基于密度与距离的钓鱼邮件检测方法

王秀娟, 张晨曦, 唐昊阳, 陶元睿
(北京工业大学信息学部, 北京 100124)

摘 要: 针对钓鱼邮件检测过程中提取特征数量愈加庞大,检测效果没有明显提升且时间成本增加这一问题,提出了一种钓鱼邮件检测方法. 该方法提出将原始的42维邮件特征转换为2个新特征,即基于密度的特征和基于距离的特征,检测准确率最高可达99.74%,分类时间仅需3.39 s,是传统算法的1/20. 实验结果表明,该方法具有较好的检测效果,并且降低了时间成本.

关键词: 机器学习; 钓鱼邮件; 特征提取; 维度缩减; 支持向量机
中图分类号: TP 393.098 **文献标志码**: A **文章编号**: 0254 - 0037(2019)06 - 0546 - 08
doi: 10.11936/bjutxb2017110027

Phishing E-mail Detection Method Based on Density and Distance

WANG Xiujuan, ZHANG Chenxi, TANG Haoyang, TAO Yuanrui
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Phishing E-mail detection methods are mostly focused on the extraction of different E-mail features, which lead the time increasing. To solve this problem, a method based on density and distance was proposed. The method replaces the 42 original mail features with 2 new ones, i.e., features based on density and distance. Then the machine learning classification algorithm was used to detect phishing E-mail. The detection accuracy of the proposed method reaches 99.74%, and time is only 3.39 s, which is 1/20 of the traditional algorithm. Results show that the algorithm has a better detection performance and saves much time.

Key words: machine learning; phishing E-mail; feature extraction; dimensionality reduction; support vector machine (SVM)

随着网络技术的发展,电子邮件成为人们日常沟通交流必不可少的一种通信方式. 据互联网数据中心(Internet Data Center, IDC)统计,到2016年,全球约有32亿人在使用电子邮件[1]. 但是,电子邮件使用率的大量增长也给人们带来了各种各样的安全隐患,例如钓鱼邮件. 钓鱼邮件是指利用伪装的电子邮件,欺骗收件人将账号、口令等敏感信息回复给指定的接收者,或引导收件人链接到特定的网页[2]. 这些网页通常会伪装成真实网站,从原来单一地仿冒淘宝等电子商务网站,到仿冒中国工商银行等银行网站,再到证券、票务、团购、网游等网站,令登录者信以为真,导致在网页上输入的银行卡号码、账户名称及密码等信息被盗[3].

据中国反钓鱼联盟统计,截至2016年11月累计认定并处理钓鱼网站385 169个,受害人数高达4 411万,损失高达200亿元[4]. 因此,钓鱼邮件检测技术的研究刻不容缓.

近年来,钓鱼邮件的形式发展迅猛,相应的钓鱼邮件检测技术也在不断地更新. 然而,目前大部分的相关文献都致力于增加邮件中的各种特征. 已有

收稿日期: 2017-11-13
基金项目: 国家重点研发计划资助项目(2017YFB0802703);国家自然科学基金资助项目(61602052)
作者简介: 王秀娟(1979—), 女, 讲师, 主要从事信号与信息处理、物联网、网络安全方面的研究, E-mail:xjwang@bjut.edu.cn

文献涉及的特征值已经达到上百种，将原始数据表示为简单且具有较大关联性的高维特征[5]，增加了分类器的训练集测试负担.

因此，本文提出一种用于钓鱼邮件检测的新的特征表示方法. 新的特征是使用基于密度的度量和基于距离的度量来表示原始邮件特征. 本文将这种钓鱼邮件检测方法命名为基于密度与距离的钓鱼邮件检测方法（phishing detection method based on density and distance，PDMBD）.

1　相关工作

1.1　钓鱼检测技术

现有的钓鱼检测技术主要有 3 类：基于黑白名单的检测技术、基于启发式规则的检测技术、基于机器学习的检测技术[6].

用基于黑白名单的检测技术维护一份黑名单列表，在该列表中记录已确认为钓鱼网站的统一资源定位符（uniform resource locator，URL）、互联网协议（Internet protocol，IP）地址或者关键词等信息. 人们可以通过黑名单准确识别钓鱼网站[7]. 这项技术简单方便，但是具有易引起漏判、更新时效低等情况[8].

基于启发式规则的检测技术[9]根据钓鱼网站之间存在的相似性设计启发式规则，指导钓鱼网站检测. 这种技术能够克服黑白名单检测技术的高漏判率问题[9]，但是对于大规模数据则存在误报率高和更新规则难的缺点[8].

