计算机科学
COMPUTER SCIENCE

一种改进的特征选择算法在邮件过滤中的应用

李永红, 汪盈, 李腊全, 赵志强

引用本文

李永红, 汪盈, 李腊全, 赵志强.一种改进的特征选择算法在邮件过滤中的应用[J]. 计算机科学, 2022, 49(11A): 211000028-5.

LI Yong-hong, WANG Ying, LI La-quan, ZHAO Zhi-qiang. Application of Improved Feature Selection Algorithm in Spam Filtering [J]. Computer Science, 2022, 49(11A): 211000028-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000028

一种改进的特征选择算法在邮件过滤中的应用

李永红[1]　汪　盈[1]　李腊全[1]　赵志强[2]
1 重庆邮电大学理学院　重庆 400065
2 重庆邮电大学软件学院　重庆 400065

摘　要　垃圾邮件一般是指未经用户请求强行发到用户电子信箱中的包含宣传资料、病毒等内容的电子邮件,它具有批量发送的特征,且会在互联网上造成巨大危害。因此,为用户过滤掉这些垃圾邮件非常重要。垃圾邮件过滤问题的实质是一个文本分类问题,具有很高的特征维度。但并不是所有特征都对分类有贡献,因此选择一个合适的能够反映整个数据集的特征子集是构造一个好的邮件分类器的基础。现有的特征选择方法存在一定的局限性,比如特征之间仍存在冗余、约简特征结果不稳定,以及计算成本高等。研究和分析现有垃圾邮件处理方法的一些优缺点,结合现有方法,提出一个新的基于信息增益方法和粒度球邻域粗糙集方法的集成特征选择方法,即IGGBNRS算法。通过在不同分类模型上的对比实验表明,该算法简化了模型,性能较好。

关键词:垃圾邮件过滤;特征选择;属性约简;文本分类;IGGBNRS
中图法分类号　TP391

Application of Improved Feature Selection Algorithm in Spam Filtering

LI Yong-hong[1],WANG Ying[1],LI La-quan[1] and ZHAO Zhi-qiang[2]
1 School of Science,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2 School of Software Engineering,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　Spam usually refers to e-mails with promotional materials,viruses and other contents that are forcibly sent to the user's e-mail address without user's request. It has the characteristics of batch sending,and will cause great harm on the Internet. Therefore,it is very important to filter out these spams for users. The essence of the spam filtering problem is a text classification problem,which has a very high features dimension. But not all features contribute to classification,so choosing a suitable subset of features that can reflect the entire data set is the basis for constructing a good email classifier. Existing feature selection methods have some limitations,such as redundancy between features,unstable result of feature reduction and high computational cost. By studying and analyzing some of the advantages and disadvantages of the existing spam processing methods,a new integrated feature selection method based on the information gain method and the granular ball neighborhood rough set method is proposed,named IGBNRS algorithm. Through the experimental comparison on different classification models,the proposed algorithm simplifies the model and has a good performance.

Keywords　Spam filtering,Feature selection,Attribute reduction,Text classification,IGGBNRS

1　引言

电子邮件(Email)是一种通过互联网提供信息交换的通信方式,在互联网中应用广泛。它有文字、图像、声音等多种形式,它的存在极大地方便了人与人之间的交流,促进了社会的发展。但随着当今互联网的飞速发展,电子邮件的发送愈发方便快捷,成本也愈低廉,因此大量虚假、无用、不健康甚至携带病毒链接的邮件也开始泛滥[1]。这些邮件未经用户许可便强行发送到用户邮箱,给用户造成了巨大困扰,被认为是"垃圾邮件"。因此,对用户接受的邮件进行识别和过滤非常必要。

2　研究现状和主要方法

在数据爆炸的今天,处理大规模数据集几乎成为行业常态,然而,这些数据集中的数据并不都是有用的。如邮件中存在大量的垃圾邮件干扰着用户对有用信息的关注,增加了邮件服务器的负担,给互联网用户带了了很大的困扰。大部分的垃圾邮件都带有商业或者其他宣传目的,因此,如何过滤垃圾邮件是一个值得深究的问题。

