DOI:10.3969/j.issn.1673-4785.201305044

基于遗传算法优化综合启发式的中文网页特征提取

沈高峰[1],谷淑敏[2]
(1.郑州轻工业学院 计算机与通信工程学院,河南 郑州 450002; 2.中原工学院信息商务学院 基础学科部,河南 郑州 450007)

摘 要:特征提取是信息检索、文本分类、文本聚类以及自动文摘生成等技术的基础。针对传统的特征提取方法不能全面有效地考查待选特征词的缺点,提出了一种基于遗传算法优化综合启发式的中文网页特征提取方法。该方法通过词频、关联度、词性以及位置等多种启发式来综合考查待选特征,并利用遗传算法来优化各启发式的权重参数。通过在不同测试集上进行对比,实验结果表明,与传统方法相比,该方法能够有效避免传统特征提取方法产生的偏差,获得具有代表性的特征集,从而使得该方法具有一定的实用价值。

关键词:特征提取;遗传算法;文本分类;文本聚类;词频;关联度
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1673-4785(2014)04-474-06

中文引用格式:沈高峰,谷淑敏. 基于遗传算法优化综合启发式的中文网页特征提取[J]. 智能系统学报, 2014, 9(4): 474-479.
英文引用格式:SHEN Gaofeng, GU Shumin. Chinese Web page feature extraction by optimizing comprehensive heuristics based on GA[J]. CAAI Transactions on Intelligent Systems, 2014, 9(4): 474-479.

Chinese Web page feature extraction by optimizing comprehensive heuristics based on GA

SHEN Gaofeng[1], GU Shumin[2]
(1.School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2. Department of Basic Subjects, College Information & Business, Zhongyuan University of Technology, Zhengzhou 450007, China)

Abstract: Feature extraction is the basis of such technologies as information retrieval, text classification, text clustering and automatic summarization. Aiming at the shortcomings of the traditional feature extraction methods which make it difficult to test feature words comprehensively and effectively, this paper proposes a method for extracting Chinese web page features by optimizing the comprehensive heuristic features based on GA. This proposed method employs comprehensive heuristics of word frequency, word correlation, parts of speech (POS) and position features to comprehensively test selected features and uses GA to optimize the weight of each heuristic parameter. The experimental results of the different test sets show that the proposed method can effectively avoid the derivations of the traditional extraction methods and obtain more representative features, and therefore it has a certain practical value.
Keywords: feature extraction; GA; text classification; text clustering; word frequency; word correlation

特征提取在自然语言处理领域有着非常广泛的应用,是信息检索、文本分类、文本聚类以及自动文摘生成等技术的关键。由于互联网资源时刻都在不断更新,中文文本呈现出“爆炸式”增长。然而,采用传统人工方式进行特征提取的方法耗时较长,且具有一定的主观性,因此快速准确地实现中文特征提取成为中文文本处理的关键。

目前,国内外学者已提出 3 类特征提取方法:基于概率统计的特征提取方法、基于传统机器学习理论的特征提取方法以及基于自然语言理解的特征提取方法。基于概率统计的特征提取方法利用文本特征的统计信息进行关键词提取,如 TFIDF[1]、词共现[2]等,该类方法具有简单、通用的特点,不需要复杂的训练过程,但准确率不高。基于传统机器学习

收稿日期:2013-05-10.
基金项目:河南省基础与前沿技术研究计划项目(102300410266);郑州轻工业学院博士科研基金资助项目.
通信作者:沈高峰.E-mail:45125301@qq.com.

