

# 基于机器学习的密码算法识别与分析

夏锐琪<sup>1</sup>, 李曼曼<sup>2</sup>, 陈少真<sup>2</sup>

<sup>1</sup>信息工程大学 网络空间安全学院 郑州 中国 450001

<sup>2</sup>密码科学技术国家重点实验室 北京 中国 100093

**摘要** 基于人工智能的密码分析技术是目前信息安全领域高度关注的问题之一, 利用机器学习的唯密文密码算法识别是其中不可或缺的关键。研究如何筛选高质量的密文特征指标提升识别模型的性能, 以及改进非固定密钥条件下密文算法识别效果是当前研究工作的难点。建立性能优异的特征工程和机器学习模型是一种理想的方案, 本文基于随机森林、Adaboosting、全连接神经网络等模型进行随机密钥条件下的密码算法识别实验, 并对特征工程使用的指标建立基于信息熵和维度标准的筛选方法, 针对诸多的特征指标进行优化研究, 全面细致地对实验现象进行理论分析。本文首先对密文随机性指标(NIST SP 800-22)根据其定义按维度大小标准分类, 依据分类结果计算各个指标的信息熵理论值, 按照信息熵的大小关系对指标性能进行排序筛选, 分析挑选出适合识别密码算法的高质量指标。由筛选结果选择9种代表性特征指标, 对包括分组密码与公钥密码在内的7种密码算法, 在随机密钥加密条件下, 建立4种机器学习模型进行识别实验。对实验现象从特征指标和模型原理等角度展开理论分析, 并结合理论和实验结果给出一类随机密钥下密码算法高效识别的结论。与先前的相关工作相比, 本文实现了在随机密钥条件下对多种类型密码算法的高效唯密文识别, 对各种算法的识别准确率提高了42%到55%, 密文所需数据量相应地降低了约40%。实验与理论结果表明, 利用几种高信息熵的多维指标作为特征数据, 识别随机密钥条件下的密码算法具有较高的识别准确率。

**关键词** 密码分析; 机器学习; 随机性指标; 信息熵; 密数据识别

中图法分类号 TN918 DOI号 10.19363/J.cnki.cn10-1380/tn.2025.01.11

## Identification and Analysis of Cryptography Algorithms based on Machine Learning

XIA Ruiqi<sup>1</sup>, LI Manman<sup>2</sup>, CHEN Shaozhen<sup>2</sup>

<sup>1</sup> Department of Cyberspace Security, Information Engineering University, Zhengzhou 450001, China

<sup>2</sup> State Key Laboratory of Cryptography Science and Technology, Beijing 100093, China

**Abstract** The cryptanalysis based on artificial intelligence is one of the most popular problems in information security. Cryptography identification using machine learning is the crucial part in the domain, which is the key step for cryptanalysis. Currently the main difficulties are that how to improve the properties of identification in the conditions of unfixed keys, and select the effective indices used for cryptography identification in order to enhance the performance of identification models. Constructing the effective feature program and establishing the suitable machine learning models are the reasonable schemes. In this work, the experiments used Random Forest, Adaboosting, fully connected neural network algorithm, etc as the models and the cipher algorithms are encrypted by the random keys. In spite of these, the theoretical analysis and selection of the features indices were performed based on the calculation of information entropy and dimensions, which improve the feature program. After that, the theoretical research about the experiments' phenomena are proposed. First, we classify the indices (NIST SP 800-22) under the criterion of dimensions based on their definitions. Second, according to the classification, we calculated the information entropy of the indices and compared the information entropy of the indices. Select the effective indices of high entropy as the feature indices fitting for cryptography identification. Then the identification experiments were based on the 9 features, 7 algorithms including block ciphers and public keys cryptography using random keys, and 4 machine learning models. Subsequently, the theoretical analysis of the experiments' results was put forward in the angles of feature indices and the principles of models. Finally, the conclusion of identifying cryptography algorithms effectively was proposed based on the theoretical analysis and experiments. The accuracy of our work is 42% to 55% higher than the previous work. Meanwhile, the ciphers data are smaller than the related work around 40%. The experiments and analysis showed that multidimensional features with high entropy can make the cryptography identification's accuracy come to a higher level.

**Key words** cryptanalysis; machine learning; randomness indices; information entropy; cipher identification

通讯作者: 李曼曼, 博士研究生, 讲师, Email: limanman15@163.com。

本课题得到数学工程与先进计算国家重点实验室开放基金课题(No. 2019A08)、资源受限环境下密码算法组件评估关键技术研究(No. 2019427)资助。

收稿日期: 2022-03-14; 修改日期: 2022-04-12; 定稿日期: 2024-11-20

## 1 引言

随着网络和计算机技术的发展, 信息安全和密码加密保护等技术成为网络空间安全领域关键的研究问题之一<sup>[1]</sup>。在现代密码学中, 从截获的密文中恢复密钥、解密得到相应的明文是密码分析者的重要工作。实际工作中, 密码分析者从公开信道中通常仅能掌握密文信息, 而对具体的加密算法往往是未知的, 这为恢复密钥、得到明文增加了困难。因此, 如何准确识别加密算法是密码分析领域的关键。近些年, 人工智能快速发展, 为解决密码分析问题提供了有力的技术支撑<sup>[2]</sup>。以机器学习为模型的密码算法识别方案对固定密钥加密条件下诸多国际标准密码算法的识别具有良好的效果, 目前广受关注, 相关成果层出不穷。

### 1.1 相关工作

基于机器学习的密码算法识别工作早期由 Ramzann<sup>[3]</sup>于 1998 年提出, 并论证了其可行性。随后, Dileep 等人<sup>[4]</sup>系统地对几种标准分组密码算法进行尝试性的识别工作, 通过建立词包模型提取密文特征, 使用支持向量机和 K 近邻聚类算法进行实验与比较, 两种算法对 DES(Data encryption standard)等算法的识别准确率均高于 50%; Manjula<sup>[5]</sup>和 Chou 等人<sup>[6]</sup>先后利用决策树等算法对包括公钥密码算法在内的多个密码算法密文进行了识别工作, 识别准确率可达 70%以上, 但在面对非固定密钥条件时结果较差。Mishra 等人<sup>[7]</sup>通过提取新的密文特征数据, 利用决策树等模型对 AES(Advanced encryption standard)、Blowfish 等分组密码算法进行了固定密钥条件下的识别工作, 进一步提高了对大数据量密文的识别准确率。近年来, Mello 等人<sup>[8]</sup>对传统的机器学习决策树算法进行了改进, 利用新的算法模型对常见的几种密码算法加以识别, 在固定密钥情况下其模型识别的准确率可接近 100%。在此基础上, 陆续呈现出多种机器学习算法模型<sup>[9-11]</sup>, 通过提取密文序列多种随机性指标作为特征数据对多种国际标准密码算法的密文进行较为详细全面的研究, 比较完整地建立了各种密码算法识别方案。

随着深度学习技术的发展, 基于人工智能的密码算法识别工作不再局限于一般机器学习方法。凭借深度神经网络模型, 对更复杂的密码算法识别问题有了新的研究成果。Souza 等人<sup>[12]</sup>通过建立深度神经网络模型, 初步探究了对小批量密文大小的 AES 算法的区分攻击, 取得了较好的准确率; Sandeep 等人<sup>[13]</sup>利用深度卷积神经网络 CNN(Convolution neural

network)等模型对多种标准分组密码算法进行了识别, 改进识别准确率达到接近 100%。进一步地, 利用 BP(Back propagation)神经网络等三种成熟神经网络模型对多个常用算法进行了基于随机性指标特征数据的深度学习实验, 在固定密钥条件下优化了原有的实验结果<sup>[14]</sup>。