2.1　特征选择方法

垃圾邮件过滤(Spam Filtering)问题是一个文本分类

基金项目:国家自然科学基金面上项目(2020YFC2003502,61876201,61901074);重庆市自然科学基金面上项目(cstc2020jcyj-msxmX0649);重庆市教委科学技术研究项目(KJQN201900636)
This work was supported by the National Natural Science Foundation of China(2020YFC2003502,61876201,61901074),Natural Science Foundation Project of Chongqing(cstc2020jcyj-msxmX0649) and Science and Technology Research Program of Chongqing Municipal Education Commission(KJQN201900636).
通信作者:李永红(liyh@cqupt.edu.cn)

(Text Categorization)问题,分类任务可以表示为获得这样一个函数:

$$f(x):D\times C\rightarrow\{T,F\} \tag{1}$$

其中,$D=\{d_1,d_2,\cdots,d_{|D|}\}$表示需要进行分类的电子邮件;$C=\{c_1,c_2,\cdots,c_{|C|}\}$表示类别集合;对$\langle d_j,c_i\rangle$来说,$T$值表示文档$d_j$属于类$c_i$,$F$值表示文档$d_j$不属于类$c_i$。一般电子邮件分为合法邮件和垃圾邮件,是一个二分类任务,则$C=\{L,S\}$。因此,所做任务就是要找到一个有效的映射函数,准确地实现域$D\times C$到T值或F值的映射,这个映射函数就是决策函数,即分类器。

由于不是所有的特征项都对分类有贡献,一方面,特征之间存在很多不相关(Irrelevant)或冗余(Redundant)的特征;另一方面,特征之间也可能相互依赖,不仅拖慢了模型的运行速度,而且导致模型非常复杂,推广能力低下。因此,降低特征属性维度是解决垃圾邮件过滤问题的关键步骤。在机器学习(Machine Learning,ML)中,通常使用特征选择(Feature Selection,FS)[2]来进行降维。FS也叫属性选择(Attribute Selection,AS),是指从已有的M个特征中选择N个特征使得系统的特定指标最优化,即从原始特征中选择最大程度上能够代替总体的最少指标(属性)。FS能剔除不相关或冗余的特征,从而达到减少特征个数、提高模型精确度以及减少运行时间的目的。在粗糙集理论(Rough Set Theory,RST)中,降低数据集特征维度的方法表述为属性约简(Attribute Reduction,AR)。粗糙集理论在处理不确定信息和模糊信息方面有着先天性优势,它的基本思想是通过对知识进行分类,在保证原数据集分类能力不变的前提下,进行属性约简,删除多余信息,最后总结出既定的规则[3]。

2.1.1 基于统计的特征选择方法

特征选择的目的是从原始的特征词集中选取一个特征子集,将信息量小或不重要的特征词剔除,从而减少特征项的个数,提高模型分类的效率。常用的特征选择方法是根据某种特征评估函数计算各个特征词的评分值,然后按评分值排序,选取评分值高的若干个特征词作为特征向量。目前常用的特征选择方法有文档频率(Document Frequency,DF)[4]、互信息(Mutual Information,MI)[5]、信息增益(Information Gain,IG)[6]等。

IG方法是包含信息的度量,是一种比较好的特征选择方法。在IG方法中,首先统计一个特征词在每个类别的所有文档中出现和未出现的文档数,然后采用式(2)计算每个特征词的权重:

$$G(t)=-\sum_{i=1}^{m}Pr(c_i)\log Pr(c_i)+Pr(t)\sum_{i=1}^{m}Pr(c_i|t)\log Pr(c_i|t)+Pr(\bar{t})\sum_{i=1}^{m}Pr(c_i|\bar{t})\log Pr(c_i|\bar{t}) \tag{2}$$

其中,$Pr(c_i)$为一篇文档在第i个类别中出现的概率;$Pr(t)$为特征词t在一篇文档中出现的概率;$Pr(\bar{t})$为特征词t不在一篇文档中出现的概率;$Pr(c_i|t)$为特征词t在第i个类别中出现的概率;$Pr(c_i|\bar{t})$为特征词t不在第i个类别中出现的概率;m为类别个数。所有的特征词将按照其计算所得的权值$G(t)$来排序,权值大的特征词被保留的可能性大。