基于机器学习的检测技术将钓鱼网站或者钓鱼邮件看作文本，使用文本分类或者聚类的方法进行检测[8]. 目前，大部分文献中使用到的钓鱼邮件识别技术都是通过增加、删除邮件中提取出来的特征，有效地检测钓鱼邮件. 2006 年，Fette 等[10]提出使用 10 种针对钓鱼邮件的特征，利用多种分类器进行训练和测试. Khonji 等[11]将特征的数量增加到 47 个. Iqbal 等[12]提取了 419 个邮件特征. 这些方法实现了较高的检测准确率，但牺牲了计算效率.

1.2　维度缩减技术

随着邮件特征维度的不断上升，邮件检测的准确率并没有得到更大的提升，反而检测时间开始下降. 因此，研究者在检测钓鱼邮件方法中引入了维度缩减技术以解决上述问题.

机器学习中的维度缩减技术主要分为两大类：特征选择和特征提取.

特征选择是从特征集合中挑选一组最具统计意义的特征，即特征选择后的特征集合是原有特征的一个子集. 常用的方法包括基于信息增益、卡方选择等[13].

特征提取将原始特征转换为一组具有明显物理意义或者统计意义的特征，即特征抽取后的新特征能够精确地表示样本信息，尽可能少丢失原有样本信息. 常用的方法有：主成分分析（principal component analysis，PCA）、线性判别分析（linear discriminant analysis，LDA）、独立成分分析（independent component analysis，ICA）等[13].

然而，研究者使用的维度缩减方法并不局限于上述提到的常用方法. 很多研究者提出了自己的特征表示方法，找出数量更少、更具有代表性的特征向量，这些向量能够更加精准地表示原始数据，从而实现维度缩减. 例如 Tsai 等[14]提出了一种基于一个三角形区域的特征提取方法，该特征向量能够更加有效地表示高维数据.

维度缩减技术在钓鱼邮件检测中的运用非常广泛，Toolan 等[15]使用了信息增益技术对特征进行选择. 但是这种降维方法只能在原始数据上进行选择，被选择的特征并不能完整地保留所有的数据信息. Zareapoor 等[16]使用 PCA 和潜在语义分析（latent semantic analysis，LSA）方法对邮件特征进行降维处理，然后再使用分类器进行分类. 这种降维方法对映射关系的选择要求较高，若选择的映射关系不能有效地对数据信息进行提取，会导致检测结果偏差较大.

本文的方法基于从邮件中提取常规内容特征，挖掘邮件数据分布的结构特征，用密度和距离来重新表征邮件，利用分类器实现钓鱼邮件检测.

2　PDMBD

PDMBD 框图如图 1 所示，本文所提的 PDMBD 算法主要分为 3 个模块：邮件特征提取模块、新特征形成模块和分类模块.

图 1　PDMBD 框图

Fig. 1　PDMBD block diagram

首先从邮件中提取出常规邮件特征，形成一个高维度的特征矩阵. 然后通过本文提出的新的维度缩减方法，将高维度的特征矩阵，转换成基于新特征

表示的低维特征矩阵. 最后调用分类器进行邮件分类,检测出钓鱼邮件.

因此,本文的核心工作是如何将高维特征矩阵替换成低维矩阵,那么寻找新特征进行有效的替换就成为关键问题.

2.1 寻找新特征

高维特征在进行分类检测的时候会造成时间消耗长、准确率下降等问题,因此,需要降低特征的维度. 为了在缩减数据维度的同时,还能够精确地保持原始特征的信息,需要寻找到有效的新特征. 密度和距离这 2 个度量值在数据挖掘的应用上具有较高的使用率. Wang 等[17]使用密度这一数值进行快速聚类. 郑金彬等[18]提出了一种基于密度的分布式聚类算法研究,算法中也使用了密度峰值. Lin 等[19]使用了基于距离的特征进行入侵检测. 马萌[20]提出了一种基于流形距离的聚类算法. 常用的机器学习算法,比如 K 均值聚类方法(K-means clustering algorithm, K-Means)、最近邻算法(K-nearest neighbor,K-NN),都使用到了距离进行聚类或分类. 因此,本文拟采用距离与密度这 2 个数据分布结构特征代表原始数据. 为了验证其可行性,作者分别统计了原始特征表征的邮件样本距离和密度分布,结果如图 2、3 所示.