然而, 为了保证加密安全性, 实际中加密密钥一般不是唯一固定的, 利用机器学习对非固定的随机密钥条件下的密文识别工作相对较少。针对随机密钥情况, 已知的相关工作实验效果较固定密钥情况差, 因此利用机器学习或深度学习进行密码算法识别有广阔的应用前景。另一方面, 大多数基于机器学习的密码算法识别是通过提取密文序列 15 种由 NIST(National institute of standards and technology)公布的密文随机性指标作为特征数据输入模型进行实验, 目前已有工作对诸多特征指标的识别效果缺乏细节分析, 导致实验效率不高, 因此对适合识别工作的高效性特征指标的筛选方法需进一步完善。综上所述, 将基于机器学习的密码算法识别推广到非固定的随机密钥条件下, 具有重要的应用价值。同时, 结合随机性指标特征性质与识别效果进行对比和理论分析, 建立有效的指标筛选方法可以很好地补充完善密码算法识别方案的工作细节, 使得这项技术更加成熟, 能够被广泛运用到各种领域。基于上述分析, 本文的主要工作如下。

### 1.2 主要工作

本文主要研究了随机密钥下机器学习密码算法识别实验与特征指标性质分析筛选方法, 以及对实验现象的理论对比分析等, 具体工作如下:

(1) 通过分析 15 种随机性指标特征的定义和性质, 首先对特征指标给出按维度标准下的等价类的划分。根据分类结果, 计算各等价类指标在随机条件下的信息熵, 分析其大小关系并予以排序。按信息熵大小和维度高低筛选能够较好识别密码算法的一类指标。本部分旨在建立有效的指标筛选机制, 挑选出适合作为密文序列提取的几种特征指标, 有效精简工作冗余, 提高密码识别工作效果, 为后续进行以随机密钥条件下的分组密码算法识别实验奠定基础。

(2) 在(1)的基础上, 选择随机森林、Adaboosting、全连接以及前馈神经网络模型, 在随机密钥条件下, 对 4 种常用分组密码、1 种序列密码和 2 种公钥密码算法的密文提取 9 种代表性特征, 对特征数据进行识别实验。实验结果与文献[4-8]相比, 识别精确率总体提高约 50%; 与文献[9-11]相比, 识别精确率普遍

提高 45%左右。同时,与文献[12-13]相比,所使用的密文数据量普遍降低 40%左右。实验结果普遍优于先前的工作,并对各指标的工作效果给出对比研究。

(3) 针对(2)中对比实验的结果,计算各特征指标的实际信息熵值与分布,系统分析并验证了几种特征实验对比结果,补充解释了机器学习模型结构、数据量等因素对实验现象的影响。基于实验结果和理论分析,最终给出在随机密钥条件下利用机器学习有效识别分组密码算法工作的结论。

### 1.3 文章结构

文章结构安排:第一部分引言,简单介绍相关背景和主要研究内容;第二部分准备工作,描述随机性指标、密码算法和机器学习模型等相关概念;第三部分随机性指标分类与筛选,对随机性指标进行分类划分,计算随机条件下每个等价类中指标的信息熵表达式,比较分析其大小关系,建立以信息熵和维度为标准的特征指标筛选机制;第四部分基于机器学习的密码算法识别实验,在前文的基础上,利用 9 种代表性特征指标和 4 种机器学习模型,对随机密钥条件下 7 种密码算法进行识别实验;第五部分,针对实验中各指标对比结果给出详细的分析,得到实验结论;第六部分,总结全文。

## 2 准备工作

基于机器学习的密码算法识别的主要工作流程如下:收集密码算法加密明文得到的密文序列;根据特征数据的定义,建立特征工程提取密文序列的特征指标作为特征数据集;将特征数据输入机器学习模型进行分类和识别;将训练测试完成的机器学习模型用于对未知算法类型密文进行算法种类的预测识别。具体过程如图 1 所示。

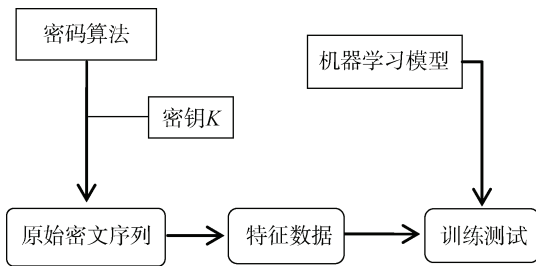


图 1 基于机器学习的密码算法识别流程

Figure 1 The flow of cryptography identification using machine learning

由图 1 可知,提取特征数据(特征工程)在密码算法识别中具有关键作用。在随机密钥条件下,原始密文本身的差异性和区分性并不明显,通过提取密文

特征能够放大密文差异性,改进机器学习模型的区分能力,有效实现模型对密文的区分和密码算法的识别。因此,提取特征指标对决定密码算法识别效果好坏有至关重要的作用。

### 2.1 密文随机性指标

密码算法的密文随机性检测指标是由 NIST 用于国际标准密码算法安全性评估和检测工作所发布的标准指标<sup>[15]</sup>,该指标共 15 种。在密码算法识别工作中,通常不计算统计检测值,只提取对应指标值。具体指标内容如表 1 所示:

表 1 密文随机性指标  
Table 1 The randomness indices

指标	含义
频率指标	提取 0,1 比特在待检测序列中的比例
块内频数指标	在待检测序列的子块中提取 0,1 比特的比率
游程指标	提取不同长度连续 1 或 0 比特序列的个数
最长 1 游程指标	提取最长 1 比特游程的长度
二元矩阵秩指标	将序列构造为若干矩阵,计算矩阵的线性相关性(秩)
离散傅里叶变换指标	检测序列的周期性,计算其与随机序列周期性的距离(差距)
非重叠块匹配指标	统计序列中某特定子序列出现的次数
重叠块匹配指标	与非重叠块匹配指标类似,其区别在于匹配过程对序列移动的步长长度不变
通用统计指标	统计匹配序列压缩前后的比特值,检测序列能否被压缩
线性复杂度指标	计算构造线性反馈移位寄存器的最小长度
序列指标	任意长度为 $m$ 的二元序列有 $2^m$ 种子序列,统计这些子序列在待检测序列中出现的次数
近似熵指标	提取 $m$ 位可重叠子序列和 $m+1$ 位可重叠子序列出现的频数
累加和指标	提取待检测序列中部分和的数值
随机游动指标	统计序列中特定子序列在一个随机长度中出现的次数,并与随机序列相比较
随机游动频数指标	与随机游动指标类似,统计特定子序列在游走过程中经过特定位置的次数

以上指标可用于提取待识别密文序列的特征数据以输入模型训练和测试。机器学习模型的性能很大程度取决于输入特征数据的质量和区分性等,不同指标体现待检测序列的特征信息量是不相同的。建立有效可行的指标筛选机制,选取能够高效识别密码算法的特征指标非常重要。因此,机器学习特征数据的指标选取是密码算法识别实验的主要工作之一。

## 2.2 机器学习算法介绍

由图 1 可知,除了特征数据的提取外,机器学习模型也是决定识别准确度的重要因素之一。具有良好泛化性、擅长处理多维数据的机器学习模型能够显著提高对未知密码算法识别和分类的精度,降低可能的误差。本文深入分析已有工作和机器学习算法的工作原理,采用随机森林算法、Adaboosting 算法(两种具有代表性的常用机器学习算法)以及全连接神经网络和前馈神经网络模型两种神经网络模型作为识别的基本模型展开研究。

### 2.2.1 随机森林模型

传统机器学习模型处理多维复杂的特征信息效果欠佳,面对高密度、大容量的信息数据工作效率不高,且准确度较差。集成学习是目前较为流行的一种机器学习策略,在诸多应用领域成效显著。随机森林算法(Random forest)是集成学习策略中最典型实用的算法之一<sup>[16]</sup>,在处理多维复杂数据格式的特征数据时具有良好的性能。随机森林算法的基本结构是通过复合叠加多层次数量的决策树基本单元组合形成。每个决策树单元独立工作,且它们判断的目标和标准是一致的。输入数据经过随机采样后,通过每一分支的决策树单元时,该决策树均会作出对当前数据对应的判断,并将结果向下一个单元传递。当到达最终输出结果层时,通过求每个决策树单元结果数值的众数(或平均数),得到模型给出的结果。随机森林模型擅长处理高维且取值范围广的数据,其泛化性能较强,但对被噪声扰动数据集的处理仍差强人意。模型示意图如图 2 所示。