但IG方法有一个缺点,它只是根据特定的信息增益阈值进行筛选,没有考虑特征之间的冗余。为了能够有效地消除特征之间的冗余,Peng等[7]提出了消除冗余性的MRMR,但其巨大的时间复杂度很难将其应用于文本分类中[8]。Lee等[9]提出了改进信息增益特征选择算法,Uysal等提出了基于特征概率性的选择方法——区别性特征选择算法(DFS)[10]。虽然这些算法能够有效地去除冗余,但都具有很高的复杂度,不能快速地进行特征选择。

2.1.2 基于粗糙集的属性约简算法

粗糙集理论(Rough Set Theory,RST)是由波兰数学家Pawlak于1982年初首次提出的[11],粗糙集是一种扩展性较强的数学模型,其理论本身在处理不确定信息和模糊信息方面有着先天性优势,是一种广泛应用于属性约简的经典方法[12-13]。约简的目的是:1)用尽可能少的属性来表示原数据集的信息;2)减少冗余属性之后可以提高分类器的可推广性。

粗糙集中的属性约简是一个NP-hard问题,大多数粗糙集算法利用数据集的先验领域知识,通过隶属度函数来处理连续属性。邻域粗糙集(Neighborhood Rough Sets,NRS)用邻域的概念代替了隶属度函数[14],使邻域粗糙集能够处理没有先验知识的情况。然而,NRS中每个目标的邻域半径是固定的,其半径的优化依赖于网格搜索。这就降低了效率和有效性,导致时间复杂度不低于$O(n^2)$。针对这些局限性,Xia等基于邻域粗糙集(NRS)和粒球计算分类器(Granular Ball Computing,GBC)[15]提出了一种时间复杂度为$O(n)$的粒球邻域粗糙集(Granular Ball Neighborhood Rough Set,GBNRS)算法[16]。GBNRS将粒度计算转化为邻域粗糙集的思想,它不需要隶属函数,不需要隶属度,也不需要通过网格搜索来优化固定的半径参数,自适应地为每个对象生成不同的邻域。

GBNRS算法可以分为两部分:1)用GBC算法得到生成正域(纯度阈值为1),再利用k-均值聚类全局微调,移出纯度不等于1的粒球之后将剩余粒球组成新的正域;2)根据生成正域判断属性的重要性,若移出某个属性后生成正域不发生变化,则说明该属性与生成正域中的属性相似,是冗余属性,应该被移出,反之,该属性应该被保留。以此来达到属性约简的目的。

该算法不优化任何额外的参数,完全自适应,与标准的NRS方法相比,具有更大的通用性、灵活性,以及更高的分类精度。这是目前最先进的基于粗糙集的属性约简算法之一。但由于生成粒度球时的初始化簇心是随机的,导致最后GBNRS算法每次约简的结果不完全相同,从而导致算法不够稳定。

2.2 垃圾邮件过滤方法

一个传统的分类问题,其本质是构造一个判别函数$f(\boldsymbol{x})$,将连续型变量映射成离散型变量。

经典的垃圾邮件过滤问题是一个二分类问题,假设每封邮件都用一个特征向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$表示,其中$\boldsymbol{x}$为一封邮件的向量空间模型表示,一封邮件中存在$m$个特征。设$L$表示合法邮件类,$S$表示垃圾邮件类,用判别函数$f(\boldsymbol{x})$对数据集进行分类。通常有两种方法:一种是使用判别函数对电子邮件进行排名,排名越靠后的邮件是垃圾邮件的可能性越大;另一种方法是根据$f(\boldsymbol{x})$设置一定数量的阈值将电子邮件分成若干类,供用户进行选择。

基于阈值的分类方法是通过一个阈值$\gamma\in[0,1]$与标准

化 $f(\boldsymbol{x})$进行比较。被阈值划分的两个域(即正域和负域)表述如下：

$$POS_{(\gamma)}(C)=\{x\mid f(\boldsymbol{x})\geqslant\gamma\}$$
$$NEG_{(\gamma)}(C)=\{x\mid f(\boldsymbol{x})<\gamma\} \tag{3}$$