理论的特征提取方法通过对大规模语料库进行学习,采用决策树[3]、贝叶斯算法[4]、最大熵模型[5]等方法对训练集进行训练,从而得到相关模型,然后再利用该模型对关键词进行提取,但该类方法较为复杂。基于自然语言理解的特征提取方法通常需要对中文文本从词、句、语义、篇章等层级进行分析,从而获得相关关键词,这类方法更加符合关键词提取的标注过程,但如何对文章进行准确的语言学分析还没有得到有效解决,该方法的抽取性能非常有限。

针对上述传统特征提取方法的特点和不足,提出了一种基于遗传算法优化综合启发式的中文网页特征提取方法。该方法首先对文本文档的分词结果进行词性标注,然后计算文档词语的词性、位置、TFIDF 以及聚集特征等综合启发式,并用遗传算法优化各启发式的权重参数,最终提取获得中文网页特征词。

1　基础知识简介

1.1　频率

TFIDF 是一种常用的信息检索方法[6]。设 N 表示给定文档集合 Ω 中的总文档数目。对于给定文档 d,采用 TFIDF 算法得到该文档中词条 t 的权重 w_t 为

$$\begin{cases} w_t = \mathrm{TF} \times \mathrm{IDF} \\ \mathrm{IDF} = \log(N/n) \end{cases} \tag{1}$$

式中:TF 表示 t 在文档 d 中出现的频率。IDF 表示文档 d 在文档集中出现的文档数目,n 表示文档集中出现特征 t 的文档数目。

从式(1)可知,如果特征 t 在文档 d 出现的次数较多而在其他文档中出现次数较少的话,那么特征 t 的权值就较大,表明该特征对文档 d 的区分能力就较强,就可以作为文档特征的候选之一。

1.2　关联度

词语的关联表现为词与词之间构成的复杂网络[7]。复杂网络方面的研究表明,汉语语言的词语之间的关联度具有高度的局部聚集性和全局连接性,能够用于表征文本特征[8]。

设 $V=\{v_1, v_2, \cdots, v_n\}$ 表示文档特征的集合,(v_i,v_j) 表示特征 v_i 和特征 v_j 之间的一条边。$G(V,E)$ 表示的是一个图,其中 V 为图的顶点集合,$E \subseteq \{(v_i,v_j) : v_i,v_j \in V\}$ 为图的边集。对于顶点 v_i,其度定义如下:

$$D_i = |\{(v_i,v_j) : (v_i,v_j) \in E, v_i,v_j \in V\}| \tag{2}$$

顶点 v_i 的聚集度 K_i 为

$$K_i = |\{(v_j,v_k) : (v_i,v_j) \in E(v_i,v_k) \in E, v_i, v_j, v_k \in V\}| \tag{3}$$

顶点 v_i 的聚集度系数 C_i 为

$$C_i = \frac{K_i}{\begin{bmatrix} D_i \\ 2 \end{bmatrix}} = \frac{2K_i}{D_i(D_i - 1)} \tag{4}$$

由式(3)和式(4)可得特征关联度为

$$CF_i = \alpha C_i / \sum_{j=1}^{N} C_j + (1 - \alpha) D_i / N \tag{5}$$

根据式(3)~(6),词语网络中节点的度和聚集度系数可以描述特征在文本中的连接特性,处于重要位置的特征往往具有较高的关联度。

1.3　词性

词性是一种浅层语言学知识的表示,该因素的获取不需要对文本进行复杂的语言学标注和分析,从而能有效避免传统采用语言学方法的缺陷。一般而言,中文文本特征的词性往往集中在名词、动词、形容词等实词中。根据人工标注结果,对特征的词性分布进行了统计分析,其结果如表 1 所示。

表 1　特征词性分布

Table 1　Characteristics distribution

词性	名词	动词	形容词	副词	其他
数目	8 431	3 405	1 830	659	675
百分比/%	56.2	22.7	12.2	4.4	4.5

从特征词性统计分布可以看到,词性能够有效表征文档的中文特征。排名前 4 位的名词、动词、形容词和副词达到关键词总数的 95.5%。因此,论文引入词性作为特征提取的重要因素之一。该因素能够有效区分停用词等,克服了传统基于统计方法无法解决高频但无实际意义的中文词语,从而提高特征提取的性能。