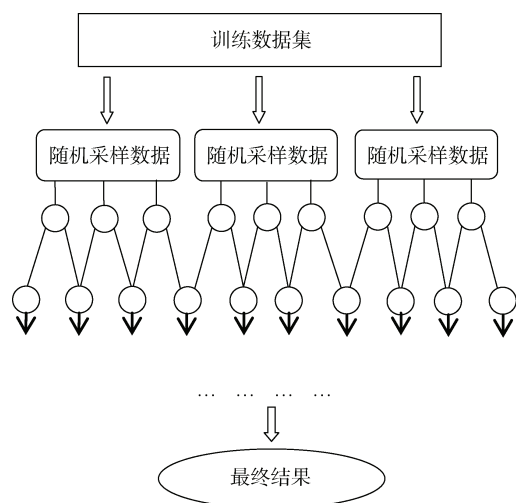


图 2 随机森林模型

Figure 2 The model of random forest

### 2.2.2 Adaboosting 模型

在集成学习领域,以 Adaboosting 算法为典型代

表的另一类模型同样具有十分优异的性能<sup>[17]</sup>。与随机森林算法相比,虽然二者在结构和原理上具有一定的相似性,但 Adaboosting 模型侧重考虑数据集中出现错误样本的调整分析。对结构中前一层训练错误的样本,将在后一层中增加其权重,以此不断调整权重分布比例,旨在优先解决划分错误的样本数据集。其模型示意图如图 3 所示。

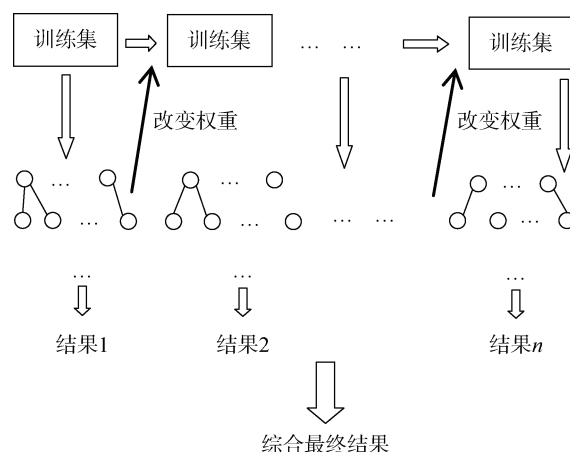


图 3 Adaboosting 模型

Figure 3 The model of Adaboosting

Adaboosting 模型是串联结构,在给每个单元模型附加权重值后,能够让效果好的分类器单元发挥更多的作用,而效果一般的则适当降低其作用。Adaboosting 算法在分类和回归工作中效果十分显著,实验表明经过多次训练调整,该模型能够将训练准确率稳定地提升到比一般机器学习模型更高的水平。

### 2.2.3 全连接神经网络模型

神经网络模型是机器学习重要的组成部分,也是深度学习技术的主要模型。全连接神经网络是其中实用且简洁的一种神经网络结构,在模式识别和回归预测等任务中具有良好的性能<sup>[18]</sup>。该模型主要由输入层、隐藏层和输出层组成。当训练数据输入后,各隐藏层中的神经元单元对各个分量进行逐个处理,与权重值加和后由激活函数计算输出值。前一个神经元将输出值传递给下一个连接层对应的神经元,以此类推。训练过程包括向前传递和向后传递,其结构如图 4 所示。

全连接神经网络模型神经元激活函数为 ReLU 函数,输入层神经元个数由输入数据的维度确定,分类判断的输出结果为一维数据。

### 2.2.4 前馈神经网络模型

前馈神经网络是较为成熟且性能优异的神经网络之一,在文本处理等方面取得显著的成效<sup>[19]</sup>。该神



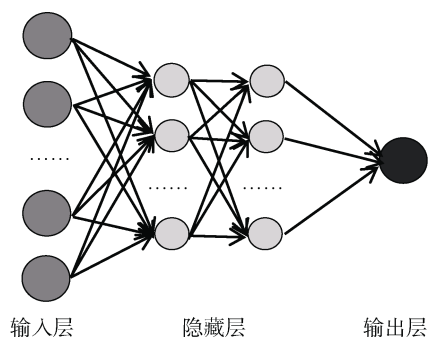


图 4 全连接神经网络模型

Figure 4 The model of Fully connected neural network

神经网络模型结构与全连接神经网络模型类似,与之不同的是,在训练过程中每个神经元的参数将不再根据训练结果的误差情况做出修正,也即神经网络结构没有反馈机制等作用。前馈神经网络的输出可以是单个神经元结果,也可为多维输出,前馈神经网络的激活函数一般为 Sigmoid 函数。该模型结构如图 5 所示。

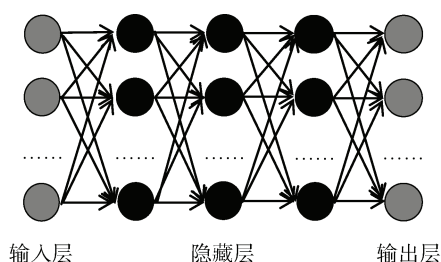


图 5 前馈神经网络模型

Figure 5 The model of feedforward neural network

## 2.3 密码算法介绍

现代密码学按密钥分配方式的不同可分为对称密码与非对称密码。对称密码由分组密码、序列密码及哈希函数组成,而非对称密码主要指公钥密码。不同的密码算法凭借其复杂的加密结构或计算原理保证明文的安全性,是信息安全领域的主要技术手段。

为了使得本文工作更具有一般性和应用价值,选取包括分组密码、序列密码和公钥密码在内的 7 种常用的国际标准算法作为实验对象。对它们的介绍如下。

DES(Data encryption standard)算法<sup>[20]</sup>。该算法由美国联邦信息处理标准于 1977 年采用,其分组长度和密钥长度均为 64 比特,加密轮数 16 轮,属于 Feistel 结构类分组密码。该算法是分组密码算法中最著名的典型算法之一。

AES(Advanced encryption standard)算法<sup>[21]</sup>。该算法是 DES 算法后由美国国家标准与技术研究院发布的高级加密标准。其分组长度为 128 比特,密钥长度分为 128/192/256 比特三种,加密轮数随密钥长度不同分为 10/12/14 轮,加密结构为 SPN(Substitution permutation network)结构。AES 算法可以有效抵抗现有的几乎所有攻击手段,并广泛用于硬件加密领域。本文所使用的加密规格为 AES-128。

KASUMI 算法<sup>[22]</sup>。KASUMI 算法是 3GPP 标准的通信加密核心算法,在通信信息保护方面具有重要的应用价值。其分组长度为 64 比特,密钥长度 128 比特,加密轮数为 8 轮,该算法属于 Feistel 结构。

PRESENT 算法<sup>[23]</sup>。PRESENT 算法为轻量级分组密码,广泛用于资源受限的设备中。加密结构为 SPN 结构,其分组长度为 64 比特,密钥长度 80/128 比特。其轮数为 31 轮。

Grain-128 算法<sup>[24]</sup>。Grain-128 算法属于非线性反馈寄存器中的标准序列密码,被大量应用于硬件安全保护等领域。其密钥长度为 128 比特。

RSA 算法<sup>[25]</sup>。RSA 算法原理基于大数分解的困难性,私钥很难由公钥直接计算获得。其密钥长度通常为 1024 比特或 2048 比特。但其加解密速率比对称密码慢。