其中，正域 $POS_{(\gamma)}(C)$包含合法的电子邮件，负域 $NEG_{(\gamma)}(C)$包含被拒绝的垃圾邮件。

根据垃圾邮件过滤技术来划分，目前主要有基于规则的过滤和基于统计的过滤两类[17]，如基于规则的决策树(DT)算法[18]、Rough Set 方法[19]等，基于统计的支持向量机(SVM)算法[20]、朴素贝叶斯分类(NBC)算法[21-22]和最近邻(KNN)算法[23]等，这两类技术基本将垃圾邮件过滤问题视为二分类问题，即判断电子邮件为合法邮件(L)或者垃圾邮件(S)[24]。

3 基于信息增益和粒球领域粗糙集的特征选择方法

本文选择 IG 方法作为特征选择的初始方法，但 IG 方法选择的特征之间仍存在一定的冗余，且升高此时的信息增益阈值会降低模型的性能。而粗糙集的属性约简算法可以在不影响最终决策分类结果的情况下，删除其中不相关或不重要的属性，但不适用于过高维度的数据集。由于 GBNRS 的不稳定性，导致了最终分类模型的不稳定。结合两者的优缺点，本文提出基于 IG 方法和 GBNRS 方法的集成特征选择算法，即 IGGBNRS(Information Gain and Granular Ball Neighborhood Rough Set)算法。在用 IG 方法对高维数据集进行初次约简后，再利用 GBNRS 算法二次约简，有效消除属性之间的冗余，而不影响最终的分类性能。

3.1 性能评价指标

电子邮件分类问题本质上是一个二分类问题，即邮件中只存在两种邮件：合法邮件(Legitimate)和垃圾邮件(Spam)。因此，可以用常用的评价指标对模型进行衡量[25]。

(1)约简率(Reduction)

维数约简是机器学习的一种必要手段，约简率是一种可以定义一个特征集合约简效果的指标。若原语料集 D 拥有 n 个指标，通过某种约简算法属性约简后的特征属性降至 m 个，且该 m 个特征属性可以较好地反映整个语料集，则定义约简率：

$$Reduction=\frac{n-m}{n} \tag{4}$$

约简率越高表明该约简算法可以去除更多的冗余属性。但约简率越高并不能表明该约简算法更好，因为该算法可能会剔除重要属性，因此，需要另外的评价指标来对算法性能进行评估，比如综合评价指标(F_1)。

(2)综合评价指标(F_1 值)

F_1 值是查全率(Recall)和查准率(Precision)的综合度量，其计算方式如下：

$$F_1=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{5}$$

其中，$Precision=\dfrac{n_{S\to S}}{n_{S\to S}+n_{S\to L}}$表示垃圾邮件的分类精度，$Recall=\dfrac{n_{S\to S}}{n_{S\to S}+n_{L\to S}}$表示垃圾邮件的召回率，$n_{S\to S}$表示被正确分类为垃圾邮件的数量，$n_{L\to S}$表示被错分的合法邮件数量，$n_{S\to L}$表示被错分的垃圾邮件数量，$n_{L\to L}$表示被正确分类的合法邮件数量。

在二分类模型中，F_1 值可以很好地衡量一个分类模型的性能，但不适用于多分类模型。

3.2 算法设计

对于语料集 $D=\{d_1,d_2,\cdots,d_n\}$，包含 n 封电子邮件以及特征集 $T=\{t_1,t_2,\cdots,t_M\}$，含有 M 个特征项。IGGBNRS 算法分为两部分：1)通过 IG 算法对特征集 T 进行初次约简；2)通过 GBNRS 算法再次约简。最终的约简结果是该特征集合 T 的约简子集。IGGBNRS 的具体算法设计如算法 1 所示。

