[文章编号] 1003－4684(2020)02-0056-05

基于纹理特征和随机森林的恶意代码分类研究

刘宇强，李 军，范志鹏
（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 为高效地识别分析恶意软件，及时防范可能的危害，提出了一种基于图像灰度纹理特征的静态分类方法。根据代码的指令长度特点，设计并提取病毒代码的多字节图像纹理，并统一成二维特征，然后将所有的特征文件作为训练集进行随机森林机器学习方法分类。利用标准数据集进行的实验表明，该方法可以达到 96.36％的精度，并分析了各个字节代码特征的重要性，进一步提出了简化的分类方法。

[关键词] 恶意代码分类；灰度纹理特征；随机森林
[中图分类号] TP309.5 [文献标识码] A

恶意代码检测技术主要分为两类[1-2]：静态的、基于代码程序结构、控制流特征的技术和动态的、基于行为特征的技术。这些技术包括建立签名数据库。主要的限制是，这些技术无法检测到一个新的恶意软件，直到它的签名被更新。动态技术在执行过程中会分析恶意软件样本。检测恶意软件是否类似报告样本的行为。然而，与静态技术相比，动态技术更为精确，因为在恶意软件执行过程中更难掩盖其行为。但风险是检测和识别过程可能已经对用户的工作造成了伤害。

近年来，许多研究人员使用机器学习[3]（Machine Learning，ML）技术动态处理不断变化的恶意软件检测行为。机器学习技术将一个标记的数据集作为训练数据集，并开发一个区分恶意软件和良性样本行为的模型。训练后的模型能够对测试样本进行分类。ML 技术可以通过大量的标记训练数据中学习并提高预测精度。

为了确定恶意代码功能属性并对其进行分类，研究人员探索了许多对恶意代码检测和识别的方法[4-5]，但面对大量使用混淆技术的恶意代码来说，传统的分析方法都存在一定的局限性[6]。为了克服加壳加密技术的影响，将恶意代码进行神经网络训练已成为了恶意代码分类检测的主流趋势。分类过程主要步骤：1）预处理，将恶意代码二进制文件进行数据预处理，构建成为符合分类器的输入模型；2）特征选择，不同的分类器有着不同的特征选择方法，依次选择特征集中影响最大的几个特征项的特征值作为特征子集，从而构建新的特征集；3）分类器训练与分类运算[7]。恶意软件分类的关键是分类模型的选择和训练阶段定义模型的参数。模型确定后，可以用于新数据的分类。这里选择随机森林模型作为分类器，因为它能够有效地处理大型和不平衡的数据集。此外，它可以处理大量的特征，而不会过度拟合。同时，考虑到恶意程序的长度、原理、以及各种技术的应用导致其代码千差万别，直接导致其代码信息很难识别，笔者提出了恶意代码的图像纹理信息作为特征数据，将其二进制信息理解为图像，设计了单字节、双字节和三字节图像纹理，达到提取特征的目的。

1 相关理论

1.1 灰度纹理图像特征

灰度共生矩阵 GLCM（Gray Level Co-Occurrence Matrixes）是研究图像像素的空间相关特性的常用方法。利用灰度纹理特征来表示大规模的图像纹理数据集可以以最小的资源占比来归纳所有的图像，Gotlied 等[8]在研究共生矩阵中研究出的一种归纳特征提取的方法，该方法后被证实对于细微纹理归纳时有良好的效果。Kancherla 等[9]提出用灰度纹理特征来对恶意代码进行分类检测并取得了 95％的准确率，在此之后研究人员逐步开始利用灰度图像来进行恶意代码研究。

通常，GLCM 是像素距离和角度的矩阵函数，它不仅能反映亮度的分布特征，还能描述给定图像

[收稿日期] 2019－11－12
[第一作者] 刘宇强（1994－），男，湖北荆门人，湖北工业大学硕士研究生，研究方向为网络安全
[通信作者] 范志鹏（1994－），男，湖北十堰人，湖北工业大学硕士研究生，研究方向为图像识别

