第44卷第3期
2012年6月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 44 No. 3
Jun. 2012

术语定义抽取的特征选择框架

潘 渭 顾宏斌 赵芷晴

(南京航空航天大学民航学院,南京,210016)

摘要:为了进一步提升航空领域术语定义抽取的精度和效率,提出了一种不依赖已有特征选择方法的特征选择框架。该框架结合了分类特征的类间分布差异和类内分布差异,更好地表达了术语定义内部各子概念间特征分布的差异对划分类别的贡献。在分析该框架和传统过滤器特征选择方法对特征分布的影响的基础上,在航空领域术语定义语料库中对实验结果进行了对比。结果表明,本文提出的方法在使用平衡随机森林方法时,取得的最好成绩为F_1-measure=0.652,F_2-measure=0.761,所需特征比例从30%~40%降低到20%~30%;在使用直接分类方法时,F_1-measure成绩提高了2.57倍,F_2-measure成绩提高了3.11倍,均优于过滤器方法和Fisher Score方法。

关键词:特征选择;不平衡语料;定义抽取;文本分类;小析取项

中图分类号:TB941 **文献标识码**:A **文章编号**:1005-2615(2012)03-0399-06

Feature Selection Framework Research in Extracting Term Definition

Pan Xu, *Gu Hongbin*, *Zhao Zhiqing*

(College of Civil Aviation, Nanjing University of Aeronautics & Astronautics, Nanjing, 210016, China)

Abstract: A feature selection framework not reling on existing feature selection method in extracting definitions is extracted from aviation professional corpus. The framework combines between-class distribution difference and within-class distribution difference of features to express contribution of small disjuncts. After analyzing influence of traditional filter method and the framework on feature distribution, experimental results are compared in corpus of term definition corpus of aviation. In BRF classification, features required to obtain the best scores F_1-measure = 0.652, F_2-measure = 0.761 is decreased from 30%—40% to 20%—30% by using the proposed framework. In SVM classification, F_1-measure of classifier using the framework is increased by 2.57 times and F_2-measure is increased by 3.11 times. The results are superior to the filter method and the Fisher Score method.

Key words: feature selection; unbalanced corpus; definition extraction; text categorization; small disjunct

随着国内航空业进入高速发展的新阶段,对从业人员的持续培训以及为航空安全、适航进行数据、知识的积累和分析成为一种常态的任务。这使得对基于计算机的培训技术(Computer based training, CBT)以及各种专业知识库的需求迅速增长,航空术语定义的抽取就是建立行业相关知识库和以知识库为基础开展智能培训的重要基础工作之一。使用分类方法处理术语定义抽取可以被看作是一个不平衡数据分类的过程[1-3],特征选择是该过程中决定分类精度和效率的关键技术之一。

有实验表明,传统的过滤器(filter)特征选择方法由于倾向于选择高频词汇,会导致不平衡数据中的少数类别被淹没。因此在不平衡语料分类上的效果不够理想[1,4-5]。Japkowicz在2003年指出,不

基金项目:中国民航局民航应用研究基金(MHRD0723)资助项目。

收稿日期:2011-06-20;**修订日期**:2012-01-15

通讯作者:顾宏斌,男,教授,博士生导师,1957年出生,E-mail:ghb@nuaa.edu.cn。

平衡数据中存在的小析取项问题是导致分类器性能不佳的重要原因[6]。本文从如何在特征选择阶段更好地选取能够反应各小析取项内部信息的特征出发，从统计学角度说明了这些特征的分布特点，并据此提出一种结合特征类间分布差异和类内分布差异的特征选择框架。该框架不依赖已有的特征选择方法，比常用filter特征选择方法更好地适应不平衡语料的分类问题。

近几年的对特征选择的进一步研究多集中于包含丰富类别信息特征词的选择上。这些方法倾向于更多地采用在类别间出现概率差异相对更大的特征；或者通过调整特征选择公式中不同类别的权重来平衡不同类别的重要性[7-10]。