1.4　位置

位置是文本特征提取的一个重要因素。根据特征所在的位置,主要包括标题、摘要和正文 3 种。根据词语所在的具体位置,还可细分为小标题、起始段、中间段、末尾段、起始句、中间句、末尾句等[9]。由于网络文本一般不存在摘要,本文主要考虑特征位于标题、起始段以及其他 3 种情况。通常特征位于标题和起始段的概率较高,因此根据文本中特征所在的位置,按照标题、起始段、其他的顺序分别赋给不同的权重。

2　论文所提方法

2.1　综合启发式

仅仅根据单词频率进行特征提取的 TFIDF 方法虽然简单,但是也存在一定的缺陷,如数据集偏

斜[10],类间、类内分布偏差[11]等。而单纯依靠复杂网络中词语之间关联度的特征提取方法,则忽略了特征本身的频率,容易造成特征提取聚集到某些无意义的高频词,如“的”等,从而导致特征提取出现偏差。研究显示,融合频率和关联特征[12]能够有效避免单一方法的缺陷,从而提高特征提取的效率。

此外,仅仅依靠统计知识容易造成特征提取偏差,特别是一些高频词如“是”、“和”等容易成为特征的候选。尽管这些词可以通过建立“停词表”对其进行过滤,但是构建合适的词表非常困难,因此引入特征的词性以及位置对特征进行进一步选取。

综合以上因素,论文采用特征的频率、关联度、词性以及位置 4 个因素来衡量待选特征。对于文本中的每个特征 w,其权重计算公式为

$$\text{score}(w) = \alpha \times W_{\text{Freq}} + \beta \times W_{\text{Loc}} + \gamma \times W_{\text{CF}} + \delta \times W_{\text{POS}} \tag{6}$$

式中:W_{Freq}表示特征的 TFIDF 启发式,W_{POS}表示特征的词性启发式,W_{CF}表示特征的关联度启发式,W_{Loc}表示特征的位置启发式。每个启发式的具体描述如表 2 所示。

表 2 特征各启发式描述

Table 2 Description of feature heuristics

类型	表示	描述
词性	POS	特征的词性信息,如名词、动词、形容词等
位置	Loc	根据特征位置,分为标题、起始段和其他 3 个部分;
频率	Freq	采用 TFIDF 值表示特征的频率信息;
关联度	CF	表征特征网络之间的链接关系;

2.2 特征提取流程

特征提取的基本流程如图 1 所示,其中虚线部分为训练模块。对于给定的输入本文,特征提取具体过程如下。

图 1 本文方法特征提取基本流程

Fig.1 Flow of feature extraction in this paper

1)预处理。将网络文本去除 HTML 格式,保留文本词语的位置信息,并对文本进行分词和词性标注。

2)各启发式计算。计算文本中每个词语的 TFIDF、关联度、位置和词性等启发式。

3)启发式融合。根据多启发式融合模型,对词语的 4 个启发式进行融合,并计算得到综合得分。

4)输出结果。最后根据各特征得分的大小进行排序,选择最优的特征并输出。

2.3 遗传算法优化权重参数

本文方法中各启发式的参数权重选择是一个典型的组合优化问题。由于遗传算法简单、易理解、易实现,且在解决组合优化问题有强大的优势[13],因此,论文采用遗传算法对式(6)中的参数权重进行优化,从而得到一定范围的最佳组合参数权重。这里限定 4 个参数权重的取值范围为(0,1),并且满足 $\alpha + \beta + \gamma + \delta = 1$。然后根据经验选取适当的初始值,并经过迭代计算,得到每个启发式的参数权重。利用遗传算法获取各特征参数权重具体过程描述如下:

1)依据经验,初始化各特征参数权重 $\alpha = 0.2$, $\beta = 0.2$, $\gamma = 0.4$, $\delta = 0.2$;

2)采用十进制编码对染色体进行编码。首先把各参数都乘以 10 或 100 使它们变成整数,然后再对它们进行编码,具体格式如下:$L = \alpha\beta\gamma\delta$。其中各参数均用 3 位十进制数来表示,例如:$\alpha = 0.2$, $\beta = 0.2$, $\gamma = 0.4$, $\delta = 0.2$,则先把它们转化为 $\alpha = 020$, $\beta = 020$, $\gamma = 040$, $\delta = 020$,则相应染色体编码为:$L = 020020040020$。

