过设置最低词频阈值去除了部分低频词，减少了 CHI 特征选择时低频词带来的干扰. 同时本文对传统的 TF-IDF 特征权重计算方法进行了改进，在特征权重计算里加入改进后的 CHI 特征选择函数，使文本的表示更合理. 通过在均衡语料和非均衡语料上的实验验证，新的方法有效提高了文本分类的效果.

关　键　词：文本分类；向量空间模型；特征选择；χ^2 统计；低频词；权重计算

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673-9868(2015)06-0137-06

文本分类是大规模信息处理的重要手段. 文本分类流程大致有以下 5 个环节[1]：原始文本预处理、文本特征降维、特征加权、训练文本分类器、用分类器进行分类.

目前，文本文档通常采用 G. Salton 提出的向量空间模型(Vector Space Model，VSM)来表示. 在向量空间模型中，文本的内容可以通过各种特征项(字、词、词组或短语等)来表示为向量的形式. 假设文档所属的类别仅与这些特征项出现的频率有关，而与这些特征项出现的顺序或位置无关，则文本可表示为 $D=(t_1, t_2, \cdots, t_i, \cdots, t_n)$，其中 t_i 表示第 i 个特征项. 在这个文本向量中，特征项通过被赋予的权重 W 来表示这个特征项在该文本中的重要程度，可表示为 $D=(W_1, W_2, \cdots, W_i, \cdots, W_n)$，其中 W_i 表示第 i 项的权重. 2 个文本 D_1 和 D_2 之间的相关程度常用它们的相似度 $\mathrm{Sim}(D_1, D_2)$ 来度量，常用的向量计算方法有内积计算、夹角余弦计算等. 可 VSM 作为英文文本的一种有效表示方法在中文文本表示上有一定的缺陷和不足[2]，在此基础上有学者提出了基于语义分析的隐含语义索引(Latent Semantic Indexing，LSI). LSI 利用词语与概念之间的映射关系，通过奇异值分析将文本中的索引词映射到低维空间中进行分析.

现有的分类算法一般是基于 VSM 的特征独立性算法，忽略了文本内词语之间的语义关系，文本被表现为分量间关系独立的向量，主要利用数学统计方法将文本分类问题转换为数学分析. 主要有：类中心向量最近距离判别算法、朴素贝叶斯(Naïve Bayes)[3]、K-邻近算法(KNN)[4]、支持向量机(SVM)[5]、神经网络(Neural Network)、决策树(DecisionTree)等，以及对多种分类方法的融合[6]. 本文针对文本分类流程里的特征选择、特征加权环节，分析了常用特征选择方法 CHI 的不足，提出了基于最低词频 CHI 的特征选择算法，并把它融入到特征权重计算里，对传统的权重计算方法进行了改进.

1　特征选择方法 CHI

对文档向量而言，文本特征空间的高维性和稀疏性对文本分类系统的分类效率和分类精度有直接的

① 收稿日期：2014-01-18
基金项目：国家自然科学基金项目(61462008)；重庆市教委科技项目(KJ120622).
作者简介：肖　雪(1981-)，女，四川射洪人，硕士，讲师，主要从事数据挖掘、图像处理、模式识别等方面的研究.

影响. 因此，除分类算法以外，特征降维对于最后的分类性能有重要的影响. 特征降维包括特征选择和特征抽取，特征选择的方法很多，其基本思想是使用某种特征评估函数对每个特征项打分，然后按照分值从高到低排序，最后取分值靠前的一定数量的特征项作为特征集合. 常用的特征选择方法有文档频率(Document Frequency，DF)、χ^2 统计(CHI)、信息增益(Information Gain，IG)、互信息(Mutual Information，MI)、期望交叉熵(expected cross entropy，CE)等[7]. 已有实验表明，IG 和 CHI 是最有效的 2 种特征选择方法[8].