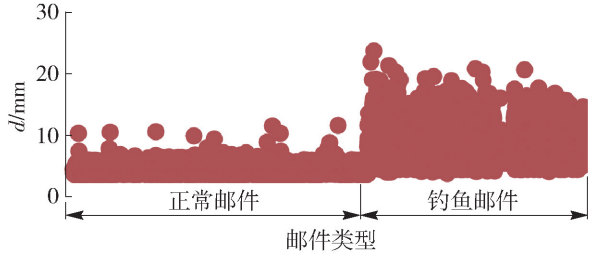

图 2 原始邮件特征距离分布散点图

Fig. 2 Scatter plot of original E-mail features distance distribution

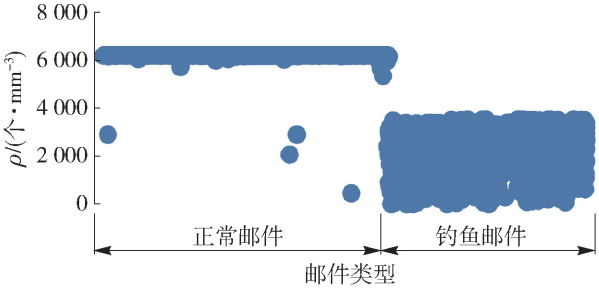

图 3 原始邮件特征密度分布散点图

Fig. 3 Scatter plot of original E-mail features density distribution

从图中可以看出,正常邮件的距离值集中在 5 左右,密度值集中在 6 000 以上. 而钓鱼邮件的距离值主要分布在 5 ~ 15,密度值主要分布在 0 ~ 3 500,表明基于距离的特征和基于密度的特征对于原始邮件有较强的区别力,且保留了较完整的信息. 这 2 个特征可以精确地替代原始数据,以达到维度缩减的目的.

本文为了进一步验证新特征的有效性,在实验部分与另外 2 种降维的方法进行了对比,即用距离特征代替原始特征和信息增益.

2.2 提取邮件特征

目前,有大量的相关文献提取了各种邮件特征进行钓鱼邮件的分类检测. 例如 Toolan 等[15]提取了 40 种邮件特征,其中包含 9 个基于邮件正文的特征,8 个基于主题的特征,13 个基于 URL 的特征,6 个 JavaScript 特征,4 个发件人特征. 比较经典的有文献[10],Fette 等提取了 10 个邮件特征,这 10 个邮件特征是钓鱼邮件检测常用的基础特征. Khonji 等[11]提取了 47 个邮件特征,除了与文献[15]类似的基于邮件正文的特征、主题的特征、URL 的特征、JavaScript 特征、发件人特征等以外,还加入了 2 类外部特征,即垃圾邮件过滤器的标记和打分机制.

本文综合了文献[10-11,15],提取出使用率较高的 42 维特征,如表 1 所示. 从邮件中提取出这 42 维原始特征,较为完整地表现出了邮件的各层信息.

表 1 特征向量说明

Table 1 Feature vector instructions

特征类别	特征名称	解释
基于主题的特征	subject has keyword bank[15]	二进制,主题中含有关键词 bank,value 为 1
	subject has keyword debit[15]	二进制,主题中含有关键词 debit,value 为 1
	subject has keyword FW[15]	二进制,主题中含有关键词 FW(转发邮件),value 为 1
	subject has keyword RE[15]	二进制,主题中含有关键词 RE(回复邮件),value 为 1
	subject has keyword verify[15]	二进制,主题中含有关键词 verify,value 为 1
	subject word number[11]	数值,主题中词的数量
	subject character number[11]	数值,主题中字符的数量

续表 1

特征类别	特征名称	解释
	subject richness[11]	数值,主题丰富度,词的数量/字符的数量
发件人特征	sender and reply-to addresses[11]	二进制,发件人地址和 reply-to 地址不同,value 为 1
	sender domain is modal domain[11]	二进制,发件人邮箱地址不是使用最频繁的域名,value 为 1
邮件正文特征	suspension[11]	二进制,邮件中含有单词“suspension”,value 为 1
	Dear[11]	二进制,邮件中含有单词“dear”,value 记为 1
	verify your account[11]	二进制,邮件中含有词组“verify your account”,value 为 1
	html emails[10]	二进制,邮件中含有 HTML 内容,value 记为 1
	multiparts[11]	二进制,邮件中含有 Multiput 的 MIME 类型,value 记为 1
	The existence of body forms[11]	二进制,邮件中含有 HTML form,value 为 1
	emial richness[11]	数值,邮件正文丰富度
	body_noCharacter	
	account number[11]	数值,邮件中含有 account 单词的个数
	suspended number[11]	数值,邮件中含有 suspended 单词的个数
	unique_word_no[15]	数值,不重复的单词个数
URL 特征	Number of dots[10]	数值,域名中包含 dot 的个数
	Here links to non-model domain	二进制,使用最频繁的链接域名中含有“click、here、link”,value 为 1
	Age if linked_to domain[10]	域名注册时间(取最小值)
	Number of external[11]	数值,external 链接的个数
	the existence of port number[11]	二进制,任意 URL 中含有端口号,value 为 1