3)利用各参数权重计算相应召回率,以召回率作为染色体的适应度函数,召回率计算公式为

$$\text{recall} = n/N$$

式中:n 代表同所标注的特征相符的特征的数目,N 代表文档集中所标注的特征总数目。

4)交叉和变异操作:遗传算法的收敛速度以及解的质量在很大程度上取决于交叉概率和变异概率。为了防止算法陷于局部最优以及加快算法搜索效率,仅让种群中较优个体参与交叉和变异,而当前种群最优个体则不参与。具体交叉概率和变异概率计算公式如下:

$$p_c = \begin{cases} a_1 \sin\left(\dfrac{\pi}{2} \times \dfrac{f_{\max} - f_c}{f_{\max} - f_{avg}}\right), & f_c \geqslant f_{\text{avg}} \\ a_2, & f_c < f_{\text{avg}} \end{cases} \tag{8}$$

$$p_m = \begin{cases} a_3 \sin\left(\dfrac{\pi}{2} \times \dfrac{f_{\max} - f_m}{f_{\max} - f_{\text{avg}}}\right), & f_m \geqslant f_{\text{avg}} \\ a_4, & f_m < f_{\text{avg}} \end{cases} \tag{9}$$

式中:a_1、a_2、a_3、a_4 为 0~1 的随机数,$f_{\max}$ 是当前群体中最优个体的适应度值,f_{avg} 是当前群体的平均适应度值,f_c 是参加交叉操作的个体中较大的适应度值,f_m 是变异个体的适应度值。

5)终止条件:当代种群最佳染色体适应度值和前

代种群最佳染色体适应度值之差绝对值不超过 10^{-5}。

采用遗传算法优化选择各启发式的参数权重,能够有效避免通过主观经验来确定参数的主观性,从而实现参数能够依据训练数据自适应地调优。下面的实验验证结果表明,采用该遗传算法获得参数权重能够使本文特征提取方法获得良好的提取效果。

3　实验验证

3.1　实验总体设置

以 Intel Core2 Duo CPU T6500、2.4 GHz、2 GB 内存和 Windows XP 2SP2 操作系统的 PC 机作为实验平台,以 MATLAB7.0 为仿真工具,进行 2 组实验:

第 1 组实验数据来自互联网抓取的 1 500 个中文文档,论文根据该数据集的来源将这些文档分为 5 个类别,分别包括新闻、财经、科技、体育和娱乐,各类文档数目分布均匀,都包含 300 篇文档。实验中选择每个类别的 200 篇文档作为训练集,剩下的 100 篇作为测试集。

第 2 组实验数据采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组构建的中文文本分类语料库作为实验数据,其下载网址为:http://www.nlp.org.cn/categories/default.php? cat_id=16。该语料库由 20 个类别的 14 378 篇文档组成,其中 6 164 篇为测试文本,8 214 篇为训练文本;各类别的测试文本集和训练文本集之间互不重叠,也即一篇文档仅属一个文本集并且每篇文本仅属于一个类别。该语料库各类别训练文档数分布极其不均匀,其中训练文档数较小的类别占大多数,约为 11 个类别,它们的训练文档数均少于 100 篇,如通信类文档数仅有 25 篇。

由于所选语料库是中文性质的,所以这 2 组实验都采用中科院计算技术研究所的"汉语词法分析系统 ICTCLAS"对它进行分词处理;分类工具软件都采用纽西兰的 Waikato 大学开发的 Weka 工具;因 KNN 分类器简单、易实现而被广泛应用,这 2 组实验选它作为实验分类器(其中距离采用向量夹角余弦来度量,$K=20$)。

为了对论文所提方法性能进行全面考查,论文对这 2 组实验分别做了不同方面的实验内容:第 1 组实验主要做特征词选择和召回率方面的实验;第 2 组主要做耗时和分类性能方面的实验。

3.2　第 1 组实验(各类别数据分布均匀)