其中 CHI 方法是基于数理统计中常用的检验 2 个变量独立性的方法[9]. 该方法中，特征 t 和类别 c 的 CHI 值为

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

其中 A 表示包含特征 t 并且属于类别 c 的文档数；B 表示包含特征 t 并且不属于类别 c 的文档数；C 表示不包含特征 t 并且属于类别 c 的文档数；D 表示不包含特征 t 并且不属于类别 c 的文档数；N 表示文档集合中的文档总数. 如表 1 所示.

表 1 CHI 公式含义

文档数	属于类别 c	不属于类别 c	总数
包含特征 t	A	B	$A+B$
不包含特征 t	C	D	$C+D$
总数	$A+C$	$B+D$	$A+B+C+D$

由于 CHI 方法衡量的是 2 个变量的相关程度，根据公式，当特征 t 和类别 c 完全独立时，$\chi^2(t,c)=0$；当特征 t 和类别 c 的相关性越强时，$\chi^2(t,c)$的值就越大.

特征 t 的全局 CHI 值可用求平均或最大值的方式求得，计算公式如下

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

$$\chi^2_{\max}(t)=\max_{i=1}^{m}\{\chi^2(t,c_i)\} \tag{3}$$

其中 m 表示文档集合中的类别总数，$P_r(c_i)$表示一篇文档在第 i 个类别中出现的概率.

2 改进的特征选择算法

公式(1)中，A 和 B 的值来源于对文档集合中包含特征 t 的文档数的统计，它只关心是否出现了特征 t，却不在乎特征 t 出现的频率. 这使得传统的 CHI 方法只通过特征出现与否来衡量特征在分类时所能提供的信息，这就可能出现这种情况：一个词在一类文章的每篇文档中都只出现了一次，其 CHI 值却大于在该类文章 99%的文档中出现了 10 次的词.

例如，给定一个文档集合中的总文档数为 300，类别 c 的文档数为 100. 特征 t_1 和 t_2 只在类 c 里出现，其中 t_1 在类 c 的每篇文档里都出现了一次，t_2 在类 c 的 99 篇文档里都各出现 10 次. 根据公式(1)不难计算出表 2 结果.

表 2 根据公式(1)计算 CHI 结果

	A	B	C	D	CHI
t_1	100	0	0	200	300
t_2	99	0	1	200	295.5

根据描述可以推论出特征 t_2 才是更具代表性的特征，但只因为它出现的文档数比前面的词少了“1”，特征选择的时候就可能筛掉 t_2 而保留了 t_1. CHI 方法由于只考虑特征是否出现这种情况，从而忽略了特征里的其他有用信息，造成对低频词有所偏袒，夸大了低频词的作用，容易在特征选择的过程中引入噪声词. 特别在类分布不均匀的情况下，传统 CHI 方法容易给一些不重要的低频词较高的评价，对最后的分类效果有较大影响.

为了弥补只考虑特征是否出现而带来的“低频词缺陷”，本文提出最低词频 CHI 算法作为对传统 CHI 方法改进. 该算法如下：

1）对某一类别 c，特征 t 在类别 c 的文档里出现时的平均词频为 $\overline{tf(t,c)}=\dfrac{\sum_{d\in c} tf(t,d)}{n_c}$，其中 $tf(t,d)$ 为特征 t 在文本 d 中出现的频数，$\sum_{d\in c} tf(t,d)$ 为特征 t 在类别 c 里出现的总词频，n_c 为类别 c 的总文档数. 若特征 t 的平均词频 $\overline{tf(t,c)}$ 大于事先给定的最低词频阈值 $F_{\min}$，那这样的特征记作 $t_{\min}$.

2）对任一类别 c_i，都能获得一个 $t_{\min}$ 的集合 T_{c_i}，最后得到 $T_{\min}=T_{c_1}\cup T_{c_2}\cup\cdots\cup T_{c_m}$，作为对原有特征集合的最低词频筛选.