ElGamal 算法<sup>[26]</sup>。ElGamal 算法原理来自有限域上离散对数的求解。该公钥密码广泛用于数字签名等领域。

## 3 随机性指标分类与筛选

### 3.1 指标分类筛选的必要性

根据准备工作,在 15 种随机性指标中,每种指标对密文序列特征提取的方法各不相同。就密码算法的随机性检测与安全性评估而言,各个指标都能发挥重要的作用。但就密码算法识别而言,并非任何一种随机性指标都适合作为模型输入的密文特征数据。

在密码算法识别实验中,机器学习模型将分析待识别算法在各自密文特征数据上的差异性,并依此作出判断。随机密钥条件下,各种算法之间密文的混淆性和扩散性都趋于完全随机,如何改进识别模型的效果是较为困难的问题。为了避免不必要的重复实验,提高密码算法识别工作的准确性、效率和实用性,对随机性指标给出准确适当的分类,建立有效的筛选标准,指出效果更好的一类指标。这是解决随机密钥条件下分组密码算法识别工作的必要前提。

### 3.2 随机性指标的分类

根据随机性指标的定义, 将 15 种指标按照返回值的数据维度分为低维指标和高维指标。(低维指标为维度不超过 2 的随机性指标, 多维指标反之) 如表 2 所示。

表 2 随机性指标分类

Table 2 The classification of randomness indices

低维指标	多维指标
频率指标	块内频数指标
最长 1 游程指标	游程指标
二元矩阵秩指标	非重叠块匹配指标
离散傅里叶指标	重叠块匹配指标
通用统计指标	序列指标
线性复杂度指标	近似熵指标
累加和指标	随机游动指标
	随机游动频数指标

特征提取往往由原始密文件按某长度划分为若干子块, 针对每个密文子块提取其特征, 形成一组数据后输入模型进行训练和测试。显然, 多维指标能够对每个子块实现多次不重复提取特征, 深层次多角度地反馈其特征, 所涵盖的信息量也较多, 低维指标则稍显逊色。另一方面, 根据机器学习模型的工作原理, 输入特征数据所提供的信息和特征越多, 其正确分类的能力越好, 发生错误的情况会更少, 同时, 多维指标能够更好地被集成学习、神经网络等模型处理。故根据分类结果, 随机密钥条件下进行密码识别实验时, 可推断多维指标的识别效果一般高于低维指标。

### 3.3 基于信息熵的指标筛选

信息熵是信息与编码学中的重要基本概念之一, 其定义如下:

**定义 1.** 信息熵<sup>[27]</sup>。信息熵是接收的每条消息中包含的平均信息量。设  $X$  为随机变量,  $P(X)$  为其概率密度函数,  $I(X) = -\log_b P(X)$  为  $X$  信息量, 则当信源为有限信息时, 信息熵可表示为:

$$H(X) = \sum_i P(x_i) I(x_i) = -\sum_i P(x_i) \log_b P(x_i) \quad (1)$$

当  $b=2$  时, 熵的单位为比特(bit)。

由 3.2 分类结果, 分别计算在随机密钥条件下密文各个特征指标的信息熵, 并根据信息熵大小关系建立筛选方法。

#### 3.3.1 低维指标的信息熵

由表 2, 低维指标共 7 种类型, 分析指标的定义和工作原理可知, 随机性统计检测指标输出值由概

率阈值  $p\_value$  及其置信区间范围  $\delta$  决定。阈值  $p\_value$  等取值同样由 NIST 规定, 可参见文献[15]。为便于表示指标类型, 将低维指标标记符号及对应的阈值标准表示为表 3。

表 3 低维指标阈值

Table 3 The  $p\_value$  of low dimensional randomness indices

低维指标	阈值
频率指标( $F_1$ )	0.109599
最长 1 游程指标( $F_2$ )	0.180609
二元矩阵秩指标( $F_3$ )	0.532069
离散傅里叶指标( $F_4$ )	0.330390
通用统计指标( $F_5$ )	0.427733
线性复杂度指标( $F_6$ )	0.845406
累加和指标( $F_7$ )	0.219194; 0.114866

则对每个指标取值概率  $P(X_i), i \leq 2$

$$p\_value - \delta \leq P(X_i) \leq p\_value + \delta \quad (2)$$

指标  $F_2$  到  $F_6$  对应一维数据, 其信息熵满足

$$H_{F_i}(X) = -p \log_2 p < 1 \quad (3)$$

其中  $H_{F_i}$  为对应指标信息熵,  $2 \leq i \leq 6$ 。对于  $F_1$  与  $F_7$ , 其信息熵表示为两项:

$$H_{F_1}(X), H_{F_7}(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \quad (4)$$

利用放缩法, 取两个指标  $p\_value$  值置信区间范围内的一个公共上界  $p\_value_{sup} = 0.3$ , 估计信息熵上界得到

$$H_{F_1}(X) \leq -2p\_value_{sup} \log_2 p\_value_{sup} \leq 1 \quad (5)$$

$$H_{F_7}(X) \leq -2p\_value_{sup} \log_2 p\_value_{sup} \leq 1 \quad (6)$$

综上所述, 低维指标在随机条件下的信息熵值分布均不超过 1 比特。最大值为频数指标  $F_1$  与累加和指标  $F_7$ , 约为 1 比特。

#### 3.3.2 多维指标的信息熵

同理可得, 将 8 种多维指标记号和对应的  $p\_value$  值表示为表 4。

设随机密钥条件下加密得到的密文比特序列为  $Y$ , 设  $Y$  长度为  $l$  bit, 特征数据的维度为  $n \in N$ 。在随机密钥条件下,  $Y$  的每一位均被视为是随机的, 且每一位比特只能取 0 或 1, 则密文序列每一比特  $y_i, i \in \{0, 1, 2, \dots, l\}$  取 0, 1 的概率  $P(y_i)$

$$P(y_i = 0) = P(y_i = 1) = \frac{1}{2} \quad (7)$$

对块内频数指标而言,  $n$  个密文子块中 0, 1 出现的概率也各为  $\frac{1}{2}$ , 故块内频数指标的信息熵为:

$$H_{F_8} = \frac{1}{n} \left( - \sum_i P(y_i = 1) \log_2 P(y_i = 1) \right) = \frac{1}{2} \quad (8)$$

同理, 对于游程分布指标, 长度为  $n=1, 2, \dots, r_{\max}$  的各个长度的游程序列在随机密钥加密下的密文序列中的概率  $P(r=n)$  为

$$P(r=n) = \binom{l}{n} \left( \frac{1}{2} \right)^{n+2} \quad (9)$$

表 4 高维指标阈值

Table 4 The  $p$ -value of high dimensional randomness indices

高维指标	阈值
块内频数指标( $F_8$ )	0.706438
游程指标( $F_9$ )	0.500798
非重叠块匹配指标( $F_{10}$ )	0.647302
重叠块匹配指标( $F_{11}$ )	0.110434
序列指标( $F_{12}$ )	0.843764; 0.561915
近似熵指标( $F_{13}$ )	0.235301
随机游动指标( $F_{14}$ )	0.573306; 0.197996; 0.164011; 0.007779; 0.778616; 0.365752; 0.790853; 0.792378
随机游动频数指标( $F_{15}$ )	0.858946; 0.794755; 0.576249; 0.493417; 0.633873; 0.917283; 0.934708; 0.816012; 0.826009; 0.137861; 0.200642; 0.441254; 0.939291; 0.505683; 0.445935; 0.512207; 0.538635

(其中  $\binom{l}{n}$  为  $l, n$  的组合数公式)