的纹理特征。可以为整个图像计算 GLCM，也可以为像素值周围的小窗口计算 GLCM。虽然给定的图像灰度为 256，但在计算灰度共生矩阵导出的纹理特征时，图像的灰度远小于 256。主要是由于矩阵维数较大，窗口尺寸较小，灰度共生矩阵不能很好地表示纹理，同时计算量大大增加。因此在计算灰度共生矩阵之前，需要对图像进行直方图化处理，以降低图像的灰度值，图像的灰度为 8 或 16。给定图像灰度共生矩阵的构造公式如下：

$$p(g_1,g_2)=\frac{p(g_1,g_2)}{R},$$

$$R=\begin{cases}N=N(N-1)\theta=0^\circ \text{ 或 } 90^\circ\\ (^N-1)2\theta=45^\circ \text{ 或 } 135^\circ\end{cases} \tag{1}$$

式(1)是对图像上保持一定距离的像素点 g_1,g_2 之间的灰度情况进行统计，根据图像中两个不同像素之间的距离为 d，方位关系度数为 θ 的两个像素点构建联合概率分布 $p(g_1,g_2\mid d,\theta)$。将距离 d 的值设置为 1，θ 设置为 0°、45°、90°和 135°

$$p(g_1,g_2)=\frac{p(g_1,g_2)}{R} \tag{2}$$

$$R=\{N(N-1)\theta=0^\circ,90^\circ\ (N-1)^2\theta=45^\circ,135^\circ$$

通常以三个角度的联合统计数据，就能够归纳出原始图像的所有特征，通过选择其中影响最大的几个特征作为特征值，可以在关键信息丢失率最低的情况下进行降维处理，GLCM 算法能够找出其相关性过大的部分进行分割，除了保存关键信息外，也能够很好地剔除掉干扰混淆的部分。

根据上述过程，当角度分别为 0°、45°、90°和 135°时，可以计算出四个 GLCM。计算结果反映了图像的纹理特征，如角二阶矩、熵、逆微分矩、惯性矩和相关性。

例如熵是对图像信息的度量。从熵的值可以看出图像纹理的不均匀程度或复杂程度，且 CLCM 散射元素越多，图像熵的值越大。二维数组数字差异变化越大，表现出的图像越复杂，具体公式为：

$$Ent=\sum_{g1=1}^{k}\sum_{g2=1}^{k}p(g_1,g_2)*\log p(g_1,g_2) \tag{3}$$

其中 k 为灰度图像尺寸大小，通过对图像当中任意像素点 g_1,g_2 构造出的灰度共生矩阵进行统计，计算出 4 个方向上的熵值，将所有方向结果上的值进行求和，可以还原出原始灰度图像的特征图像。

1.2　随机森林 RF(Fandom Forest)分类器

随机森林算法是一种能够对大量数据进行准确分类的新型分类技术[10]。它既可以用于故障的分类，也可以用于故障的回归类型。基于树的学习算法是机器学习和数据科学中应用最广泛的学习方法之一。

由于随机森林分类器建立了多个决策树，并根据这些树的投票结果对最终结果进行评估，从而消除了单决策树方法中存在的过度拟合问题。合并树的过程称为集成方法，从每棵树中对向量进行分类，并将其视为类的投票，然后选择投票最多的分类器作为向量。它是以分治方法为基础的集成模型分类器。一组个体的弱学习者可以通过这个过程共同形成一个强学习者。