3）对 $T_{\min}$ 中的任一特征 $t_{\min}$，通过公式(1)计算其 CHI 值，最后通过公式(2)取得该特征的全局 CHI 平均值

$$\chi^2_{\text{avg}}(t)=\sum_{i=1}^{m}P_r(c_i)\chi^2(t,c_i)$$

3　改进的特征权重计算

目前最普遍的赋权重的方法是运用文本的统计信息，即统计的方法，主要是采用词频、逆文档频率来计算特征项的权重. 比如布尔权重、TFIDF 型权重、WIDF 型权重等. 其中 TFIDF 权重的计算公式为

$$W(t,d)=\frac{tf(t,d)\times\log\left(\dfrac{N}{n_t}+0.01\right)}{\sqrt{\sum_{t\in d}\left[tf(t,d)\times\log\left(\dfrac{N}{n_t}+0.01\right)\right]^2}} \tag{4}$$

其中，$tf(t,d)$ 为特征 t 在文本 d 中出现的频率，n_t 为文本集中含有 t 的文本的数量，N 为文本集合中总的文本数量.

通过公式(4)可以看出，在传统的特征权重计算方法里，特征选择函数并未纳入权重表示，导致权重计算和特征选择过程完全割裂，不能精确反映特征的重要性. 文献[10]通过实验指出特征函数能够起到过滤噪音特征的作用，因此权重由 TF,IDF 和特征选择函数 3 者组合在一起，会达到比较好的分类效果. 本文根据这种思想，在传统权重公式(4)的基础上，加入了本文所采用的特征选择函数 CHI 进行特征权重计算，公式(5)如下：

$$W(t,d)=\frac{tf(t,d)\times\log\left(\dfrac{N}{n_t}+0.01\right)\times\chi^2(t)}{\sqrt{\sum_{t\in d}\left[tf(t,d)\times\log\left(\dfrac{N}{n_t}+0.01\right)\times\chi^2(t)\right]^2}} \tag{5}$$

4　实验结果及分析

实验采用的语料来自于复旦大学李荣陆教授提供的中文文本分类语料库，语料库分为 20 个类别，共有文档 19 637 篇，其中训练文档 9 804 篇，测试文档 9 833 篇，基本按照 1∶1 比例划分. 但各类别所包含的文档数并不均衡，最多的经济类共有 3 201 篇，最少的通信类仅有 52 篇. 本次实验按各类别文档数从多到少排列，选择文档数大于 1 000 的前 8 个类别：经济、计算机、体育、环境、政治、农业、艺术、太空. 在去除无效文档后，本次实验抽取均衡语料和非均衡语料 2 个语料集，具体分布如表 3、表 4 所示.

表 3　均衡语料集文档分布

类别	训练集文档数	测试集文档数	类别	训练集文档数	测试集文档数
经济	576	200	政治	550	200
计算机	582	200	农业	546	200
体育	565	200	艺术	543	200
环境	556	200	太空	540	200

表4 非均衡语料集文档分布

类别	训练集文档数	测试集文档数	类别	训练集文档数	测试集文档数
经济	1 390	200	政治	550	200
计算机	1 080	200	农业	546	200
体育	565	200	艺术	543	200
环境	556	200	太空	60	60

实验采用中国科学院的汉语词法分析系统ICTCLAS对训练文本集合中的文本进行分词，采用了性能较好的KNN分类算法，其中采用向量间的夹角余弦值作为文本间的距离，K值取10.

经反复实验，本文取最低词频阈值$F_{\min}=4$，特征数量为1 000. 实验1和实验2的结果评估采用$F1$测试值作为指标. $F1$测试值公式为

$$F1\text{测试值}=\frac{\text{准确率}\times\text{召回率}\times 2}{\text{准确率}+\text{召回率}}$$

实验1 实验基于均衡语料，采用本文所提出的最低词频CHI特征选择＋改进的特征权重计算方法(见公式(5))，与最低词频CHI特征选择＋传统特征权重计算方法、传统CHI特征选择＋传统特征权重计算方法的分类效果做比较，结果如表5所示.