将(9)式代入信息熵定义式, 计算后得到

$$H_{F_9} = 1 - \frac{\binom{n}{1} + \binom{n}{2} - 3}{2^n} \quad (10)$$

对于扑克检测指标, 当选定的子序列长度为  $n$  时, 密文子序列考查的类型就有  $2^n$  种。由于密文序列整体是随机的, 故每个维度  $n$  下子序列的概率  $P_n$  即为

$$P_n = \frac{1}{2^n} \quad (11)$$

将(11)式代入信息熵定义式, 得到

$$H_{F_{12}} = 2 - \frac{n+2}{2^n} \quad (12)$$

指标  $F_{10}$ 、 $F_{11}$  以及  $F_{13}$  工作原理与  $F_{12}$  类似, 在指标提取形式上无本质差异, 故几种指标的信息熵理论值相同。

最后, 针对指标  $F_{14}$ 、 $F_{15}$ , 根据表 4 阈值取值范围计算得到信息熵取值范围:

$$2 < H_{F_{14}}, H_{F_{15}} < 3 \quad (13)$$

综上所述, 指标  $F_8$  信息熵在随机条件下最小, 为 0.5 比特, 比较式(10)与(12), 当维度  $n > 2$  时,

$$H_{F_9} < 1 < H_{F_{12}} < 2 \quad (14)$$

再由式(13)分析知:

$$H_{F_{12}} < 2 < H_{F_{14}}, H_{F_{15}} \quad (15)$$

因此, 针对多维指标的信息熵大小关系, 得到结论如下:

$$H_{F_8} < H_{F_9} < H_{F_{12}}, H_{F_{10}}, H_{F_{11}}, H_{F_{13}} < H_{F_{14}}, H_{F_{15}} \quad (16)$$

### 3.3.3 特征指标筛选方法

由上文结论, 15 种随机性指标在随机条件下的信息熵大小关系为:

$$\{H_{F_2}, H_{F_3}, H_{F_4}, H_{F_5}, H_{F_6}, H_{F_8}, H_{F_9}\} < H_{F_1}, H_{F_7} \approx 1$$

$$1 < H_{F_{12}}, H_{F_{10}}, H_{F_{11}}, H_{F_{13}} < H_{F_{14}}, H_{F_{15}} \quad (17)$$

式(17)表示了不同维度指标间信息熵的大小关系, 是指标筛选方法的关键。图 6 表示了各个指标的信息熵分布情况, 从该图观察知: 除块内频数分布指标与游程指标外, 多维指标的信息熵均高于低维指标的信息熵, 这说明本文对指标的分类标准是基本合理的。

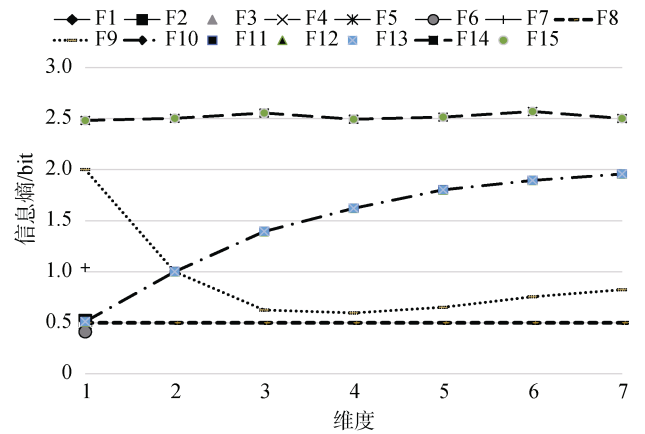


图 6 指标信息熵分布

Figure 6 The distributions of indices' entropy

信息熵用于衡量不同数据之间信息量的大小, 反映数据信息的不确定性和冗余度等性质。另一方面, 机器学习等模型的工作性能主要由信息增益决定<sup>[28]</sup>, 而信息增益同样由输入数据的信息熵决定。信息熵越大, 则信息增益越明显, 机器学习性能越突出。因此, 就特征指标而言, 其信息熵越大, 所能反映原密文序列的信息量也越大, 作为待识别密文的提取特征更有效。对机器学习模型而言, 具有更高信息熵的数据集携带的信息量更大, 模型处理该类数

据的性能越强,能够更精准地完成目标任务。因此,利用指标维度的分类和信息熵大小关系,可推断筛选多维高信息熵的指标(如序列指标、随机游动指标等)能够显著提高密码算法识别的效果。

## 4 基于机器学习的密码算法识别

根据上文的筛选方法及结果,实验选择 6 种多维指标和 3 种低维指标作为对比实验,考查不同类型指标的性能并验证筛选方法的合理性与可靠性。实验选择的指标如表 5 所示。

表 5 实验指标类型

Table 5 The species of indices

高维指标	低维指标
块内频数指标( $F_8$ )	频率指标( $F_1$ )
游程指标( $F_9$ )	二元矩阵秩指标( $F_3$ )
序列指标( $F_{12}$ )	累加和指标( $F_7$ )
近似熵指标( $F_{13}$ )	
随机游动指标( $F_{14}$ )	
随机游动频数指标( $F_{15}$ )	

实验所使用机器学习模型为随机森林模型、Adaboosting 模型、全连接神经网络模型与前馈神经网络模型,实验对象为 7 种国际标准密码算法: DES、AES-128、KASUMI、PRESENT、Grain-128、RSA 与 ElGamal。

实验条件为随机密钥加密。实验选取的明文来源于美国国家开放语料库(OANC)常用文本,从其中随机选择文本作为明文。实验的相关环境为:硬件配置 Intel Core i7,软件配置为 Python3.7 程序设计语言环境,Scikit-learn 版本 0.23.1、Tensorflow 1.14.0 与 Keras 2.3.1。

实验设计思路为在随机密钥加密不同明文条件下,收集各种算法的密文序列,分组密码算法的工作模式为 ECB 模式,将每个算法对应的密文存储为密文件。按随机固定的密文子块长度划分并提取每个子块的 9 种特征指标,随后建立随机森林、Adaboosting 模型、全连接神经网络模型与前馈神经网络模型,输入特征数据进行实验,并得到实验评价指标。

### 4.1 密码算法识别流程

随机密钥条件下基于机器学习的密码算法识别实验遵从机器学习训练测试的工作流程,按收集原始数据、特征工程(提取特征指标)、输入模型训练测试、得到模型指标评估等步骤进行,具体为:

(1) 对 7 种分组密码算法利用随机生成的密钥加

密随机明文数据,按实验数据量要求分别得到 7 个对应的密文文件,其大小为  $2^{21}$  比特量级,约 1 GB。设各密文总量为  $S$ 。

(2) 随机固定分段长度  $L$  ( $L > 0.001S$ ),将各个算法的密文件按该长度进行划分,得到  $m = \lfloor S/L \rfloor$  组密文序列,这些密文序列为特征指标提取的密文子序列。

(3) 按照 9 种随机性指标特征数据的定义,对密文子序列提取对应的特征数据:

1) 块内频数指标:按照分块长度  $l$  对密文子序列划分为  $n = \lfloor L/l \rfloor$  个子块,统计每个子块的频数,得到该子序列的特征指标。

2) 游程指标:首先统计  $m$  组密文子序列中各组最长游程的长度,取各组最长游程的最大值  $r_{\max}$  作为游程分布特征提取的统一维度。按此维度提取各组密文子序列的各长度游程的数量,若不含有此长度的游程,则该维度记为 0。

3) 序列指标:由序列指标的定义,对每一组密文子序列提取长度为  $n = 2, 3, 4, 5, 6$  的子块数量,例如当  $n = 2$  时,在密文子序列中提取 00,01,10,11 四种类型的子块数量,构成 4 维向量,依此形成 8,16,32 和 64 维的不同维度的扑克指标,对不同维度的序列指标先进行实验对比取最高准确率作为该指标结果。

4) 近似熵指标:比较并计入相邻两组长为  $k$  ( $0 < k < L$ ) 子序列之间重叠子序列模式的个数(频数),无重叠的模式该维度记为 0。

5) 随机游动指标:将每个密文比特子块转化为累加和序列(0 变成 -1, 1 不变),对累加和序列求和,统计和为 -4, -3, -2, -1, 1, 2, 3, 4 的个数。若无该状态则记为 0。