在该组实验中,论文对比了基于频率的特征提取方法、基于关联度的特征提取方法以及本文方法性能。

3.2.1　特征词选择实验结果

分别采用上面 3 种方法计算全部词语的 4 个启发式值,并根据不同启发式权重进行排序,最后提取得分最高的前 10 个词语作为最后的关键词。表 4 为实验对比结果。其中,基于频率的方法用 TFIDF 来表示,基于关联度的方法用 CF 来表示,本文方法用 Multi 来表示。

表 3　3 种方法下召回率对比结果

Table 3　Comparison results of recall rate on three methods

方法	关键词	召回率/%
参考答案	负增长、收入、中央、降、季度、财政、利润、降低、涨幅、增长	—
TFIDF	负增长、收入、中央、季度、数据、降、影响、进口、今年、同比	50
CF	负增长、收入、中央、财政、季度、降低、都、进口、随后、做	60
Multi	负增长、收入、中央、财政、降低、进口、利润、季度、累计、财政部	70

从表 4 可以看出,对于"都"、"随后"这类词,本文方法能够有效地滤除。由于这类单词在文本中通常具有较高的频率,很难通过统计的方法有效去除。而且本文方法召回率能够达到 70%,表现出较好的提取性能。此外,比较特征词自动提取和人工选择,3 种提取方法都得到了"进口"这个特征词,但人工标注却忽略了这个词语。通过查看原文,"进口"确实应该标注为特征词,反映出人工选择带有较强的主观性,这种主观性很容易产生实验误差。同样也反映出特征词自动提取能够在一定程度上克服这种主观性的缺点。

3.2.2　召回率实验结果

针对测试集的不同类别,论文分别对比不同特征词提取方法的性能。由于不同类别的多启发式融合参数不同,论文利用每个类别的训练语料分别训练得到各个类别的多启发式融合参数。各特征词提取方法性能采用该类别测试集上的平均召回率表示,实验结果如图 2 所示。

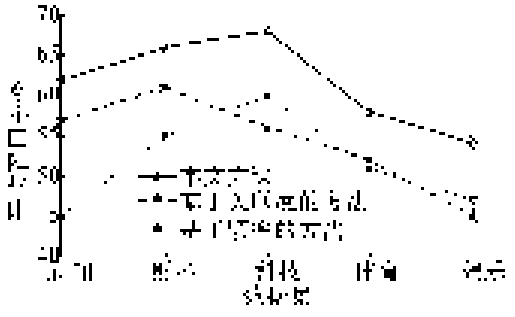

图 2　各特征提取方法在各类别下的召回率对比结果

Fig.2　Comparison results of recall rate on feature extraction methods

从图 2 可以看出,采用本文方法在各个测试集上的平均召回率均高于基于关联度的方法和基于频率的方法的性能,这说明该方法提取特征词的性能稳定,在各个类别的提取效果均得到明显提高。

3.3 第 2 组实验(各类别数据分布极其不均匀)

在这组实验中,采用宏平均 F_1 值和微平均 F_1 值作为分类性能评价标准,使用 3 种经典的特征提取方法:信息增益(IG)、x^2 统计量(CHI)、互信息(MI)与本文所提特征提取方法作比较。

3.3.1 耗时实验结果

在实验中,记录了各特征提取方法从开始执行到执行结束整个过程所消耗的时间,其结果如图 3。

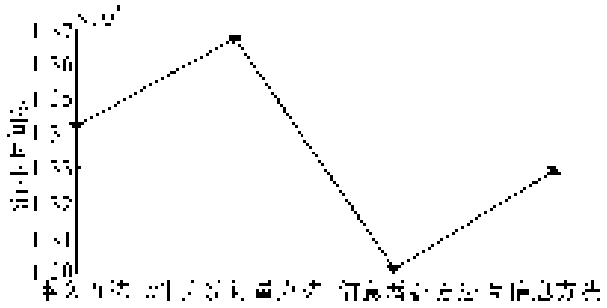

图 3 各方法消耗的时间

Fig.3 Comparison results of consuming time

由于本文方法采用了多个指标以及组合方法,耗时有所增加。从图 3 可以看出,在该组实验中,本文方法在消耗时间方面劣于互信息方法和信息增益方法,但优于最耗时的 x^2 统计量方法,但它们耗时相差不大,这也使得本文方法有一定的实用价值。