续表 1

特征类别	特征名称	解释
URL 特征	Number of port[11]	数值,包含端口号的 URL 个数
	Number of @[11]	数值,链接中出现“@”的个数
	Number of IP address[10]	数值,包含 IP 地址的 URL 的个数
	IP_based URLs[10]	二进制,URL 使用 IP 地址
	Number of domains[10]	数值,连接使用域名数量
	Number of links[10]	数值,链接数量
	Nonmatching URLs[10]	二进制,超链接与 link text 存在差异;link text:文本描述、图片描述等
	url click[11]	二进制,任意文本连接中含有 click,value 为 1
	url here[11]	二进制,任意文本连接中含有 here,value 为 1
	internal link number[11]	数值,internal 链接的个数;internal 链接:指可以在电子邮件中指向邮件内部资源的链接
	linked image number[11]	数值,邮件中图片链接的个数;图片链接:需要点击图片进行跳转
JavaScript 特征	onClick JavaScript events[11]	二进制,邮件中含有“onClick” JavaScript 事件,value 为 1
	javascript from an unmodal domain[11]	二进制,邮件中有 Javascript 加载不是常见域名的外部网页,value 为 1
	change status[11]	二进制,邮件中含有 JavaScript 代码改变状态栏,value 为 1
	pop-up window[11]	二进制,邮件中含有 JavaScript 代码打开弹出窗口,value 为 1
	the existence of javascript[15]	二进制,邮件中包含 Javascript,value 为 1

2.3　新特征定义

通过 2.1 可知,密度和距离这 2 个特征可以有效并精确地将 42 维原始特征缩减成 2 维. 因此,用基于密度的特征 ρ_i 和基于距离的特征 D_i 组成的二

维特征(ρ_i,D_i)代替原始邮件特征.

2.3.1 基于距离的特征定义

聚类可以发现数据分布的特性,相关文献表明,聚类中心和最近邻居点可以有效表征数据点的分布信息. Wang 等[21]使用数据点与最近邻居点的距离和数据点与聚类中心的距离 2 个值进行了入侵检测,检测效果达到了 97.04%,超过了支持向量机(support vector machine,SVM)和 K-NN 的检测率. 因此,本算法采用文献[21]中距离特征的定义方法.

对数据点 A 基于距离的特征的定义为

$$D_A = d_{1A} + d_{2A} \tag{1}$$

式中 d_{1A}表示数据点 A 到所有聚类中心的距离之和,定义为

$$d_{1A} = d(AC_1) + d(AC_2) + \cdots + d(AC_N) = \sum_{k=1}^{N} d(AC_k) \tag{2}$$

式中 $d(AC_k)$表示 A 点与第 k 个聚类中心 C_k的距离. 聚类中心可使用机器学习中的 K-Means 聚类算法计算得到.

K-Means 聚类算法中的 K 值代表类别数,在算法计算过程中 K 值由使用者根据需要进行定义. 本文中使用 K-Means 聚类算法对邮件数据集进行聚类,找到每个类的聚类中心. 邮件数据集中只包含正常邮件和钓鱼邮件 2 种类型,因此,聚类中心有 2 个,定义 K 等于 2.

d_{2A}表示数据点 A 到其最近邻居点的距离,本文采用欧氏距离计算方法. 则 d_{2A}定义为

$$d_{2A} = d(A, \text{neigh}_{N(A)}) \tag{3}$$

A 的邻居点用 neigh(A)表示,$\text{neigh}_{N(A)}$表示 A'点的最近邻居点,定义公式为

$$\text{neigh}_{N(A)} = \underset{A_i}{\text{argmin}}\, d(A, A_i) \tag{4}$$

即与 A 距离最短的邻居点为 A 的最近邻居点.