6) 随机游动频数指标:与随机游动指标类似,构造密文子序列的累加和序列,累加和特殊状态增加为 -9, -8, -7, -6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6, 7, 8。

7) 频率指标:对每个密文子序列,统计 0,1 所占总比特数的比例,形成二维向量形式。

8) 二元矩阵秩指标:根据定义,对  $m$  组密文子序列按照每组序列组合成对应大小规格的矩阵(矩阵各元素为 0,1),计算每组矩阵的秩作为该子序列的特征指标,构成特征数据。

9) 累加和指标:同样将密文子序列转化为累加和序列,比较加和后结果的正负,为正则记为 1,否则记为 -1,当其中一个维度非零时,另一个维度记为 0。

(4) 将提取完成的特征数据及对应算法类型标签输入 4 种机器学习模型进行训练,按 6:4 分为训练集与测试集。通过调整模型参数等,对分类识别结



果进行优化, 取 10 次实验的最佳结果。训练集和测试集的明文与密钥是不同的, 对识别模型而言, 这是两套完全不同的数据集。训练集中对应标签为已知的, 而测试集中标签将被抹去。从实际应用的角度看, 训练过程可认为由己方承担进行, 测试过程可认为由敌方提供数据验证。(密码算法对应的标签由 0~6 数字标记, 如表 6 所示)

表 6 实验算法对应标签  
Table 6 The labels of algorithms

密码算法	标签
DES	0
AES-128	1
KASUMI	2
PRESENT	3
Grain-128	4
RSA	5
ElGamal	6

(5) 收集并统计各个特征数据在 4 种不同模型下的识别训练测试结果, 对多种维度类型的指标取最佳输出状态的维度为结果。

需要强调的是, 实验设计和实验流程中, 机器学习模型以数据组为最小单元进行训练和测试, 训练集和测试集每组使用的明文与密钥都是不同的, 对每组密文提取特征指标后输入机器学习模型进行实验, 因此, 这一步骤排除了明文、密钥的选择对实验结果的影响, 即明文/密钥的选择不影响特征提取。

4.2 实验结果

实验结果按使用的机器学习类型共分为 4 组, 每组收集 9 种指标在随机密钥条件下识别 7 种算法的测试集精确率与召回率, 即每种指标共有 7 个识别精确率与 7 个召回率。精确率(Precision)与召回率

(Recall)定义如下:

**定义 2.** 精确率. 精确率描述预测为正确的结果中真正正确的样本比例, 设  $TP$  为模型正确判断的正例数,  $FP$  为模型判断为正确实际错误的错例数, 则精确率为:

$$P = \frac{TP}{TP + FP} \quad (18)$$

**定义 3.** 召回率. 召回率描述模型预测正确的样本例数占全部正确样本的比例。设  $FN$  为被模型误认为是错例实则为正确样本的个数,  $TP$  同样为模型正确判断的正例数。则召回率为:

$$R = \frac{TP}{TP + FN} \quad (19)$$

精确率与召回率是描述机器学习识别模型工作效果的两个重要指标, 从“查全”与“查准”两个角度刻画识别效果。理想情况下往期望两种指标越高越好, 但实际情况中常常精确率升高则召回率有所下降, 因此实验中应当综合分析二者的情况。

4.2.1 随机森林模型实验结果

由表 7、8 和图 7、8 可知, 随机密钥加密条件下, 随机森林模型低维指标的测试集精确率集中于 15%~25%之间, 略高于平均精确率 14.3%, 块内频数指标( $F_8$ )高于低维指标, 但低于其余多维指标精确率, 其精确率分布于 30%~50%之间。序列指标等( $F_{12}, F_{13}, F_{14}, F_{15}$ )精确率大多高于 80%, 且最高能达到 99%的水平, 其效果十分显著。游程分布指标( $F_9$ )介于块内频数指标( $F_8$ )与序列指标( $F_{12}$ )等之间, 精确率分布于 70%~80%之间, 但在部分算法上, 如 AES-128、PRESENT 等的识别精确率也不低于其他高信息熵的指标。相对应的, 召回率与精确率相比有所降低, 序列指标等( $F_{12}, F_{13}, F_{14}, F_{15}$ )召回率普遍高于其他指标, 低维指标召回率与精确率相差无几。

表 7 随机森林模型精确率  
Table 7 The validate precision of random forest

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.174	0.189	0.186	0.201	0.166	0.218	0.225
$F_3$	0.195	0.184	0.191	0.234	0.192	0.235	0.244
$F_7$	0.147	0.154	0.198	0.178	0.205	0.188	0.184
$F_8$	0.386	0.361	0.420	0.386	0.414	0.365	0.371
$F_9$	0.729	0.856	0.832	0.801	0.794	0.711	0.758
$F_{12}$	0.891	0.794	0.964	0.994	0.997	0.925	0.897
$F_{13}$	0.823	0.881	0.745	0.806	0.879	0.841	0.882
$F_{14}$	0.865	0.844	0.866	0.786	0.919	0.841	0.802
$F_{15}$	0.788	0.846	0.859	0.887	0.862	0.894	0.918

表 8 随机森林模型召回率

Table 8 The validate recall of random forest

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.102	0.166	0.195	0.145	0.200	0.251	0.250
$F_3$	0.225	0.204	0.247	0.224	0.154	0.230	0.261
$F_7$	0.212	0.230	0.352	0.188	0.202	0.157	0.194
$F_8$	0.360	0.285	0.301	0.251	0.284	0.451	0.212
$F_9$	0.643	0.664	0.488	0.423	0.485	0.752	0.741
$F_{12}$	0.574	0.800	0.632	0.295	0.702	0.793	0.662
$F_{13}$	0.592	0.710	0.582	0.292	0.659	0.900	0.705
$F_{14}$	0.488	0.563	0.716	0.546	0.718	0.712	0.742
$F_{15}$	0.566	0.463	0.481	0.627	0.566	0.801	0.884

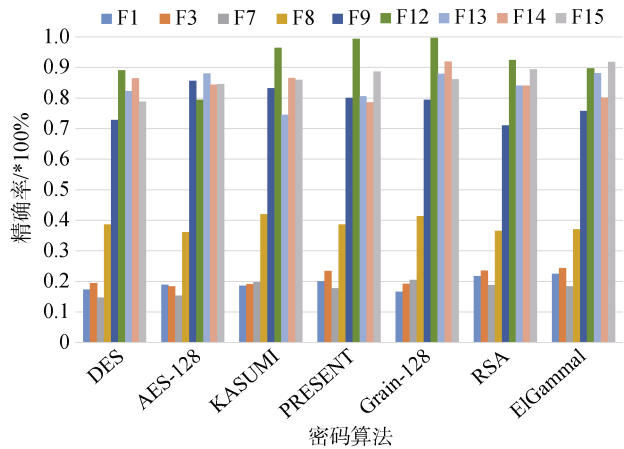


图 7 随机森林模型精确率  
Figure 7 The precision of random forest

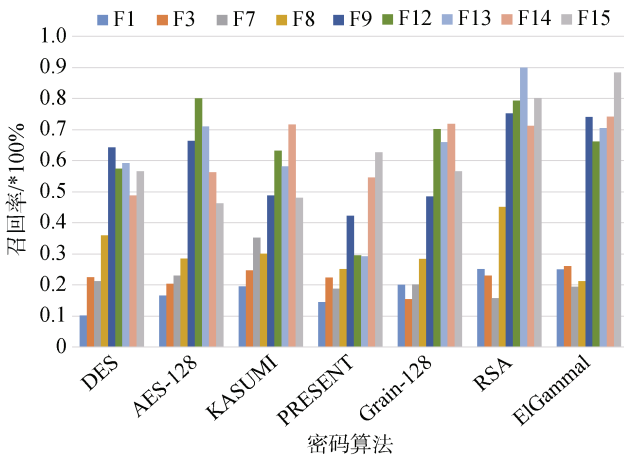


图 8 随机森林模型召回率  
Figure 8 The validate recall of random forest

4.2.2 Adaboosting 模型实验结果

由表 9、10 与图 9、10 可知, Adaboosting 模型训练效果整体高于随机森林模型, 有多种算法在不同指标下的识别精确率超过 90%。多维指标的识别精确率、召回率仍明显高于低维指标。随机游动指标