3.3.2 宏平均和微平均实验结果

从各个特征提取方法所获得的特征集(其中的特征已按权重逆序进行了排序)中,分别选取相应数目的特征对实验数据进行宏平均 F_1 和微平均 F_1 计算,具体结果如图 4 和图 5 所示。

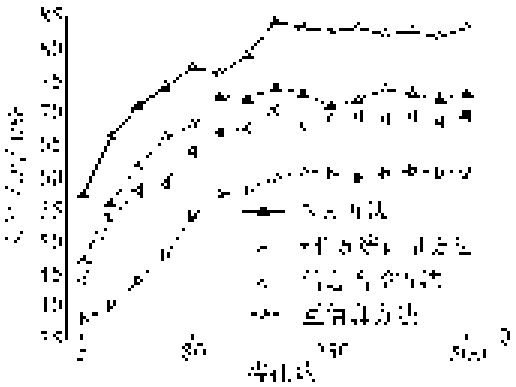

图 4 宏平均 F_1 实验结果

Fig.4 Comparison results of macro-average F_1

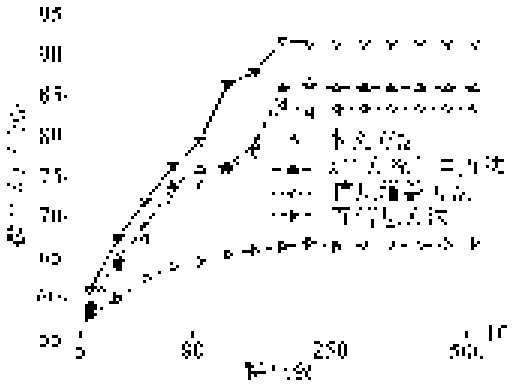

图 5 微平均 F_1 实验结果

Fig.5 Comparison results of micro-average F_1

利用特征数目的变化来考查分类器的性能,可以比较准确地反映出该分类器数对据样本变化是否敏感。图 4 表明:随着特征数目的递增,宏平均 F_1 值不断增加,但是由于实验数据中各类别样本分布极其不均匀而有所波动;图 4 表明:随着特征个数的不断增加,微平均 F_1 值也递增并趋于一个相对较稳定的值。

从图 4 和图 5 可以看出:在本文方法所选的前 1 500 个特征上 KNN 分类器性能最佳,宏平均 F_1 值约为 84%,微平均 F_1 值约为 92%;在 CHI 方法所选的前 1 500 个特征上 KNN 分类器性能最佳,宏平均 F_1 值约为 74%,微平均 F_1 值约为 86%;在 MI 方法所选的前 1 500 个特征上 KNN 分类器性能最佳,宏平均 F_1 值约为 70%,微平均 F_1 值约为 84%;在 IG 方法所选的前 2 000 个特征上 KNN 分类器性能最佳,宏平均 F_1 值约为 61%,微平均 F_1 值约为 67%。这表明在该组实验中,这 4 个特征提取方法的优劣依次为本文方法、CHI、MI、IG。原因在于:本文方法在选择特征时,不但考查了特征的词性和词频还考查了特征的位置和关联度,从而有效地对待选特征进行全面考查,这使得本文方法受类别分布影 响较小,因此所选特征集较具代表性。CHI 方法在选择特征时不但考查了特征在文档中存在的情况而且还考查了特征不在文档中的情况,MI 方法仅考查了特征在文档中存在的情况,但它们都没能有效地消除冗余特征。因此,这 2 个方法要劣于本文方法,但是 CHI 方法要优于 MI 方法;由于实验中所用语料库中各类别样本分布相差较大,而 IG 方法对类别样本分布极其敏感,因此,在此情况下 IG 方法所选择的特征集代表性最差。