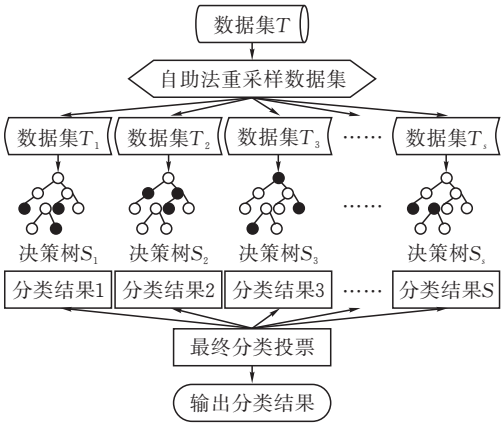


图 1　随机森林整体模型

假设数据集 T 具有 M 个特征，n 个数据。T 表示为 $\boldsymbol{X}_1,\boldsymbol{Y}_1;\boldsymbol{X}_2,\boldsymbol{Y}_2;\cdots;\boldsymbol{X}_n,\boldsymbol{Y}_n$。其中 $\boldsymbol{X}_i=\{A_{i1},A_{i2},\cdots,A_{iM}\}$ 为 M 个特征值创建的第 i 个向量，$\boldsymbol{Y}_i$ 为对应向量的输出类。通过自助法重采样技术将原始数据集 T 有放回的重复抽取 n 个样本，形成新的训练集样本 T_i，新的训练集样本大小与原始训练集样本大小相同，这一步骤重复 S 次形成 S 个数据集：$T_1,T_2,\cdots,T_S$，通常随机森林分类器使用输入数据的 2/3 作为训练集，1/3 作为测试集，这一类数据称为包外数据。对于一组在数据集 T_i 上被选择的向量 $\boldsymbol{X}_i,\boldsymbol{Y}_i$，在进行重构数据集时，可以被重新用来创造新的数据集 Tj，由于随机采样是通过替换完成的，任何向量 $\boldsymbol{X}_i,\boldsymbol{Y}_i$ 都可以被不同的数据集 T_i 选择多次，并且存在一些从未被任何 T_i 选择的向量，这种情况被称为 bagging，它基于引导聚合产生[11]。对于每个数据集 T_i 都会形成一个决策树 S_i，通过决策树对输出向量 $\boldsymbol{V}_i$ 进行分类，最后统计 $V_1,V_2,\cdots,V_s$ 的输出结果，取最大的分类结果来决定 V_i 的类别。

1.3　K-MEANS 聚类分类方法

K-means 聚类是一种基于相似性将数据对象分为 K 个簇的分块聚类方法[12]。在算法中，必须指定集群的数量 K。最初选择 K 个质心。每个数据对象都被分配给包含其最近质心的簇。初始质心的选择是随机的。用欧几里德距离、余弦相似性来衡量与质心和数据对象的接近程度。初始分组完成

后,计算每个簇的新质心以及每个数据点到每个中心的距离。根据距离重新分配数据点。如果该点与簇的所有成员之间的距离之和不能再最小化,则将簇中的点视为质心。K-means 聚类的主要目的是最小化聚类成员与其质心之间的距离之和。

假设数据集 $X_1, X_2, \cdots, X_n$ 中,每一个样本 $\boldsymbol{X}_i$ 均为 d 维实向量,k-means 方法就是将这 n 个样本划分到 k 个集合当中,其中 $k \leqslant n$,同时满足划分后的聚类平方和最小为 K_s ,具体公式为:

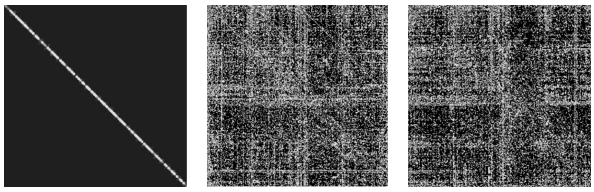
$$K_s = \operatorname{argmin} \sum_{k=1}^{k} \sum_{i=1}^{n} || \boldsymbol{X}_i - u_i ||^2 \tag{4}$$

其中 u_i 为数据集 $X_1, X_2, \cdots, X_n$ 中所有点的平均值。

2 改进的灰度纹理图像特征

恶意软件中的单个操作码与普通代码并无太大差异,而较长的操作码具有预测现象发生的能力。每个恶意软件文件的二进制代码长度不一,经过文本可视化后[3],可以看到恶意软件代码可以理解为由众多的 1 字节 16 进制数构成的 1 维向量,数据集中最长的长度为 405 248×16B,最短向量的长度为 8950×16B。若直接理解为图像,显然图像大小不一,带来后续训练和检测的困难,因此,需要提取每个恶意软件图像的纹理特征,并形成统一大小的特征纹理图像。考虑到代码的顺序性,只采用了水平方向的步长,而不考虑其他方向。

