doi:10.3969/j.issn.1000-436x.2013.z2.001

基于灰度—梯度共生矩阵的图像型垃圾邮件识别方法

冯兵[1,2]，李芝棠[1,2,3]，花广路[1,2]

（1. 华中科技大学 计算机科学与技术学院，湖北 武汉 430074；2. 下一代互联网接入系统国家工程实验室，湖北 武汉 430074；
3. 华中科技大学 网络与计算中心，湖北 武汉 430074）

摘 要：为了逃避基于文本的垃圾邮件系统的检测，越来越多的垃圾邮件制造者将文本信息嵌入到图像中。为了有效地检测出图像型垃圾邮件，提出了一种基于灰度—梯度共生矩阵(GGCM, gray-gradient co-occurrence matrix)的图像型垃圾邮件识别方法。先通过灰度—梯度共生矩阵提取图像的特征信息，然后运用最小二乘支持向量机（LS-SVM, least squares support vector machines）进行分类。实验表明，该方法具有较高的分类精度和较好的实时性。

关键词：图像型垃圾邮件；灰度—梯度共生矩阵；最小二乘支持向量机；纹理特征

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0001-04

Image spam identification method based on gray-gradient co-occurrence matrix

FENG Bing[1,2], LI Zhi-tang[1,2,3], HUA Gang-lu[1,2]

(1.School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;
2. National Engineering Laboratory for Next Generation Internet Access System, Wuhan 430074, China;
3. Network and Computing Center, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: In order to avoid the detection of the spam system based on text, more and more spammers have embedded text information into the image. An image spam identification method based on gray-gradient co-occurrence matrix (GGCM) was proposed to detect image spam effectively. The feature of image was extracted through GGCM firstly, and then LS-SVM was used to do classification. The test results show that this method has higher classification accuracy and better real-time performance.

Key words: image spam; gray-gradient co-occurrence matrix; LS-SVM; texture feature

1 引言

在互联网快速发展的进程中，电子邮件凭其快捷、方便、低成本的优势成为了互联网的关键应用之一，给人们的生活和工作带来了极大的便利。然而，某些别有用心的人利用电子邮件大量散布各种欺诈信息、反动言论、商业广告等垃圾信息。中国互联网协会反垃圾信息中心 2011 年第 4 季度反垃圾邮件状况调查报告表明，中国网民平均每周收到垃圾邮件比例为 35.5%[1]。

针对垃圾邮件的泛滥，学者们提出了许多有效的过滤检测方法，但这些方法大多是基于文本内容的检测。为了逃避传统的过滤检测，垃圾邮件制造者将垃圾信息嵌入到不易于提取分析的图像中，这种垃圾邮件一般被称为图像型垃圾邮件[2]。

目前针对图像型垃圾邮件的过滤技术主要有：基于文本提取（光学字符识别 ORC）、基于图像特征、指纹识别等方法。

光学文字识别(ORC)法[3]是通过分离出图像中的文字区域，然后提取图像中的文本信息，再对文本进行信息检测过滤。干扰因素对 ORC 法识别准确率影响很大，若图像中的文字经过模糊、扭曲处理后，ORC 法就很难识别。

基于图像特征的过滤方法[4~6]是利用图像的各种特征来识别垃圾邮件，如颜色或纹理特征、文本区域特征、边缘特征以及元数据特征。这种方法的

收稿日期：2013-09-06

识别率较高，但时间复杂度也相对较高。

指纹识别技术是通过提取能唯一标识邮件的特征，形成指纹 DNA，并与指纹库中的 DNA 进行对比，从而得到该邮件的垃圾评分等级。其不足之处就是全球性的指纹库非常大，不易维护。

为了逃避过滤，垃圾邮件制造者利用模板和随机化处理技术不断研究出新的图像型垃圾邮件，经过欺骗性处理后，使得一般识别方法的识别率将明显下降。

本文提出一种基于灰度—梯度共生矩阵的图像型垃圾邮件识别方法，利用灰度—梯度共生矩阵提取图片的纹理特征，然后用 LS-SVM 对特征向量进行分类，经过一定量的图像训练后，取得了较好的实验效果。