2.3.2 基于密度的特征定义

在几何计算中,通常将空间区域中点的数目与区域面积的比值称为空间密度. 对于一个点,它的空间局部密度是一定距离内的空间邻近域中点的个数与该邻近域面积的比值. 为了方便计算局部密度,Wang 等[21]使用了一种全局密度计算方法计算数据点的局部密度特征,即每一个数据点在一定范围内包含数据点的总个数. 具体而言,局部密度定义为

$$\rho_i = \sum_j X(d_{ij} - d_c) \tag{5}$$

式中:d_c为一个截断距离,由人工选择;d_{ij}为 i 点到 j 点的距离. 以数据点 i 为圆心,d_c为半径画一个圆,在圆内的数据点个数即为该点 i 的局部密度,即计算每个点与 i 点的距离,取小于 d_c的数据点个数.

但是该定义方法存在一定缺陷. 图 4 中,◦表示一个聚类 1,· 表示另一个聚类 2,根据定义式(5)计算 i 点的局部密度,$\rho_i = 6$. 若以同样的截断距离 d_c计算 j 点局部密度和 z 点的局部密度,$\rho_j = 0$,$\rho_z = 5$. 通过结果分析,i 点的局部密度与 z 点更加接近,与 j 点差别较大,i 点和 z 点应为同一类别,但实际上却是 i 点和 j 点为同一类别. 因此,这种全局中的局部密度计算方法对类似于 i 点的数据点并不能很好地进行定义.

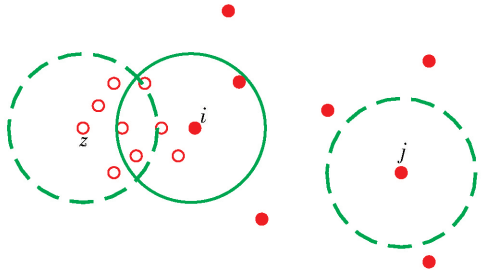

图 4 局部密度计算方法

Fig. 4 Local density calculation method

本文在这种局部密度定义上进行改进,提出另一种局部密度计算方法——聚类中的密度计算方法. 聚类中的密度是指每一个数据点在一定范围内包含的其所属聚类中数据点的个数,定义为

$$\rho_i = \sum_j X(d_{ij} - d_c), i, j \in C_k \tag{6}$$

式中:d_c依然表示一个截断距离,由人工选择;d_{ij}表示 i 点到 j 点的距离,这里 i 点和 j 点属于同一个 K-Means 聚类之后的类别.

2.4 分类

本文使用机器学习中常用的 K-NN 分类器、SVM 分类器、随机森林分类器对 2.3 中得到的二维数据进行分类.

3 实验

3.1 实验设置

3.1.1 数据集选择

本文采用了钓鱼邮件检测领域常用的数据集验证算法的性能,其中钓鱼邮件从 mokey.org 下载. 正常邮件来自于安然邮件数据集(Enron E-mail dataset). 该数据集通过认知学习助手和组织(cognitive assistant that learns and organizes, CALO)

项目收集整理，由约 150 个用户的数据组成一个文件夹，用户主要是安然公司的高级管理人员. 该数据集共包含约 0.5 MB 的消息，由联邦能源管理委员会调查并张贴到网络上. 本文从这 2 个数据集中，提取出 42 维特征数据进行量化归一处理后，分别计算每个样本的 D_i 和 ρ_i 值，得到数据集的最终特征矩阵.

3.1.2　分类过程

本文使用分类算法对二维特征向量(ρ_i，D_i)进行检测. 检测过程使用十折交叉验证方法(10-fold cross-validation)，将数据随机分成 10 组，重复做 10 次实验，每次取 1 组作为测试集，9 组作为训练集. 分类算法选择 *K*-NN 分类器、SVM 分类器和随机森林分类器(random forests，RF).

3.1.3　评价标准

准确率、检测率和虚警率是最常见的性能评价标准. Devaraju 等[22]使用检测率和虚警率评价算法性能. Tsai 等[14]使用准确率和检测率评价算法性能. Wang 等[23]使用准确率、检测率和虚警率评价算法性能. 因此，PDMBD 算法的性能评价使用的评价标准公式为

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{7}$$

$$\mathrm{Det}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{8}$$

$$\mathrm{Fal}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{9}$$

式中：Acc 表示准确率；Det 表示检测率；Fal 表示虚警率；TP 表示钓鱼邮件被正确归类为钓鱼邮件的数量；TN 表示普通邮件被正确归类为普通邮件的数量；FP 表示普通邮件被错误归类为钓鱼邮件的数量；FN 表示钓鱼邮件被错误归类为普通邮件的数量.