首先选择步长 1、2、3 建立灰度共生矩阵。原因如下:在操作系统以及汇编指令手册的分析中可以知道,计算机代码中大部分由 1 字节、2 字节、3 字节指令构成,如分类 1:没有操作数的指令,指令长度为 1 字节 ;分类 2:操作数只涉及寄存器的指令,长度为 2 字节;分类 3:操作数涉及内存地址的指令,长度为 3 字节等。因此,在灰度共生矩阵中采用了 1 字节、2 字节和 3 字节的灰度共生矩阵。首先分别以 1 字节、2 字节、3 字节为单位切割恶意软件代码行向量并做统计。通常的灰度共生矩阵考虑的是距离为 d 的 2 字节同时出现的统计,在大多数文献中[13-14]均为 2 字节矩阵。对于 1 字节,行列坐标为 0—255,统计每个字节中对应的数值出现个数。对于 2 字节灰度矩阵,则行代表第一字节,列代表第二字节,如:EB 3C 代表 EB 行,3C 列的值加 1,直至循环遍历整个恶意软件代码。其中,1 字节和 2 字节矩阵均可形成 256×256 的标准输入矩阵,1 字节灰度共生矩阵为主对角对称矩阵。以样本文件 di5lC6uMRX8hJ3BQtIVf.bytes 为例,通过图像可视化得到三个纹理图像(图 2)。



(a)样本单字节纹理 (b)样本二字节纹理 (c)样本三字节纹理

图 2 样本文件不同字节纹理图像

3 实验和仿真

3.1 数据集

本文采用的数据集为微软 2015 年恶意代码分类大赛中使用的数据集,BIG2015 数据集包含 9 个恶意家族的 21 741 个样本,其中 10 868 个样本为带标签的训练集,其他为不带标签的测试集。训练集中,每一个样本名为一个 20 字符的哈希 ID,以及对应的一个整数值作为家族标签,分别为 Ramnit、Lollipop、Kelihos ver3、Vundo、Simda、Tracur、Kelihos、ver1、Obfuscator.ACY 和 Gatak。对于每个类别,对恶意代码图像分别做 1 字节,2 字节和 3 字节纹理提取。

3.2 RF 实验结果

在这项工作中,根据第 2 部分生成的灰度共生矩阵生成方法,对每个恶意代码文件重新构成了 3 个 256×256 的共生矩阵 CSV 文件。并根据随机森林分类算法,将样本与百分比分割(80%)使用。其余 20%样本向量用作测试数据集。

$$\mathrm{T} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2), \cdots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$$

在随机森林分类器训练过程中,首先从 10 棵决策树开始进行训练,通过图 3 可以看出,随着决策树的增加,分类准确率逐步提升,但超过 30 棵后,准确率在 96%左右变化,不再增加。准确率随着深度增加而逐步提高,但超过 10 棵后增加不明显。通过图 4 可以得出,随机森林算法还可以评估所有变量的重要性,无需顾虑变量的多元共线性问题。现实情况下,一个数据集中往往有成百上千个特征,如何在其中选择对结果影响最大的那几个特征,以此来缩减建立模型时的特征数目可以提高算法的效率。这样的方法其实很多,比如主成分分析,lasso 等等。可以通过计算每个特征在随机森林中的每颗树上做了多大的贡献,然后取平均值,最后比较特征之间的贡献大小。该方法通常采用基尼指数来评价奉献率。变量重要性评分(variable importance measures)用 VIM 来表示,将基尼指数用 $Gini$ 来表示,在分类问题中,假设有 k 个类,样本点属于第 k 类的概率为 P_k ,则概率分布的 $Gini$ 指数的定义为:

$$Gini(p) = \sum_{k=1}^{k} P_k(1-P_k) = 1 - \sum_{k=1}^{k} p_k^2 \tag{5}$$

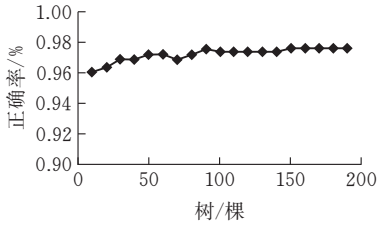


图 3　树的数目对正确率影响

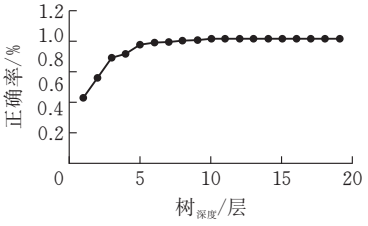


图 4　树的深度对正确率影响

基于图像纹理的双字节特征相对重要性见表 1。基尼系数越大，说明该变量对代码特征的分类重要性越高，经过实验，本方案得出的基于代码特征的指标重要性排序为"0000" ＞"BC66" ＞"474E" ＞"4E49" ＞"69C3"＞…"0001"。由表 1 可知，取前 600 个参数就可以达到 96%的累计重要性比率，因此可以进一步简化模型，分类代码时，无须每次计算全部的 256×256 个矩阵参数，而只需要计算列表中 600 个参数，即可达到近似的效果。经过实验，表格 1 中 GLCM-RF 简化版，可以达到 91%。

表 1　各列重要性排序表

排序	重要性	列名	代码
0	0.3017	0	0000
1	0.1092	35942	8C66
2	0.0759	18254	474E
3	0.0383	20041	4E49
4	0.0374	27075	69C3
5	0.0306	17863	45C7
6	0.0258	24037	5DE5
7	0.0239	51712	CA0
8	0.0204	59472	E850
9	0.0201	32129	7D81
10	0.0138	1674	068A
…	…	…	
599	0.0001	1	0001
…	…	…	…
65535	0	2	0002

3.3　KNN 聚类结果

为了检验基于恶意代码图像纹理特征提取的效果，继续采用 KNN 分类方法来验证该特征提取方式的有效性。并将随机森林中得到的重要性特征排序进行聚类可视化排序。由图 5 可知，各个类别在这些重要的特征上表现出了较强的聚类现象。

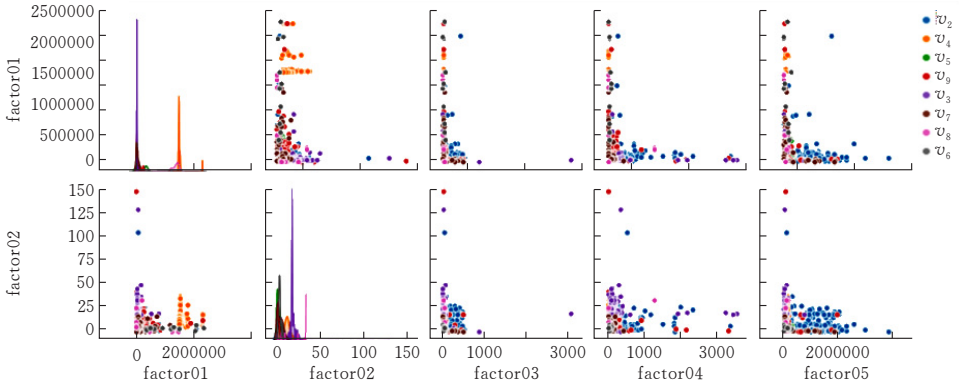


图 5　前 2 列特征聚类情况分析

由于按 GLCM 聚类的维数较多，达到 65 536 维，为了更好地的显示结果，采用了 TSNE 可视化方法。TSNE 是一种非线性降维算法，非常适用于高维数据降维到 2 维或者 3 维，图 6 为采用默认的 T 分布后 9 类别映射到二维后的结果。每种不同的演示代表了不同的种类，可以看出，红色和绿色的种类聚类特征明显，其他类则较为分散。

为了比较采用 GLCM 后对分类算法带来的影响，直接提取恶意代码文件的前 64K 字节作为数据集，用同样的分类方法来进行比较，通过分析统计数据可以看出，采用了图像纹理特征提取后的分类方法均比以前有了显著的提高，其中，GLCM-RF 随机森林方法准确率达到了 96.36%，较未采用图像特征提取的 RF 方法提高了约 10%，对于传统的 KNN 方法也有了较大的提高，分类效果明显。