2 灰度—梯度共生矩阵模型

灰度—梯度共生矩阵[7]模型集中反映了图像中各像素点的灰度和梯度的相互关系，各像素点的灰度是构成一幅图像的基础，梯度则是构成图像边缘轮廓的要素，梯度值大的像素点是边缘的可能性也大，将图像的梯度信息加入到灰度共生矩阵中，构成灰度—梯度共生矩阵，可用灰度和梯度的综合信息表征纹理特征。

2.1 灰度、梯度矩阵及正规化

图像的灰度矩阵表示为 $\boldsymbol{f}(i,j)$，梯度矩阵表示为 $\boldsymbol{g}(i,j)$。本文采用拉普拉斯算子计算灰度图像中各像素点的梯度值，其计算公式如下

$$\boldsymbol{g}(i,j)=4\boldsymbol{f}(i,j)-f(i+1,j)-f(i-1,j)-f(i,j+1)-f(i,j-1) \tag{1}$$

在不影响图像纹理特征的情况下，对灰度矩阵、梯度矩阵进行正规化处理以减少计算量。

对灰度矩阵进行正规化变换：

$$F(i,j)=\mathrm{INT}[\boldsymbol{f}(i,j)\times N_f/f_{\max}]+1 \tag{2}$$

其中，INT 表示取整运算，$f_{\max}$ 为图像中的最大灰度值，N_f 为正规化的最大灰度值，本文取 N_f=64。

对梯度矩阵进行正规化变化

$$G(i,j)=\mathrm{INT}[g(i,j)\times N_g/g_{\max}]+1 \tag{3}$$

其中：INT 表示取整运算，$g_{\max}$ 为图像中的最大梯度值，N_g 为正规化的最大梯度值，本文取 N_g=64。

2.2 灰度−梯度共生矩阵生成及正规化

在正规化后的灰度图像 $F(i,j)$和正规化后的梯度图像 $G(i,j)$中，计算同时使 $F(i,j)=x$ 和 $G(i,j)=y$ 的像素点数，即得到灰度—梯度共生矩阵 $\boldsymbol{H}(x,y)$的第 (x,y)个元素值。

对灰度—梯度共生矩阵进行正规化处理，使其各元素之和为 1。变换公式为

$$\hat{H}(x,y)=\frac{\boldsymbol{H}(x,y)}{\sum_{x=0}^{N_f-1}\sum_{y=0}^{N_g-1}\boldsymbol{H}(x,y)} \tag{4}$$

2.3 特征提取

通过灰度—梯度共生矩阵,可以充分利用图像的灰度和梯度的综合信息提取纹理特征。常用的数字特征参数有 15 种，包括小梯度优势、大梯度优势、灰度分布不均匀性、梯度分布不均匀性、能量、灰度平均、梯度平均、灰度均方差、梯度均方差、相关、灰度熵、梯度熵、混合熵、惯性、逆差矩。本文计算这 15 个特征参数，形成一个 15 维的特征向量。

3 最小二乘支持向量机(LS-SVM)

相对于经典的支持向量机(SVM)而言，LS-SVM[8]采用最小二乘线性系统作为损失函数，并用等式约束替换不等式约束，使求解过程成为了解等式方程组，加快了计算速度。

设 $D=\{(x_k,y_k)|k=1,2,\cdots,N\}$，其中，$x_k\in R^n$ 为输入数据，$y_k\in R$ 为输出类别。在 w 空间中最小二乘支持向量机分类问题可描述为

$$\min_{w,b,e}\phi(a,b,e)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}w+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2 \tag{5}$$

约束条件为

$$y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]=1-e_k,e_k\in R^k,k=1,\cdots,N \tag{6}$$

定义拉格朗日函数

$$L(w,b,e,a)=\phi(w,b,e)-\sum_{k=1}^{N}\alpha_k\{y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]-1+e_k\} \tag{7}$$