在采用新思路建立适合不同平衡度数据的特征选择框架方面，近年来的主要成果包括靖红芳等人通过调整少数类别和普通类别的权重，定义的基于类别分布的特征选择框架(Category distribution-based feature selection，CDFS)[11]。徐燕等人基于特征的类别区分能力定义的高性能特征选择算法[8]。崔自峰等人基于特征的不同相关程度，建立Markov Blanket理论和特征相关性之间的联合，并结合卡方检验提出的特征选择方法[12]。Zheng等人提出的显式近似最优组合正特征和负特征能够提高特征选择在非平衡语料上的效果。但是这种框架依赖于其他方法，只有与它所依赖的方法能够很好地选择正特征和负特征时此方法才能得到很好的效果[13-14]。文献[1,2]的研究也表明，在处理极不平衡的术语定义抽取语料时，先构建平衡数据集，再进行特征选择，比在原数据集上进行特征选择的效果更好。

1 基于类间和类内分布差异的特征选择框架

在进行术语定义抽取时，除了数据分布类间不平衡外，还存在类内不平衡问题，即小析取项(Small disjuncts)。它反应了同一类别中若干子概念之间的学习样本分布的不平衡性。在术语学研究中，冯志伟[15]将术语定义大致分为6种类型和3种定义方式，其用词和造句各不相同，但每种类型和定义方式又有各自的规律。这些不同类型的术语定义，构成了术语定义大类别中的子概念。为解决以上问题，本文着眼于提升对定义句内部各子概念的特征表达，着力于寻找具有较强类别区分能力的特征，分别用于解决不平衡分类的类间不平衡和类内不平衡问题。

1.1 特征选择框架定义

本文提出的结合两类分布差异的特征选择框架由3个部分组成，分别对应特征的类间分布差异、类内分布差异、数据分布加权。论述如下：

特征t在类别A中出现频率较高，而在类别B出现频率较低。这种情况构成了特征t的类间分布差异。由此可得定义如下：

定义1 令特征t在类别C_i中出现的频率为$TF_{(i,t)}$，特征t在样本空间中出现的频率的均值为$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}TF_{(i,t)}$，则特征$t$的类间分布差异函数对应类别$C_i$的部分可以定义为

$$BC_{(i,t)}=\log\frac{E\{[TF_{(i,t)}-E(TF_{(i,t)})]^2\}}{\overline{X}}=\log\frac{\sum_{i=1}^{n}TF_{(i,t)}^2-n\times\overline{X}^2}{n\times\overline{X}} \tag{1}$$

特征t在类别A的某个子概念的实例中集中出现，而在类别B的实例中均匀出现，这种不同构成了特征t的类内分布差异。由此可得定义如下：

定义2 令特征t在属于类别C_i的文档d中出现的次数为$TF_{(d,t)}$，在文档d中出现的频率为$F_{(d,t)}=\frac{TF_{(d,t)}+1}{|d|}$，其在类别$C_i$中出现的频率的均值为$\mathrm{mean}F_{i(d,t)}=\frac{1}{n}\sum_{d\in i,d=1}^{n}F_{(d,t)}$，则特征$t$的类内分布差异可以用特征在不同类别内部出现频率的标准差来进行比较$\mathrm{std}F_{i(d,t)}=\sqrt{\frac{\sum_{i=1}^{n}(F_{(d,t)}-\mathrm{mean}F_{i(d,t)})^2}{n-1}}$。特征$t$的类内分布差异函数对应类别$C_i$的部分可以定义为

$$WC_{(i,t)}=\log\frac{\mathrm{std}F_{i(d,t)}}{\mathrm{mean}F_{i(d,t)}}=\log\frac{\sqrt{\frac{\sum_{i=1}^{n}(F_{(d,t)}-\mathrm{mean}F_{i(d,t)})^2}{n-1}}}{\frac{1}{n}\sum_{d\in i,d=1}^{n}F_{(d,t)}}=\log\frac{n\sqrt{(n-1)\sum_{i=1}^{n}(F_{(d,t)}-\mathrm{mean}F_{i(d,t)})^2}}{(n-1)\sum_{d\in i,d=1}^{n}F_{(d,t)}} \tag{2}$$