算法 1 IGGBNRS

输入：数据集 D(包含特征集 $T=\{t_1,t_2,\cdots,t_M\}$)，IG 阈值 th_{IG}，GBNRS 算法

输出：一个特征子集 T′

1. T′初始化为 T，这里 $T'=\{t_1,t_2,\cdots,t_M\}$
2. FOR i=1,2,…,M DO
3. 通过式(2)计算 t_i 的 IG 值 g_i
4. IF $g_i<th_{IG}$ THEN
5. 从 T′移除 t_i
6. ELSE
7. 保留 T′中的 t_i
8. END IF
9. END FOR
10. T′更新为 $T'=\{t_1',t_2',\cdots,t_{M_1}'\}$，其中 $M_1\leqslant M$，$\{t_1',t_2',\cdots,t_{M_1}'\}\subseteq\{t_1,t_2,\cdots,t_M\}$
11. GBNRS 算法约简 T′，得到特征子集 $T_1'=\{t_1'',t_2'',\cdots,t_{M_2}''\}$
12. 再次更新 T′为 T_1'，其中 $M_2\leqslant M_1$，$\{t_1'',t_2'',\cdots,t_{M_2}''\}\subseteq\{t_1',t_2',\cdots,t_{M_1}'\}\subseteq\{t_1,t_2,\cdots,t_M\}$

IGGBNRS 算法需要 GBNRS 算法进行属性约简，由于 GBNRS 算法在利用 2-means 选择初始簇心时会随机选择，因此算法仍然不够稳定，故该算法约简一次的结果不一定是最佳的属性子集。因此本文将对 IGGBNRS 算法实验 10 次，计算每一次的模型分类性能，选择性能最好的那一组属性子集作为 IGGBNRS 算法最终的属性子集。

3.3 时间复杂度

在 IGGBNRS 算法的第一步中，IG 方法会计算所有样本中每个特征的信息增益值，时间复杂度为 $O(Mn)$，其中参数 n 为样本数，M 为特征数量，一般看作常数。在第二步中，主要特征选择算法依赖于粒度球邻域粗糙集(GBNRS)算法，该算法的时间复杂度在于粒球的生成，即 k-means 算法，其时间复杂度为 $O(nkt)$，其中参数 k 表示聚类次数，t 代表迭代次数。由于电子邮件分类为二分类问题，因此使用 2-means 聚类方法，它具有较快的收敛速度，可近似看成线性算法，故时间复杂度接近 $O(n)$。因此，IGGBNRS 算法的时间复杂度为 $O(n)$。由于该算法继承了 GBNRS 算法鲁棒性好的优点，因此，低的复杂度并不会降低模型的精度。

4 实验与评估

4.1 数据集准备

本文采用 UCI 数据集中专门用于二分类的垃圾邮件数据库(Spambase，SB)进行实验[1)]。该语料集包含 1813 封合法

1) http://archive.ics.uci.edu/ml/datasets/Spambase

邮件和2788封垃圾邮件，以及57个特征。把语料集分为10个部分，采用10折交叉验证，轮流将其中9份作为训练数据，1份作为测试数据，进行实验。这里选择综合评价指标 F_1 值作为模型的性能评价指标，最终的模型性能取10次实验结果的平均值。

由于并不是所有特征都具有良好的分类能力，通过IG方法进行初步筛选，选择保留模型性能开始明显下降前的特征个数。由于不同的分类模型对同一数据集会表现出不同的分类性能，因此本文采用8种不同的分类模型：BP神经网络(BP)、K近邻(KNN)、朴素贝叶斯分类(NBC)、支持向量机(SVM)、分类决策树(CART)、Logistics回归(LR)、梯度提升决策树(GBDT)以及其改进算法(XGBoost)对数据集进行分类，在不断移除信息量少的属性过程中，以 F_1 值来评判模型的分类性能。

由图1可以看出，在剩余39个特征时，NBC模型性能开始出现明显下降，其他模型在剩余15个属性时性能开始下降。由于在剩余39个属性时大部分模型的分类性能只略高于80%，性能不够理想，因此保留信息增益值最大的39个特征而不是保留15个特征。