其中，拉格朗日乘子 $\alpha_k\in R^n$。根据优化条件

$$\frac{\partial L}{\partial w}=0,\ \frac{\partial L}{\partial b}=0,\ \frac{\partial L}{\partial e_k}=0,\ \frac{\partial L}{\partial a_k}=0 \tag{8}$$

可得到

$$\alpha_k = \gamma e_k,\ \ y_k[\boldsymbol{w}^{\mathrm{T}}\varphi(x_k)+b]-1+e_k=0,$$
$$\sum_{k=1}^{N}\alpha_k y_k = 0,\ \ w=\sum_{k=1}^{N}\alpha_k y_k \varphi(x_k) \tag{9}$$

上式消去 w 和 e，可化为求解如下矩阵方程

$$\begin{bmatrix} 0 & \boldsymbol{Y}^{\mathrm{T}} \\ \boldsymbol{Y} & \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}+\gamma^{-1}\boldsymbol{I} \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ \boldsymbol{I} \end{bmatrix} \tag{10}$$

其中，

$$\boldsymbol{Z}=\left[\varphi(x_1)^{\mathrm{T}}y_1,\cdots,\varphi(x_N)^{\mathrm{T}}y_N\right]^{\mathrm{T}}$$
$$\boldsymbol{I}=[1\cdots 1]^{\mathrm{T}},\ \ \boldsymbol{Y}=[y_1\cdots y_N]^{\mathrm{T}}$$
$$\boldsymbol{\alpha}=[\alpha_1\cdots\alpha_N]^{\mathrm{T}}$$

令核函数为

$$K(x,x_k)=\varphi(x)^{\mathrm{T}}\varphi(x_k) \tag{11}$$

则可得最小二乘支持向量机分类决策函数为

$$y(x)=\mathrm{sgn}\left[\sum_{k=1}^{N}\alpha_k y_k K(x,x_k)+b\right] \tag{12}$$

4　实验及结果

4.1　实验环境

实验图像数据集采用比较有公信力的 Dredze[9] 图像集，其中正常邮件图像集 Personal ham 2 022 幅，垃圾邮件图像集 Personal spam 3 299 幅。从中选取 2 500 幅图像作为训练样本，其中正常邮件图像 1 000 幅，垃圾邮件图像 1 500 幅。另外再选取 1 000 幅图像作为测试样本，其中正常邮件图像 500 幅，垃圾邮件图像 500 幅。

实验在 MATALAB8.0 上进行，LS-SVM 分类器采用 LS-SVMlab Toolbox 1.8，使用径向基核函数。

4.2　实验过程

实验过程如图 1 所示，在特征提取前，若样本图像不是灰度图像，还需转为灰度图像。LS-SVM 分类器有 2 个很重要的参数：正则化参数和径向核宽，它们的取值对分类器的识别精度影响较大，可以通过 LS-SVMlab Toolbox 的 tunelssvm 函数对训练样本进行交叉验证和网格搜索，来得到最优的参数。

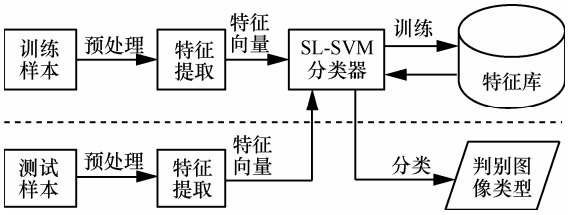

图 1　实验流程

4.3　实验结果

采用召回率（recall）、正确率（precision）、精确度（accuracy）3 个最常用的指标来评价检测性能。召回率是指实际垃圾邮件图像中被正确识别出的垃圾邮件图像的比例。正确率是指被检测为垃圾邮件图像中确实为垃圾邮件图像的比例。精确率是指对所有图像，包括垃圾邮件图像和正常图像的判对率。

本文方法和 OCR 法、颜色矩法、灰度共生矩阵法的识别性能对比如表 1 所示，从中可以看出，本文采用的基于灰度—梯度共生矩阵的方法对图像型垃圾邮件有较高的识别率。