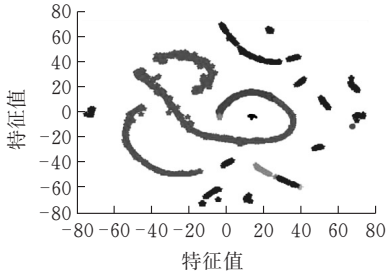


图 6　恶意软件分为 9 类并采用 TSNE 后的聚类显示

表 2 基于 GLCM 的 RF 与传统 KNN 方法比较

方法	正确率	召回率
KNN	61.1	0.42
GLCM-KNN	77.1	0.68
RF	85.36	85.69
GLCM-RF	96.36	0.96
GLCM-RF(简化版)	90.2	0.90

4 总结

本研究提出一种基于恶意代码图像纹理的随机森林分类方法,这种方法的优点是能够快速高效的识别恶意代码。并通过随机森林分析的特征重要性排序,可以简化图像特征维数,加快分类识别时间。研究结果表明,图像纹理提取简化了代码维数,提高了识别准确率。

[参 考 文 献]

[1] Nataraj L, Karthikeyan S, Jacob G, et al. Malware images: visualization and automatic classification[C]// Proceedings of the 8th International Symposium on Visualization for Cyber Security(VizSec'1), New York, USA, 2011.

[2] Fairuz, Amalina, Narudin, et al. Evaluation of machine learning classifiers for mobile malware detection [J]. Soft Computing, 2016,20(1):343-357.

[3] 任卓君, 陈光,卢文科. 基于 N-gram 特征的恶意代码可视化方法[J]. 电子学报, 2019, 47(10): 2108-2115.

[4] Yan H, Zhou H, Zhang H. Automatic malware classification via PRICoLBP[J]. Chinese Journal of Electronics, 2018, 27(4): 852-859.

[5] 乔延臣, 云晓春, 张永铮, 等. 基于调用习惯的恶意代码自动化同源判定方法[J]. 电子学报, 2016, 44(10): 2410-2414.

[6] Gandotra E, Bansal D, Sofat S. Malware analysis and classification: A survey[J]. Journal of Information Security, 2016, 5(2): 56-64.

[7] 陈艳秋, 孙培立. 一种基于类别强信息特征和贝叶斯算法的中文文本分类器[J]. 计算机应用与软件, 2014 (8): 330-333.

[8] Gotlieb C C, Kreyszig H E. Texture descriptors based on co-occurrence matrices [J]. Computer Vision, Graphics, and Image Processing, 1990, 51(1): 70-86.

[9] Kancherla K, Mukkamala S. Image visualization based malware detection [C]//2013 IEEE Symposium on Computational Intelligence in Cyber Security (CICS). IEEE, 2013: 40-44.

[10] 邓煜, 李明, 周稻祥. 基于两阶段随机森林的螺丝锁附结果判别研究[J]. 太原理工大学学报, 2020, 51(2): 198-205.

[11] Quentin Brabant, Miguel Couceiro, Didier Dubois, Henri Prade, Agnès Rico. Learning rule sets and Sugeno integrals for monotonic classification problemsEB/OL]. [2020-02-13] https://doi. org/10. 1016/j. fss. 2020.01.006.

[12] 刘建花. K-means 聚类算法的改进与应用[J]. 太原师范学院学报(自然科学版), 2020, 19(1): 81-83.

[13] 韩晓光, 曲武, 姚宣霞, 等. 基于纹理指纹的恶意代码变种检测方法研究[J]. 通信学报, 2014, 35(8): 125-136.

[14] 张晨斌, 张云春, 郑杨, 等. 基于灰度图纹理指纹的恶意软件分类[J]. 计算机科学, 2018, 45(6A): 383-386.

Classification of Malware Based on Texture Feature and Random Forest

LIU Yuqiang, LI Jun, FAN Zhipeng

(*School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: The identification and defense of malware becomes more difficult with the progress of science. In view of the difficulty of the recognition analysis in preventing the possible harm in time, a static classification method based on the grayscale texture features of the image is proposed. According to the instruction length of the code, the multi byte image texture of the virus code is designed and extracted, and unified into two dimensional features. And then, all feature files are used as training sets to classify random forest machine learning methods. Experiments with standard datasets show that the accuracy of this method is 96.36%. By analyzing the importance of each byte code feature, a simplified classification method is proposed.

Keywords: malware classification; grayscale texture feature; random forest

[责任编校: 张岩芳]