表5 均衡语料上的$F1$(%)值比较

类别	最低词频CHI特征选择＋改进的特征权重计算	最低词频CHI特征选择＋传统特征权重计算	传统CHI特征选择＋传统特征权重计算
经济	85.92	81.70	78.31
计算机	91.35	89.28	86.95
体育	90.78	88.91	87.83
环境	86.13	82.29	79.58
政治	87.57	83.16	82.78
农业	88.50	86.81	84.42
艺术	88.49	83.85	83.64
太空	86.43	84.42	82.03

由表5可知，在语料较均衡的情况下，“经济”类的分类效果最差，这与本类中的特征词模糊有关. 当采用传统的特征权重计算方法时，最低词频CHI特征选择方法的总体分类效果比传统CHI特征选择有所提高，说明通过抑制低频词能减少低频词对分类效果的干扰. 在此基础上，采用改进的特征权重计算方法时，分类效果有较大提高，特别是“经济”类提高最大，证明在特征权重因子中加入CHI特征选择函数能更好地表示当前特征，以此提高分类效果.

实验2 实验基于非均衡语料，采用本文所提出的最低词频CHI特征选择＋改进的特征权重计算方法，与最低词频CHI特征选择＋传统特征权重计算方法、传统CHI特征选择＋传统特征权重计算方法的分类效果做比较，结果如表6所示.

表6 非均衡语料上的$F1$(%)值比较

类别	最低词频CHI特征选择＋改进的特征权重计算	最低词频CHI特征选择＋传统特征权重计算	传统CHI特征选择＋传统特征权重计算
经济	72.86	66.83	60.46
计算机	88.73	78.64	69.42
体育	90.64	82.57	76.86
环境	82.15	77.90	70.93
政治	83.89	79.65	73.48
农业	81.21	77.16	71.57
艺术	84.27	80.02	74.34
太空	68.78	64.30	62.02

由表 6 中的数据可知，在非均衡语料情况下，采用传统的特征权重计算方法时，使用最低词频 CHI 特征选择方法的分类效果比传统 CHI 特征选择有明显提高，特别是“计算机”类别，证明在样本分布不均匀时，通过对低频词的过滤，能很好地降低对 CHI 特征选择的干扰. “太空”类分类性能提高不大，经过分析，在样本数过小的情况下，某些“太空”类的特征词作为低频词被筛选掉，影响了分类效果. 在此基础上，当采用改进的特征权重计算方法时，分类效果也有所提高，同样证明在特征权重因子中加入特征选择函数能更好地表示当前特征.

同时，通过表 6 与表 5 的实验结果对比分析，可看出 CHI 特征选择方法在语料均衡时分类效果较好，在语料不均衡的情况下分类效果有明显下降. 可知在样本分布不均匀时，很多小类文本被误分到了大类，影响了大类文本的准确率及小类文本的召回率.

实验 3　实验采用本文所提出的最低词频 CHI 特征选择＋改进的特征权重计算方法，在均衡语料和非均衡语料上，对最低词频阈值 $F_{\min}$取不同值时的宏平均 $F1$ 值进行比较，比较结果如表 7 所示. 宏平均 $F1$ 值评价的是整个数据集上的分类性能，公式为

$$F1_{\text{Macro}}=\frac{1}{m}\sum_{i=1}^{m}F1_i$$

其中 m 为文档类别总数.

表 7　本文方法在不同最低词频阈值下的宏平均 $F1$(%)值比较

最低词频阈值 $F_{\min}$	均衡语料	非均衡语料	最低词频阈值 $F_{\min}$	均衡语料	非均衡语料
2	84.66	71.14	7	84.41	72.65
3	87.83	79.33	8	83.29	70.81
4	88.15	81.57	10	82.70	67.77
5	88.18	79.98	20	79.57	61.20
6	87.04	76.62			

由表 7 中的数据可知，最低词频阈值 $F_{\min}$取不同值时，在均衡语料上的分类性能总体要高于非均衡语料；在同一语料的情况下，非均衡语料采用本文方法的优势更加明显，在 $F_{\min}$值设置适当时分类性能有很大提高，但不同 $F_{\min}$值造成的分类效果差异较大. 并且在本实验语料上 $F_{\min}=4$ 时整体分类效果最好. $F_{\min}$值设置过小，对低频词的抑制效果不明显，一些在大量文档里都出现、不具分类意义的特征项没有得到过滤，对分类过程存在干扰；$F_{\min}$值设置过大，会把部分只在少量文章中高频出现的具有代表性的特征项(比如一些专有词汇)过滤掉，影响分类效果. 因此，阈值的设置和语料的复杂性密切相关，可根据反复实验获得最优值.