表 1　识别性能对比

分类方法	召回率/%	正确率/%	精确度/%
OCR	74.0	75.1	76.1
颜色矩法[10]	75.0	72.2	73.0
灰度共生矩阵法[11]	72.8	71.7	72.0
本文方法	90.2	88.8	89.4

在现实应用中，垃圾邮件过滤方法的实时性要求比较高。本文方法的实时性测试数据（每幅图像平均处理时间）如表 2 所示，可以看出平均每幅图像的总处理时间在 300 ms 左右，故本文采用的方法实时性也较好。

表 2　实时性测试数据

测试数据	值
特征提取时间/ms	101.63
训练时间/ms	114.24
分类时间/ms	95.57

5　结束语

本文利用灰度—梯度共生矩阵提取图像型邮件的纹理特征，用 LS-SVM 分类器对图像型邮件进行分类，通过实验证明该方法是一个实时性能和识别性能都较好的图像型垃圾邮件过滤方法，具有一定实用价值。

在垃圾邮件过滤系统中，单个特征可能会存在过度拟合的缺点，采用多种特征组合的过滤方式，将会具有更好的分类效果和推广性。

参考文献：

[1]　中国互联网协会反垃圾邮件信息中心. 2011 年第四季度中国反垃圾邮件状况调查报告[EB/OL]. http://www.anti-spam.cn.2012.

The Anti-Spam Center of Chinese Internet Association. The investigation report of the Chinese anti spam in the fourth quarter of 2011[EB/OL]. http://www.anti-spam.cn.

[2] DREDZE M, GEVARYAHU R,ELIAS-BACHRACH A. Learning fast flassifiers for image spam[A].CEAS 2007-Fourth Conference on Email and AntiSpam[C]. Berlin:CEAS, 2007.487-493.

[3] FURMERA G, PILLAI I, ROLI F. Spam Filtering Based on the Analysis of Text Information Embedded into Images[M]. Berlin: Springer, 2006: 2699-2720.

[4] ZUO H Q, HU W M, WU O. Detecting image spam using local invariant features and pyramid match kernel[A]. Proceedings of the 18th international conference on World Wide Web[C]. Madrid, 2009.

[5] NHUNG N P, PHUONG T M. An efficient method for filtering image-based spam[A]. 2007 IEEE International Conference on Research, Innovation and Vision for the Future[C]. 2007. 96-102.

[6] 刘峤, 秦志光. 基于颜色和边缘特征直方图的图像型垃圾邮件分类模型[J]. 计算机应用研究, 2010, 27(7):2608-2609
LIU Q, QIN Z G. Image spam classification model based on color and edge histogram statistics[J]. Application Research of Computers, 2010, 27(7):2608-2609.

[7] HONG J G. Gray level-gradient cooccurrence matrix texture analysis method[J]. Acta Automatica Sinica, 1984, 10(1) : 22-25.

[8] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9:293-300.

[9] DREDZE M, GEVARYAHU R, ELIAS-BACHRACH A. Learning fast classifiers for image spam[A]. CEAS 2007-Fourth Conference on Email and AntiSpam[C]. 2007.

[10] FU Y, WANG Y W, WANG W Q. Content-based natural image classification and retrieval using SVM[J].Chinese Journal of Computers, 2003, 26(10):1261-1265

[11] 肖靓. 基于支持向量机的图像分类研究[D]. 上海: 同济大学, 2006.
XIAO L. Support Vector Machine-Based Image Classification Research[D]. Shanghai: Tongji University, 2006.

作者简介：

冯兵（1986-），男，湖南沅江人，华中科技大学硕士生，主要研究方向为网络与信息安全。

李芝棠（1951-），男，湖北监利人，华中科技大学教授、博士生导师，主要研究方向为计算机系统结构、网络与信息安全、P2P 网络。

花广路（1989-），男，安徽淮南人，华中科技大学硕士生，主要研究方向为网络与信息安全。