在处理类似定义抽取这样的不平衡数据分类时，为使本文提出的特征选择框架能够适用于不同的分类策略，可定义一个用于不同策略中的相应调

整权重。

定义 3　令特征 t 在训练数据的 C_i 类和其他类别中出现次数的比值为 W_p，令训练数据中 C_i 类实例数量和其他类别的数量比为 W_i，则定义数据平衡度权重为

$$\mathrm{Weight}t_d = (W_p - W_i)^2 \tag{3}$$

定义 4　特征选择框架函数可定义为

$$FS_{(t)} = \mathrm{Weight}t_d \times \sum_{i=1}^{n}(BC_{(i,t)} \times WC_{(i,t)}) = (W_p - W_i)^2 \times \sum_{i=1}^{n}\left(\log\frac{\sum_{i=1}^{n}TF_{(i,t)}^2 - n\times\overline{X}^2}{n\times\overline{X}} \times \log\frac{n\sqrt{(n-1)\sum_{i=1}^{n}(F_{(d,t)} - \mathrm{mean}F_{i(d,t)})^2}}{(n-1)\sum_{d\in i,d=1}^{n}F_{(d,t)}}\right) \tag{4}$$

特别的，对于二分类问题，由于特征在两个类别间出现频率的方差是相同的，这时式(4)中的特征选择框架函数由于类间分布差异函数 $BC_{(i,t)}$ 相同，会退化为只考虑特征类内分布差异的特征选择方法，因此在处理二分类问题时，修改特征的类间分布差异函数定义如下。

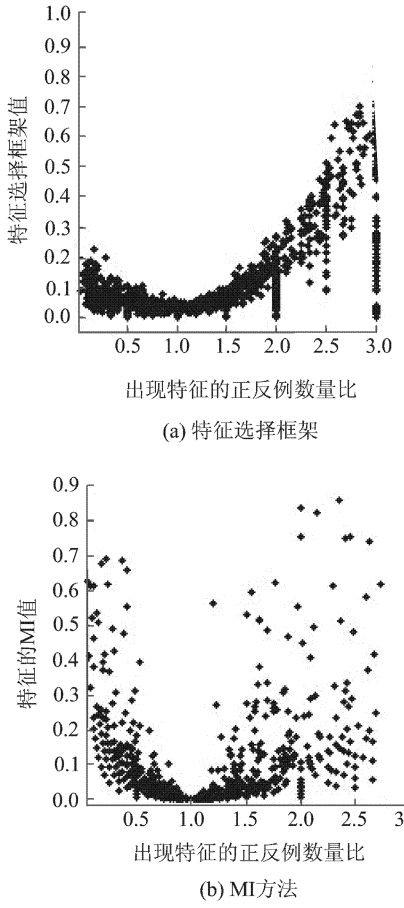
定义 5　令特征 t 在类别 C_i 中出现的频率为 $TF_{(i,t)}$，特征 t 在样本空间中出现的频率的均值为 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}TF_{(i,t)}$，则特征 t 的类间分布差异对应类别 C_i 的部分可以定义为

$$BC_{(i,t)} = \frac{TF_{(i,t)}}{\overline{X}} \tag{5}$$

1.2　特征选择框架特性分析

由以上定义可以看出，式(4)通过加权组合的方法结合两类分布的特点。使其具有较高类内分布差异的特征，比在仅考察特征类间分布差异时，获得更高的特征值将自身从其他特征中分离出来，便于选取。同时，分别计算特征的 $BC_{(i,t)}$ 和 $WC_{(i,t)}$ 函数在不同类别中的值之后再加和，使正特征和负特征对分类的贡献都得到体现。

图 1 是以本文语料库为例，$FS_{(t)}$ 框架和以互信息(Mutual information，MI)方法为代表的传统 filter 方法在 BRF 分类策略下，训练集样本的特征值分布图，图中的特征选择函数值经过归一化处理，分布图类似开口向上的抛物线。训练集中多数特征的类间分布差异不明显，表现为其值密集分布在抛物线底部区域，并随着抛物线向上延伸变得稀疏。在 $FS_{(t)}$ 框架(图 1(a))中，有较高类内差异的特征获得较高的值，从而和其他特征分离，减少了多个特征对应相同特征值的情况，便于筛选。直接分类策略的训练样本特征值分布也有相同的趋势。在以 MI 为代表的传统 filter 方法(图 1(b))中，特征值在抛物线底部密集分布，存在多个特征对应相同特征值的情况，不易筛选。