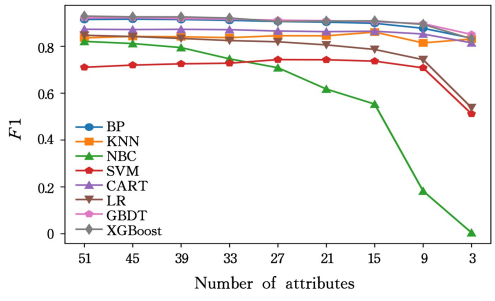

图1 SB语料集中所有特征的分类能力

Fig. 1 Classification ability of all features in SB corpus

本文利用IG，GBNRS，IGGBNRS这3种算法筛选合适的特征子集，并用不同的分类算法对由筛选后的特征子集组成的数据集进行分类，比较其性能。

4.2 实验

SB语料集中包含57个特征，4601条电子邮件。用IG，GBNRS，IGGBNRS这3种不同的属性约简算法来约简属性。这里在使用IG方法时保留39个属性，GBNRS方法对57个特征进行约简；而IGGBNRS算法将对用IG方法方法选择的39个属性再次使用GBNRS算法进行二次约简。选择BP，KNN，NB，SVM，CART，LR，GBDT，XGBoost这8种不同的分类器，此时SB语料集的约简率和 F_1 值如表1、表2所列。

表1 SB语料集在3种约简算法下模型的约简率

Table 1 Reduction of model under 3 reduction algorithms

(单位：%)

分类器	约简		
	IG	GBNRS	IGGBNRS
BP	31.58	22.81	38.60
KNN	31.58	29.82	43.86
NB	31.58	19.30	38.60
SVM	31.58	33.33	47.37
CART	31.58	19.30	42.11
LR	31.58	22.81	42.11
GBDT	31.58	22.81	40.35
XGBoost	31.58	22.81	42.11

表2 SB语料集在3种约简算法下模型的 F_1 值

Table 2 F_1 value of model under 3 reduction algorithms

(单位：%)

分类器	约简		
	IG	GBNRS	IGGBNRS
BP	91.61	92.16	91.89
KNN	84.07	85.46	84.65
NB	79.46	81.89	78.73
SVM	72.52	73.52	74.61
CART	87.19	88.50	87.92
LR	83.37	84.88	83.71
GBDT	91.84	92.39	92.16
XGBoost	92.55	92.76	92.32

由上面的实验结果可以得到以下结论。

(1)约简率

IGGBNRS算法的约简率明显高于IG和GBNRS这两种算法，因为该算法是经IG算法对属性集进行初次约简后，利用GBNRS算法再次约简后的结果。该方法在IG方法上进行优化，剔除了剩余属性之间的冗余属性，而不影响模型的性能，简化了模型。

(2) 分类性能

对于 F_1 值，相比GBNRS算法，IGGBNRS的分类性能出现小幅度下降，这是因为在用IG方法进行约简时，剔除了一小部分可能包含较多信息的属性，但就损失的性能来看，剔除的该部分属性造成的损失是可以接受的；相比IG算法，IGGBNRS是对IG方法的优化，在提高了约简率的同时，模型性能不仅没有降低，反而得到了提升，达到了我们需要的效果。

因此IGGBNRS算法可以说是优于IG和GBNRS这两种算法，该算法使模型性能得到了一定的提升。

结束语 本文基于目前最先进的属性约简算法之一GBNRS算法，以及电子邮件分类问题提出了一种新的解决方法——IGGBNRS算法，在不影响模型分类能力的前提下，成功选择一个数量更少的特征子集，简化了模型，使模型的使用范围更广。但由于该算法仍未解决GBNRS算法约简属性不稳定的问题，在往后的工作中，将对IGGBNRS算法进行优化，并结合现有的分类模型，构造出一个稳定的、更具鲁棒性和说服力的垃圾邮件过滤模型。

参 考 文 献

[1] BHOWMICK A, HAZARIKA S M. E-Mail Spam Filtering: A Review of Techniques and Trends[M]. Singapore: Springer, 2018.

[2] GUYON I M, ELISSEEFF R. An Introduction to Variable and Feature Selection[J]. The Journal of Machine Learning Research, 2003, 38(3): 1157-1182.

[3] LI H M, WANG J Y. Research on knowledge discovery based on knowledge dependency reduction[J]. Software Guide, 2015, 14(6): 135-137.