3.2　实验结果

在计算密度特征的时候，截断距离 d_c 的选择是人工进行的，d_c 的选择不同，会影响分类结果. 为了选取合适的参数值，在实验中统计点与点之间的距离取值范围，分别选取了 2% 的最大距离值、20% 的最大距离值、40% 的最大距离值、60% 的最大距离值、80% 的最大距离值作为截断距离 d_c 计算局部密度，得到的分类结果如图 5 所示.

从图中可以看出，3 种分类器的准确率都是在截断距离取值为最大距离值 60% 时达到最大值，*K*-NN 的准确率为 99.74%，SVM 的准确率为 94.45%，RF 的准确率为 98.69%，之后逐步下降.

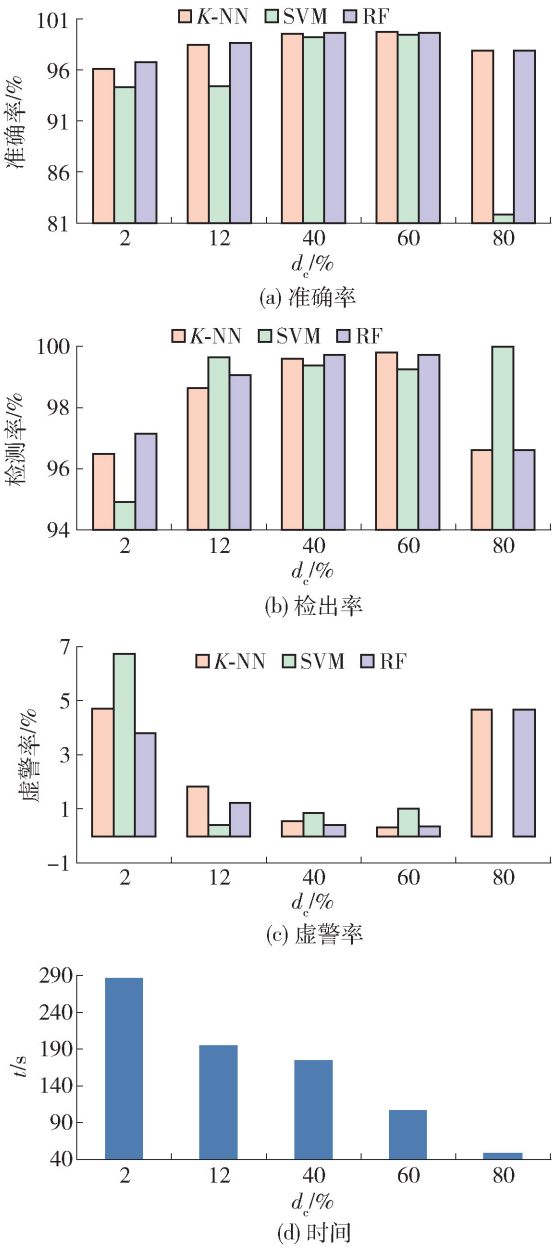

图 5　截断距离取值性能对比

Fig. 5　Different truncated distance performance

而 *K*-NN 和 RF 分类器的检测率和虚警率都与准确率一样，在截断距离取值为最大距离值 60% 时达到最大值. *K*-NN 的检测率为 99.77%，虚警率为 0.31%；RF 的检测率为 99.72%，虚警率为 0.38%. SVM 分类器不稳定，但是也在截断距离取值为最大距离值 60% 时得到相对较高的值，检测率为 99.25%，虚警率为 1.00%. 分类时间随着截断距离取值增大而减少. 因此，截断距离 d_c 取值为最大距离值 60% 时，分类效果最好.

为了验证 PDMBD 的有效性，本文选取了几组对照组进行结果比较. 将本文提出的 PDMBD 作为

实验组,从图 5 分析得知,截断距离选取最大距离值 60% 时效果最佳,因此,在验证 PDMBD 有效性时,使用该截断距离下的分类结果. 对没有经过任何处理的原始 42 维特征矩阵使用分类器进行分类,作为对照组Ⅰ. 为了验证密度特征向量的有效性,只保留 PDMBD 中距离这一维特征进行分类,作为对照组Ⅱ. 最后,使用信息增益方法对原始特征进行信息选择,为了和 PDMBD 进行对比,只采用增益值最大的前两维特征构造降维后的特征矩阵,将该特征矩阵的分类结果作为对照组Ⅲ. 对比结果如图 6 所示.