(a) 特征选择框架

(b) MI方法

图 1　BRF 分类策略下各特征选择函数比较

2　分析与实验

2.1　实验环境设置

本文通过在二分类的不平衡语料上进行分类实验来验证提出的特征选择框架。使用的语料总计 16 627 个句子，约 55 万字，其中定义句 1 359 个，含特征词 7 923 个，为 BRF 方法准备的平衡训练集含特征词 4 300～4 700 不等[2]。采用了两种不同的分类策略，一种是改进的 BRF 方法[2]，另一种是直接在全部语料上使用 SVM 方法，取十折验证结果。以上两种分类策略中使用的分类器均为 Waikato 分析环境(Waikato environment for knowledge analysis，WEKA)自带的方法，实验在 32 位的 Linux 平台上进行。实验中的 Fisher Score 按照文献[16]

的公式计算。

本文实验使用的评价方法包括召回率(Recall),准确率(Precision),F-measure,定义如下

$$\text{Recall} = \frac{\text{被正确分类的实例数量}}{\text{该类型实例的总数}} \tag{6}$$

$$\text{Precision} = \frac{\text{被正确分类的实例数量}}{\text{被分为该类的实例数量}} \tag{7}$$

$$F_\beta = \frac{(1+\beta^2)\times P \times R}{\beta^2 \times P + R} \tag{8}$$

式(8)中β的取值由实验中召回率和准确率的重要性来决定,当β取值为1的时候(F_1-measure指标),认为召回率和准确率同等重要;当β取值为2的时候(F_2-measure指标),认为召回率比准确率更加重要。

在平衡随机森林(Balanced Handom forest,BRF)实验中,经过重采样的训练集中正反例的比例为1∶1,权重W_i设置为1;在直接分类实验中,正例占实例总数的8.17%,权重设置为$W_i = \frac{8.17\%}{1-8.17\%} = 0.89$。

2.2 实验结果及分析

图2为使用框架$FS_{(t)}$配合改进的BRF方法[2]在定义抽取试验中取得的最好结果,F_1-measure和F_2-measure指标随着聚合树数量增长而增长。当聚合树数量达到15颗之后F_2-measure指标的增长趋势平缓,最高点在0.761;F_1-measure指标则呈现波动态势,最高点在0.652。

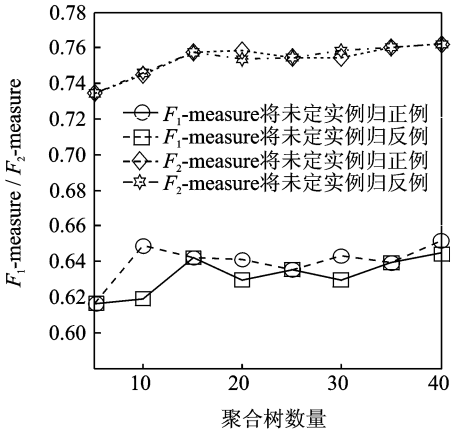


图2 特征选择框架选取25%特征的分类结果

图3为使用Fisher Score选取不同比例特征配合改进BRF方法在定义抽取实验中的结果。图3(a)选取的特征比例和框架$FS_{(t)}$最佳结果选取比例相同,此时使用Fisher Score的实验F_1-measure在聚合树数量超过35以后趋于稳定并达到最大值0.346;F_2-measure在聚合树数量达到20时取得最大值0.334,之后呈现出下降趋势。图3(b)为使用Fisher Score的最好成绩,F_1-measure从0.618开始震荡上行,并在聚合树数量达到10时取得最高值0.654,在聚合树数量为30时达到0.649并开始逐步下行;F_2-measure从0.733开始逐步提高,在聚合树数量达到20时取得最高值0.762,之后稳定在0.760左右。