( $F_{14}$ )和随机游动频数指标( $F_{15}$ )等综合效果最佳, 对 7 种算法都能达到高于 80%的识别精确率, 且最高精确率为序列指标( $F_{12}$ )和近似熵指标( $F_{13}$ ), 达到 97.4%。对诸如 AES-128、Grain-128 等算法, 相对低熵的多维指标, 如序列指标( $F_{12}$ )也比高熵指标如随机游动指标( $F_{14}$ )等识别效果更好。低维指标的精确率与召回率分布在 20%左右, 最低识别精确率为二元矩阵指标 18.5%, 但均高于平均精确率。

4.2.3 全连接神经网络模型实验结果

由图 11、12 和表 11、12 可知, 全连接神经网络在同样条件下识别精确率并未高于前者, 多维指标的识别精确率多分布于 80%~90%之间, 最高精确率为随机游动指标 91.9%。低维指标与前两者相比也略有下降, 其精确率最小值为二元矩阵秩指标( $F_3$ )14.4%。综合精确率与召回率分布观察得知, 低维低熵指标的召回率、精确率仍低于高维高熵指标。以神经网络模型为代表的深度学习与普通机器学习模型相比, 其在较小数据集上的表现并不如意, 本文将在后续分析部分就数据量对实验结果的影响给出更细致的分析。

4.2.4 前馈神经网络模型实验结果

由图 13、14 和表 13、14 可知, 前馈神经网络模型实验结果与全连接神经网络模型相比精确率整体提高约 5%~10%, 对 AES-128、KASUMI 算法等识别效果有显著提高, 在召回率上也有所提高。同样地, 具有高信息熵的多维指标识别效果最佳, 相对的低信息熵的低维指标较差。序列指标( $F_{12}$ )的识别精确率最高, 为 94.5%。而最低精确率为 15.3%, 均高于平均精确率。指标  $F_9, F_{12}, F_{13}, F_{14}, F_{15}$  对 7 种算法的识别精确率、召回率分布在 60%以上, 低维指标  $F_1, F_3, F_7$  则主要分布于 20%左右。同样, 游程指标( $F_9$ )在 AES 等多个算法中识别精确率均不同程度地高于部分高熵指标, 最高识别精确率可达 92.1%。

表 9 Adaboosting 模型精确率

Table 9 The validate precision of Adaboosting

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.195	0.188	0.192	0.206	0.229	0.213	0.24
$F_3$	0.185	0.201	0.225	0.264	0.216	0.224	0.236
$F_7$	0.229	0.224	0.247	0.251	0.259	0.223	0.217
$F_8$	0.329	0.401	0.338	0.318	0.355	0.374	0.422
$F_9$	0.782	0.815	0.856	0.842	0.836	0.852	0.839
$F_{12}$	0.815	0.856	0.826	0.941	0.974	0.902	0.871
$F_{13}$	0.847	0.810	0.889	0.923	0.957	0.841	0.974
$F_{14}$	0.885	0.815	0.896	0.902	0.984	0.956	0.933
$F_{15}$	0.915	0.996	0.845	0.887	0.895	0.923	0.905

表 10 Adaboosting 模型召回率

Table 10 The validate recall of Adaboosting

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.150	0.157	0.157	0.260	0.252	0.232	0.235
$F_3$	0.162	0.200	0.250	0.242	0.266	0.247	0.211
$F_7$	0.194	0.245	0.272	0.211	0.299	0.231	0.273
$F_8$	0.394	0.391	0.381	0.408	0.300	0.284	0.512
$F_9$	0.582	0.745	0.712	0.742	0.660	0.72	0.719
$F_{12}$	0.759	0.626	0.746	0.741	0.854	0.900	0.671
$F_{13}$	0.747	0.681	0.729	0.523	0.769	0.805	0.854
$F_{14}$	0.682	0.715	0.669	0.592	0.881	0.814	0.713
$F_{15}$	0.591	0.810	0.547	0.697	0.802	0.726	0.885

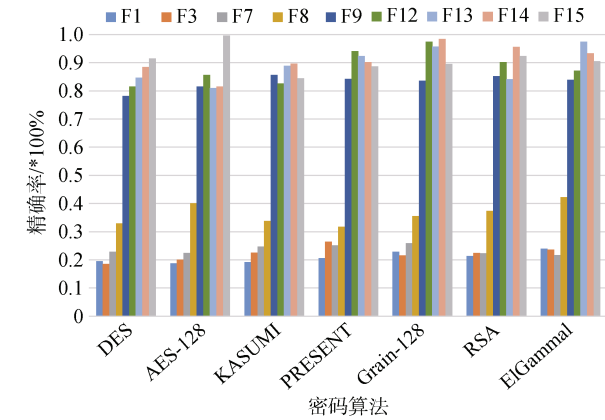


图 9 Adaboosting 模型精确率

Figure 9 The validate precision of Adaboosting

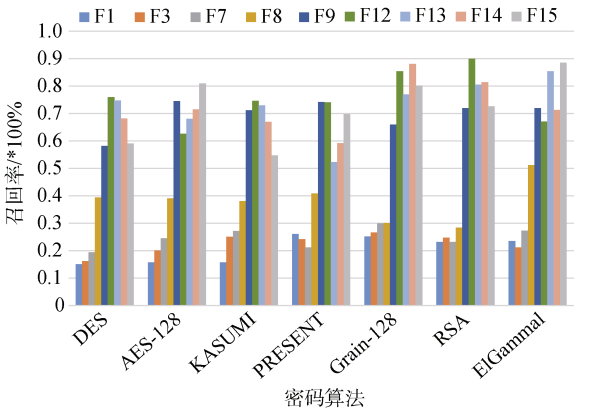


图 10 Adaboosting 模型召回率

Figure 10 The validate recall of Adaboosting

表 11 全连接神经网络识别精确率

Table 11 The validate precision of fully connected neural network

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.158	0.166	0.202	0.194	0.233	0.221	0.275
$F_3$	0.235	0.204	0.212	0.20	0.196	0.253	0.221
$F_7$	0.192	0.245	0.258	0.188	0.215	0.197	0.144
$F_8$	0.426	0.262	0.324	0.356	0.344	0.405	0.321
$F_9$	0.682	0.755	0.712	0.741	0.692	0.701	0.758
$F_{12}$	0.821	0.83	0.765	0.894	0.882	0.726	0.774
$F_{13}$	0.836	0.788	0.841	0.852	0.719	0.901	0.892
$F_{14}$	0.865	0.844	0.866	0.786	0.919	0.841	0.802
$F_{15}$	0.788	0.846	0.859	0.887	0.862	0.894	0.918