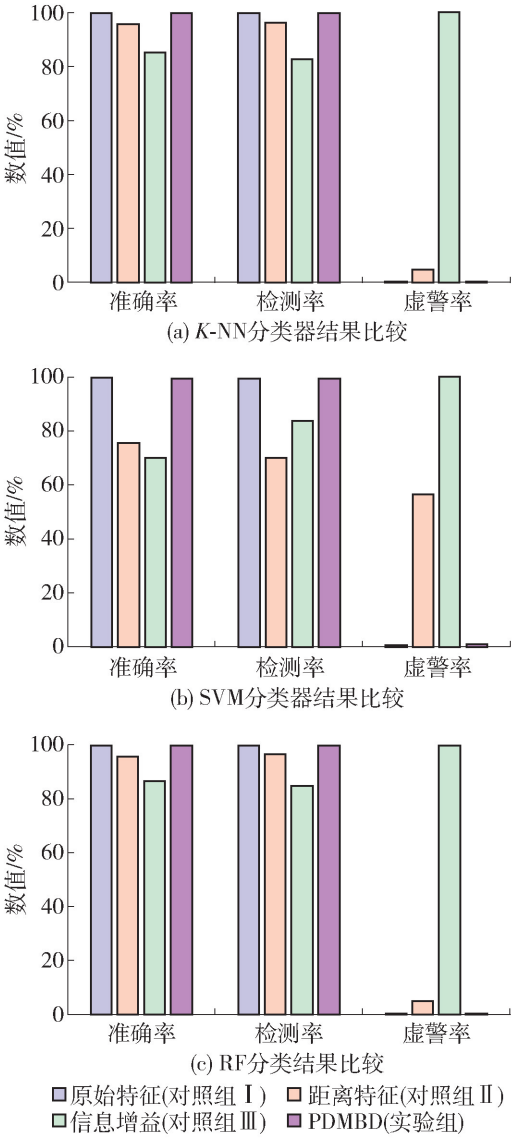

图 6 不同特征分类对比

Fig. 6 Results of different features

从图中可以看出,无论是哪种分类器,PDMBD 的准确率、检测率都要高于距离特征和信息增益特征选择方法,虚警率是 4 种方法中最低的. 虽然原始特征方法的准确率、检测率在 4 种方法中最好,但是 PDMBD 与其相差甚微,甚至在 *K*-NN 分类器中 PDMBD 的准确率和检测率还要优于原始特征. 表 2 列出了几组方案的分类时间对比结果,可以看出,PDMBD 的时间消耗最少,且只有原始特征的 4.3%,因此,PDMBD 极大提高了钓鱼邮件的检测效率.

表 2 不同方法的消耗时间对比

Table 2 Time-consuming comparison among different methods

项目	原始特征方法	距离特征方法	信息增益方法	PDMBD
t/s	78.25	8.53	98.17	3.39
维度缩减比例	1:1	1:42	1:21	1:21

4 结论

1) 本文提出了一种钓鱼邮件检测方法 PDMBD. 该方法首先使用 *K*-Means 算法,找出聚类中心,然后计算每个样本特征向量到聚类中心的距离和最近邻居点距离,从而得到了样本点的距离特征. 此外,本文还定义了一种局部密度作为样本点的密度特征,有效地将原始特征矩阵替代为简单而具有代表性的二维向量.

2) 本文利用 monkey 数据集和安然数据集进行训练和测试,验证了 PDMBD 的性能. 实验结果表明,本文提出的钓鱼邮件检测方法有效地提高了准确率、检测率和虚警率,并且时间性能也比原始特征矩阵有明显提高.

参考文献:

[1] 中文互联网数据资讯中心. IDC: 预测 2016 年全球网民用户数达 32 亿人[R/OL]. [2016-12-22]. http://www.199it.com/archives/422330.html.

[2] CHOWDHURY M U, ABAWAJY J H, KELAREV A V, et al. Multilayer hybrid strategy for phishing email zero-day filtering[J]. Concurrency & Computation Practice & Experience, 2016, 29(23): 623-639.

[3] 杨明, 杜彦辉, 刘晓娟. 网络钓鱼邮件分析系统的设计与实现[J]. 中国人民公安大学学报(自然科学版), 2012(72): 214-226.
YANG M, DU Y H, LIU X J. The design and implementation of phishing email analysis system[J]. Journal of Chinese People's Public Security University (Natural Science Edition), 2012(72): 214-226. (in

Chinese)

[4] 中国反钓鱼联盟. 中国反钓鱼联盟 2016 年 11 月月报[R/OL]. [2016-12-22]. http://www.apac.cn/.