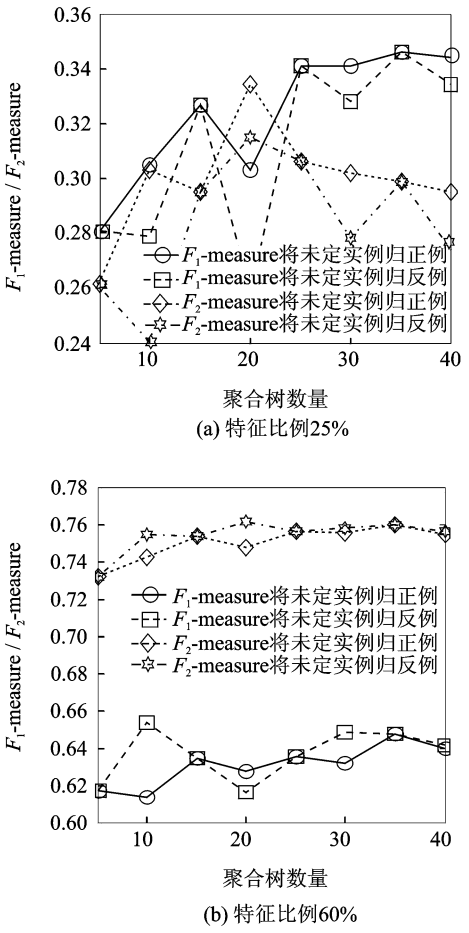
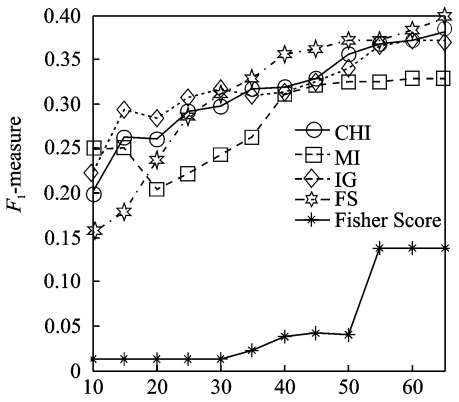
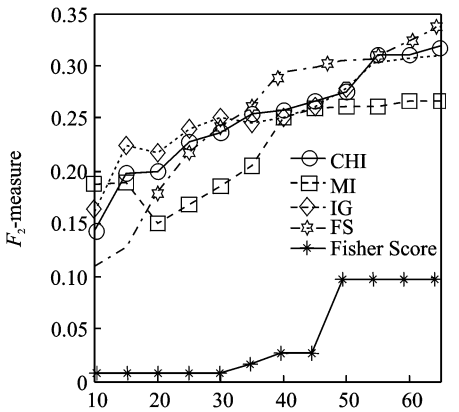


图3 使用Fisher Score选取不同比例特征的结果

图4中给出了直接分类策略下SVM分类实验的结果,对比的方法分别为卡方检验(Chi-squaretest,CHI)、信息增益(Information gain,IG)和MI。各方法的初始成绩都很差,但随着使用特征比例提高而提高,使用$FS_{(t)}$的分类器在选取特征超过20%后,成绩超过MI方法;在选取特征超过30%后,成绩领先于其他所有方法。比较F_1-measure,CHI从0.202提升到0.381,提升1.89倍;IG从0.224提升到0.372,提升1.66倍;MI从0.203提升到0.328,提升1.62倍;$FS_{(t)}$从0.155提升到0.398,提升2.57倍。比较F_2-measure,CHI从0.144提高到0.319,提升2.22倍;IG从0.161提



(a) F_1-measure成绩



(b) F_2-measure成绩

图 4　直接分类策略下 SVM 分类结果

升到 0.310，提升 1.93 倍；MI 从 0.155 提升到 0.266，提升1.76 倍；$FS_{(t)}$从0.109 提升到0.339，提升 3.11 倍。Fisher Score 方法相比其他方法，成绩差距较大。由于实验硬件平台限制，在选取特征比例超过 65%以后，无法完成实验。

由以上数据可知，在采用BRF 分类策略时，相比使用 filter 特征选择方法的结果[2]，聚合成绩相当，但达到最佳成绩所需的特征比例从30%～40%降低到 20%～30%，单颗决策树训练时间从平均 520 s 下降到330 s。和使用Fisher Score 方法相比，两者最好成绩相当，所需特征比例从 60%下降到 25%左右。

使用SVM 方法的成绩相比 BRF 方法差距较大，特别是在使用特征较少时成绩不佳，缺乏实用意义。随着使用特征数量增加，各方法成绩均有提高，其中 $FS_{(t)}$框架的成绩增长最快，并在选用特征超过 30%以后，优于其他方法，更具实用意义。

对于单个分类器而言，由于数据集中绝大部分特征密集分布于图 1 的抛物线底部区域，所以当选取的特征比例达到 20%～30%时，$FS_{(t)}$框架的选取范围进入图中抛物线的谷底部分，除了抛物线两侧类间分布差异较大的特征之外，那些由于类内差异较大而获得更高的特征函数值的特征，由于被从抛物线底部分离出来也开始被选中。在使用传统 filter 方法的实验中，按照 20%～30%的比例选取特征时，尚无法触及分布于抛物线底部的特征，因此在分类成绩上落后于使用 $FS_{(t)}$框架的方法。