5　结　论

针对 CHI 特征的特点，本文提出了最低词频 CHI 特征选择算法，用来弥补 CHI 特征选择对低频词有所偏袒的缺陷. 同时在特征权重计算里加入特征选择函数，对传统的 TFIDF 特征权重计算方法进行了改进，使文本的表示更合理. 通过在均衡语料和非均衡语料上的实验验证，新的方法有效提高了中文文本的分类效果.

参考文献：

[1]　张玉芳，万斌侯，熊忠阳. 文本分类中的特征降维方法研究 [J]. 计算机应用研究，2012，9(7)：2541－2543.

[2]　宋胜利，陈　平，王少龙. 面向文本分类的中文文本语义表示方法 [J]. 西安电子科技大学学报：自然科学版，2013，40(2)：109－119.

[3]　JOACHIMS T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization [C]// Proceedings of the International Conference on Machine Learning. Nashville，US：CiteSeerX，1996：143－151.

[4]　王小青. 基于并行遗传算法的 KNN 分类方法 [J]. 西南师范大学学报：自然科学版，2010，35(2)：103－106.

[5]　KWOK J T Y. Automated Text Categorization Using Support Vector Machine [C]//In Proceedings of the International Conference on Neural Information Processing. Japan：ICONIP，1998：347－351.

[6] 李健苹，游中胜. 一种新的多分类器融合方法 [J]. 西南师范大学学报：自然科学版，2014，39(1)：126－130.
[7] SANTANA L E A, DE OLIVEIRA D F, CANUTO A M P, et al. A Comparative Analysis of Feature Selection Methods for Ensembles with Different Combination Methods [C]// Proceedings of International Joint Conference on Neural Networks. Orlando, Florida, USA: IEEE, 2007: 643－648.
[8] YANG Y M. An Evaluation of Statistical Approaches to Text Categorization [J]. Journal of Information Retrieval, 1999, 1(2): 67－88.
[9] 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.
[10] 张爱华，靖红芳，王 斌，等. 文本分类中特征权重因子的作用研究 [J]. 中文信息学报，2010，24(3)：97－104.

Research on Feature Selection Algorithm Based on the Lowest Term Frequency of CHI

XIAO Xue[1], LU Jian-yun[1,2], YU Lei[3], GONG Heng[4]

1. *College of Software, Chongqing College of Electronic Engineering, Chongqing 401331, China;*
2. *College of Computer Science, Chongqing University, Chongqing 400044, China;*
3. *College of Computer and Information Science, Chongqing Normal University, Chongqing 400044, China;*
4. *College of Applied Electronic, Chongqing College of Electronic Engineering, Chongqing 401331, China*

Abstract: CHI is an important method of feature selection in text categorization. In order to overcome the deficiencies of this method, a new feature selection algorithm based on the lowest word frequency of CHI is proposed in this paper. This new approach removes some low-frequency terms by setting a threshold value of the lowest frequency terms, and thus reduces the interference of the low-frequency terms in CHI feature selection. Meanwhile, the classical feature weighting method of TF-IDF is improved in this study, and the addition of an improved feature selection function of CHI to the feature weighting method makes the text more reasonable. The experimental results on the corpora of even distribution and uneven distribution show that the new approach has effectively improved the quality of text categorization.

Key words: text categorization; vector space model; feature selection; Chi-square statistic; low-frequency term; term weighting

责任编辑 崔玉洁