[5] WU L, DU X, WU J. Effective defense schemes for phishing attacks on mobile computing platforms[J]. IEEE Transactions on Vehicular Technology, 2016, 65(8): 6678-6691.

[6] CHOWDHURY M U, ABAWAJY J H, KELAREV A V, et al. Multilayer hybrid strategy for phishing email zero-day filtering[J]. Concurrency & Computation Practice & Experience, 2016, 29(23): 56-74.

[7] PRAKASH P, KUMAR M, KOMPELLA R R, et al. Phishnet: predictive blacklisting to detect phishing attacks [C]// Proceedings of IEEE International Conference on Computer Communications. Washington DC: IEEE Computer Society, 2010: 1-5.

[8] 邹学强，张鹏，黄彩云，等. 基于页面布局相似性的钓鱼网页发现方法[J]. 通信学报，2016(增刊 1)：116-124.
ZOU X Q, ZHANG P, HUANG C Y, et al. Phishing Web page discovery method based on similarity of page layout [J]. Journal of Communication, 2016(Suppl 1): 116-124. (in Chinese)

[9] VARSHNEY G, MISRA M, ATREY P K. A survey and classification of Web phishing detection schemes [J]. Security & Communication Networks, 2016, 9: 6266-6284.

[10] FETTE I, SADEH N, TOMASIC A. Learning to detect phishing emails[C]// International Conference on World Wide Web, WWW 2007. New York: ACM, 2007: 649-656.

[11] KHONJI M, IRAQI Y, JONES A. Enhancing phishing e-mail classifiers: a lexical URL analysis approach [J]. International Journal to Information Security Research, 2013, 3(1): 236-245.

[12] IQBAL F, BINSALLEEH H, FUNG B C M, et al. Mining writeprints from anonymous e-mails for forensic investigation[J]. Digital Investigation, 2010, 7(1/2): 56-64.

[13] 潘锋. 特征提取与特征选择技术研究[D]. 南京：南京航空航天大学，2011.
PAN F. Research on feature extraction and feature selection [D]. Nanjing: Nanjing University of Aeronautics & Astronautics, 2011. (in Chinese)

[14] TSAI C F, LIN C Y. A triangle area based nearest neighbors approach to intrusion detection [J]. Pattern Recognition, 2010, 43(1): 222-229.

[15] TOOLAN F, CARTHY J. Feature selection for spam and phishing detection [C]// Ecrime Researchers Summit (Ecrime). Washington DC: IEEE Computer Society, 2010: 1-12.

[16] ZAREAPOOR M, SHAMSOLMOALI P, ALAM M A. Highly discriminative features for phishing email classification by SVD [J]. Advances in Intelligent Systems & Computing, 2015, 339: 649-656.

[17] WANG S, WANG D, CAOYUAN L I, et al. Clustering by fast search and find of density peaks with data field [J]. Chinese Journal of Electronics, 2016, 25(3): 397-402.

[18] 郑金彬，卓义宝. 基于密度的分布式聚类算法研究[J]. 计算机工程，2008，34(17)：65-67.
ZHENG J B, ZHUO Y B. Density based distributed clustering algorithm [J]. Computer Engineering, 2008, 34(17): 65-67. (in Chinese)

[19] LIN W C, KE S W, TSAI C F. CANN: an intrusion detection system based on combining cluster centers and nearest neighbors[J]. Knowledge-Based Systems, 2015, 78(1): 13-21.

[20] 马萌. 基于流形距离的聚类算法研究及其应用[D]. 西安：西安电子科技大学，2009.
MA M. Research and application of clustering algorithm based on manifold distance [D]. Xi'an: Xi'an Electronic and Science University, 2009. (in Chinese)

[21] WANG X J, ZHAN C X, ZHENG K F. Intrusion detection algorithm based on density, cluster centers, and nearest neighbors[J]. China Communications, 2016, 13(7): 24-31.

[22] DEVARAJU S, RAMAKRISHNAN S. Detection of attacks for IDS using association rule mining algorithm [J]. Iete Journal of Research, 2015, 61(6): 624-633.

[23] WANG F N. Solving the intrusion detection problem with KPCA-RVM [C]// Design, Manufacturing and Mechatronics. Singapore: World Scientific, 2015: 520-527.

(责任编辑　梁　洁)