3　结束语

本文结合特征的两类分布差异提出了特征选择框架 $FS_{(t)}$。通过实验证明，该框架可以在体现特征的类间分布差异的同时，也体现出特征类内差异对分类的贡献，从而在一定程度上研究了术语定义中的小析取项对抽取结果的影响。在不平衡语料的直接分类和 BRF 分类中能取得比传统过滤器方法更好的成绩，通过简单的权重设置就能够适用于不同的分类策略，能够有效提升术语定义抽取的效率和精度。

参考文献：

[1] Pan Xu, Gu Hongbin, Sun Chanjuan. A classification approach to identify definitions in aviation domain[C]// Proc of CCPR2009. Nanjing, China: [s. n.], 2009:663-669.

[2] 潘渭，顾宏斌，赵芷晴. 采用改进重采样和BRF 方法的定义抽取研究[J]. 中文信息学报，2011，25 (3)：30-37.
Pan Xu, Gu Hongbin, Zhao Zhiqing. Definition extraction with improving re-sampling and BRF[J]. Journal of Chinese Information Processing, 2011, 25 (3):30-37.

[3] Luukasz Kobylinski, Adam Przepiórkowski. Definition extraction with balanced random forests[C]// Proc of the 6th International Conference on Advances in Natural Language Processing. Gothenburg, Sweden: [s.n.], 2008:237-247.

[4] Mladenic D, Grobelnik M. Feature selection for unbalanced class distribution and Naive Bayes [C]// Proc of ICML 99. San Francisco : Morgan Kaufmann, 1999 : 258-267.

[5] 周茜，赵明生，扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报，2004，18(3)：17-23.
Zhou Qian, Zhao Mingsheng, Hu min. Study on feature selection in Chinese text categorization. [J]. Journal of Chinese Information Processing, 2004, 18 (3):17-23.

[6] Japkowicz N. Class imbalance: Are we focusing on the right issue?[C] //Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets. Washington DC:AAAI Press,2003.

[7] 吴迪，张亚平，殷福亮，等. 基于类别分布差异和VPRS特征选择的文本分类方法[J]. 电子与信息学报，2007，29(12)：2880-2884.
Wu Di, Zhang Yaping, Yin Fuliang, et al. Feature selection based on class distribution difference and VPRS for text classification[J]. Journal of Electronics & Information Technology, 2007, 29 (12): 2880-2884.

[8] 徐燕，李锦涛，王斌，等. 基于区分类别能力的高性能特征选择方法[J]. 软件学报，2008，19(1)：82-89.
Xu Yan, Li Jintao, Wang Bin, et al. A category resolve power-based feature selection method [J]. Journal of Software, 2008, 19 (1): 82-89.

[9] Li S, Zong C. A new approach to feature selection for text categorization [C] // Proc of IEEE NLP2KE. Beijing: Beijing University of Posts and Telecommunications Press, 2005: 626-630.

[10] How B C, Narayanan K. An empirical study of feature selection for text categorization based on term weight age [C] // Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC: IEEE Computer Society, 2004: 599-602.

[11] 靖红芳，王斌，杨雅辉，等. 基于类别分布的特征选择框架[J]. 计算机研究与发展，2009，46(9)：1586-1593.
Jing Hongfang, Wang Bin, YangYahui, et al. Category distribution-based feature selection framework [J]. Journal of Computer Research and Development, 2009, 46 (9): 1586-1593.

[12] 崔自峰，徐宝文，张卫丰，等. 一种近似Markov Blanket最优特征选择算法[J]，计算机学报，2007，30(12)：2074-2081.
Cui Zifeng, Xu Baowen, Zhang Weifeng, et al. An approximate Markov Blanket feature selection algorithm[J]. Chinese Journal of Computers, 2007, 30 (12): 2074-2081.

[13] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data [C]// Proc of ACM SIGKDD Explorations Newsletter. New York: ACM, 2004: 80-89.

[14] Zheng Z, Srihari R. Optimally combining positive and negative features for text categorization[C]// Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets. Washington DC: AAAI Press, 2003.

[15] 冯志伟. 现代术语学引论[M]. 北京：北京语言文化出版社，1997：31-38.

[16] Richard O Duda, Peter E Hart, David G Stork. 模式分类[M]. 第一版. 北京：机械工业出版社，2003：96-99.