[4] AZAM N, YAO J. Comparison of Term Frequency and Document Frequency based Feature Selection Metrics in Text Categorization[J]. Expert Systems with Applications, 2012, 39(5): 4760-4768.

[5] YANG Y. A Comparative Study on Feature Selection in Text Categorization[C]// Proceedings of International Conference on Machine Learning. 1997.

[6] ZHAI J C, QIN Y P, CHE W W. Improvement of Information

Gain in Spam Filtering[J]. Computer Science,2014,41(6):214-216.

[7] PENG H,LONG F,DING C. Feature Selection based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy [J]. IEEE Transactions on Pattern Analysis/Machine Intelligence,2005,27(8):1226-1238.

[8] SHANG C,LI M,PENG S,et al. Feature selection via maximizing global information gain for text classification[J]. Knowledge-Based Systems,2013,54(4):298-309.

[9] LEE C,LEE G G. Information Gain and Divergence-based Feature Selection for Machine Learning-based Text Categorization [J]. Information Processing/Management, 2006, 42(1): 155-165.

[10] UYSAL A K,GUNAL S. A Novel Probabilistic Feature Selection Method for Text Classification [J]. Knowledge-Based Systems,2012,36(13):226-235.

[11] PAWLAK Z. Rough sets[J]. International Journal of Computer/Information Sciences,1982,11(5):341-356.

[12] LI Y,FAN B,GUO J,et al. Attribute Reduction Method Based on k-prototypes Clustering and Rough Sets[J]. Computer Science,2021,48(6A):342-348.

[13] YANG Y,CHEN D,HUI W. Incremental Perspective for Feature Selection Based on Fuzzy Rough Sets [J]. IEEE Transactions on Fuzzy Systems,2018,26(3):1257-1273.

[14] HU Q,ZHANG L,ZHOU Y,et al. Large-Scale Multimodality Attribute Reduction with Multi-Kernel Fuzzy Rough Sets[J]. IEEE Transactions on Fuzzy Systems,2018,26(1):226-238.

[15] XIA S,LIU Y,DING X,et al. Granular Ball Computing Classifiers for Efficient,Scalable and Robust Learning [J]. Information Ences,2019,483(10):136-252.

[16] XIA S,ZHANG Z,LI W,et al. GBNRS:A Novel Rough Set Algorithm for Fast Adaptive Attribute Reduction in Classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2022,34(3):1231-1241.

[17] CHEN Z X Survey on Spam Filtering Technology[J]. Application Research of Computers,2009,26(5):1612-1615.

[18] SUBASI A,ALZAHRANI S,ALJUHANI A,et al. Comparison of Decision Tree Algorithms for Spam E-mail Filtering[C]// 2018 1st International Conference on Computer Applications/Information Security(ICCAIS). 2018.

[19] LIU Y,DU X P,ZHOU S,et al. Intelligent Analysis and Filtering of "Spam" and Discussion on Rough Sets [C]// Network and Data Communication Academic Conference of China Computer Federation. China Computer Federation,2022.

[20] DRUCKER H,WU D,VAPNIK V N. Support Vector Machines for Spam Categorization [J]. IEEE Transactions on Neural Networks,2002,10(5):1048-1054.

[21] WANG Q S,WEI R Y. Bayesian Chinese Spam Filtering Method Based on Phrases[J]. Computer Science,2016,43(4):256-259,269.

[22] WANG L,LI Z W,ZHU C D,et al. Research on spam filtering based on NB algorithm[J]. Transducer and Microsystem Technologies,2020,39(9):46-48,52.

[23] DONG M G,HUANG Y Y,JING C. K-Nearest Neighbor Classification Training Set Optimization Method Based on Genetic Instance and Feature Selection [J]. Computer Science, 2020, 47(8):178-184.

[24] BO Y,XU Z B. A Comparative Study for Content-based Dynamic Spam Classification Using Four Machine Learning Algorithms[J]. Knowledge-Based Systems,2008,21(4):355-362.

[25] ZHOU Z H. Machine Learning[M]. Beijing:Tsinghua University Press,2016.

LI Yong-hong, born in 1970, B.S, professor. His main research interests include combinatorial optimization, fuzzy matroid and data processing.