4 结束语

基于统计方法和基于语言学的特征提取方法已

经被广泛应用于特征词提取。本文结合 2 种方法的优点,提出了一种基于遗传算法优化综合启发式的中文网页特征提取方法。该方法能够有效利用词语的内在属性和词语之间的链接关系,通过多种启发式表征中文文本的特征,对特征词进行较全面的考查。实验结果表明该方法能够有效融合不同因素的优点,与传统方法相比,该方法具有一定的优势,从而使得该方法在文本挖掘方面有一定的实用价值。

由于不同类别的文档的因素分布不尽相同,论文接下来的工作将继续研究不同领域内采用该方法的特征词提取的性能。另外通过实验发现,对于人工标注的结果,主观性因素的影响依然存在。论文还将进一步研究合理的标注方式,对现有网页数据进行处理,减少主观因素带来的实验误差。

另外,本文方法虽然采用了十进制编码以及自适应交叉变异操作等措施来确保遗传算法的性能,进而保证本文特征抽取方法的性能,但是目前有些智能优化算法比遗传算法优秀,例如粒子群优化算法、蜂群优化算法等,如果把它们用于本文方法的参数权重优化,效果可能会优于遗传算法。为此,作者下一步研究工作就是尝试把其他智能优化算法用于本文方法的参数权重优化,以进一步提高本文方法的性能。

参考文献:

[1] GHEYAS I A, SMITH L S. Feature subset selection in large dimensionality domains[J]. Pattern Recognition, 2010, 43(1): 5-13.

[2] NGUYEN M H, TORRE F D. Optimal feature selection for support vector machines[J]. Pattern Recognition, 2010, 43(3): 584-591.

[3] ZHAO Zheng, WANG Lei, LIU Huan. On similarity preserving feature selection[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 619-632.

[4] JAVED K, BABRI H A, SAEED M. Feature selection based on class-dependent densities for high-dimensional binary data[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 465-477.

[5] WU Xindong, YU Kui ,DING Wei. Online feature selection with streaming features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(5): 1178-1192.

[6] LEE S, PARK C, KOO J Y. Feature selection in the Laplacian support vector machine[J]. Computational Statistics and Data Analysis, 2011, 55(1): 567-577.

[7] SONG Qinbao, NI Jingjie, WANG Guangtao. A fast clustering-based feature subset selection algorithm for high-dimensional data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 1-14.

[8] CHUANG L Y, YANG C H, LI J C. Chaotic maps based on binary particle swarm optimization for feature selection[J]. Journal of Applied Soft Computing, 2011, 11 (1): 239-248.

[9] 李纲,戴强斌. 基于词汇链的关键词自动标引方法[J]. 图书情报知识, 2011,12(3): 67-71.
LI Gang, DAI Qiangbin. Keywords automatic indexing based on lexical chains[J]. Document, Information and Knowledge, 2011, 12(3): 67-71

[10] 朱颢东, 李红婵. 基于互信息和粗糙集理论的特征选择[J].计算机工程, 2011, 37 (15): 181-183.
ZHU Haodong, LI Hongchan. Feature selection based on mutual information and rough set theory[J]. Computer Engineering, 2011, 37 (15): 181-183.

[11] JEONG Y S, KANG I H, JEONG M K. A new feature selection method for one-class classification problems[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2012, 42(6): 1500-1509.

[12] LIU Z, LIU Q. Balanced feature selection method for Internet traffic classification[J]. Networks, 2012, 1 (2): 74-83.

[13] MAHROOGHY M, YOUNAN N H, ANANTHARAJ V G. On the use of the genetic algorithm filter-based feature selection technique for satellite precipitation estimation[J]. Geoscience and Remote Sensing Letters, 2012, 9(5): 963-967.

作者简介:

沈高峰,男,1978 年生,讲师,主要研究方向为数据库应用、数据挖掘。通过省级成果鉴定 8 项,先后发表学术论文 11 篇,参与编写教材 4 部。