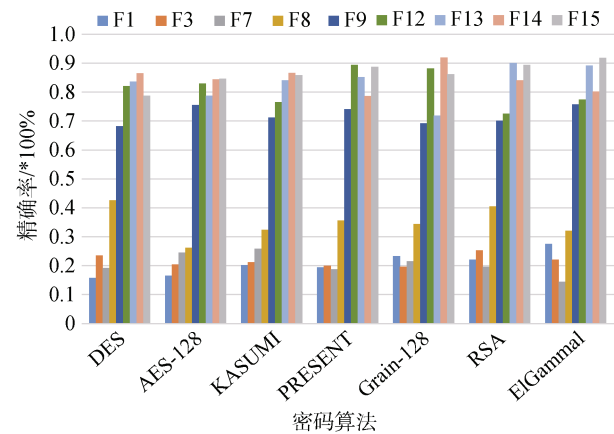


图 11 全连接神经网络模型精确率

Figure 11 The validate precision of fully connected neural network

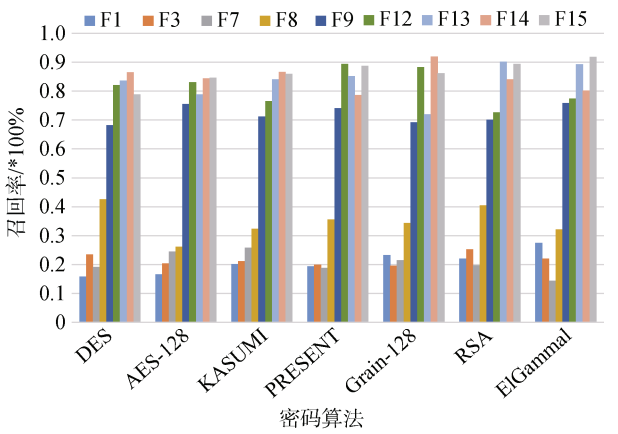


图 12 全连接神经网络模型召回率

Figure 12 The validate recall of fully connected neural network

表 12 全连接神经网络识别召回率

Table 12 The validate recall of fully connected neural network

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.144	0.140	0.220	0.152	0.212	0.232	0.150
$F_3$	0.210	0.225	0.217	0.230	0.153	0.240	0.213
$F_7$	0.174	0.236	0.243	0.151	0.200	0.157	0.146
$F_8$	0.458	0.216	0.335	0.334	0.34	0.424	0.254
$F_9$	0.512	0.619	0.700	0.643	0.612	0.755	0.450
$F_{12}$	0.459	0.715	0.652	0.664	0.742	0.778	0.582
$F_{13}$	0.723	0.739	0.629	0.802	0.659	0.800	0.691
$F_{14}$	0.745	0.602	0.62	0.710	0.723	0.814	0.592
$F_{15}$	0.611	0.582	0.719	0.615	0.596	0.852	0.711

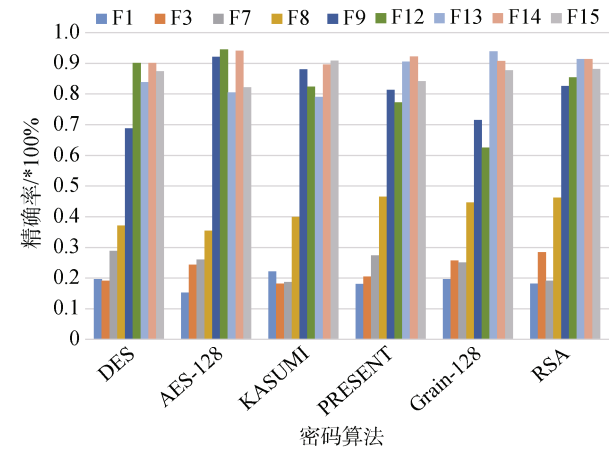


图 13 前馈神经网络模型精确率

Figure 13 The validate precision of feedforward neural network

5 实验分析

结合上文的实验与结果, 将从特征指标与模型性质等方面给出进一步的分析。通过计算实验特征

数据及其具体信息熵分布, 比较各个指标的信息熵分布关系, 解释分析不同指标的识别效果。同时, 从机器学习的模型原理角度给出不同模型识别准确度和稳定性差异的分析, 通过实验数据量和结构等方

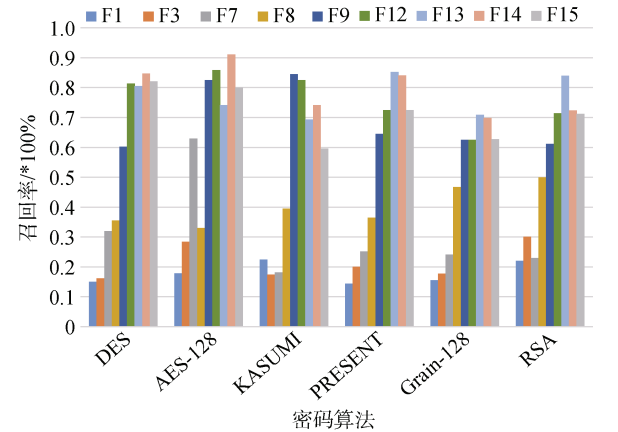


图 14 前馈神经网络模型召回率

Figure 14 The validate recall of feedforward neural network

表 13 前馈神经网络模型精确率

Table 13 The validate precision of feedforward neural network

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.197	0.153	0.222	0.181	0.196	0.182	0.276
$F_3$	0.191	0.244	0.182	0.205	0.257	0.284	0.211
$F_7$	0.289	0.260	0.187	0.274	0.251	0.191	0.173
$F_8$	0.371	0.354	0.399	0.465	0.447	0.462	0.372
$F_9$	0.688	0.921	0.881	0.814	0.715	0.826	0.648
$F_{12}$	0.901	0.945	0.824	0.773	0.625	0.854	0.819
$F_{13}$	0.839	0.805	0.791	0.906	0.939	0.914	0.893
$F_{14}$	0.902	0.941	0.896	0.922	0.908	0.914	0.932
$F_{15}$	0.874	0.822	0.909	0.842	0.877	0.882	0.910

表 14 前馈神经网络识别召回率

Table 14 The validate recall of feedforward neural network

指标	DES	AES-128	KASUMI	PRESENT	Grain-128	RSA	ElGamal
$F_1$	0.150	0.178	0.225	0.144	0.156	0.220	0.266
$F_3$	0.162	0.284	0.174	0.200	0.177	0.301	0.205
$F_7$	0.320	0.629	0.182	0.252	0.241	0.230	0.199
$F_8$	0.355	0.330	0.395	0.365	0.467	0.500	0.321
$F_9$	0.602	0.825	0.845	0.645	0.625	0.612	0.608
$F_{12}$	0.814	0.859	0.825	0.725	0.625	0.714	0.809
$F_{13}$	0.805	0.741	0.693	0.852	0.709	0.840	0.719
$F_{14}$	0.847	0.911	0.741	0.841	0.698	0.724	0.628
$F_{15}$	0.821	0.800	0.596	0.725	0.627	0.712	0.718

面的对比, 给出不同模型下各个指标识别不同密码算法的准确度现象的解释, 使得实验工作更加完善。最后, 通过实验分析给出实际工作中有效识别密码算法的结论。

5.1 特征指标信息熵及其分布

从每个指标数据量角度分析, 实验每种算法的密文数据量均为约 1 GB, 对每个密文件按实验流程提取 9 种特征后, 得到对应的特征指标数据, 具体特征数据量如下:

由表 15 可知, 尽管低维指标的数据量普遍高于多维指标, 但其准确率却并不是最高的。这说明就特征指标而言, 一定程度上数据量并不是决定准确率的绝对因素, 每种指标本身的性质等具有更加关键的作用。

为计算每种指标的信息熵并拟合其分布, 对每种指标的特征数据任意选择 1/4 数据量。计算各条指标的信息熵值, 拟合为箱线图如下。

观察图 15 可知, 除块内频数指标  $F_8$ 、游程指标  $F_9$  外, 低维指标的信息熵总体小于多维指标。频率指标  $F_1$ 、二元矩阵秩指标  $F_3$ 、累加和指标  $F_7$  与块内频数指标  $F_8$  的信息熵分布于 0.5 比特附近, 游程指标的信

息熵值在 1 比特左右, 而序列指标  $F_{12}$  等多维指标的信息熵则基本在 2 比特左右。因此, 在任意选取的数据量考查范围内, 9 种指标的信息熵大小关系基本与理论方法的结果吻合。一方面说明了筛选方法的合理性与正确性, 另一方面验证了多维高信息熵指标在区分不同密码算法的优越性能。

尽管信息熵和特征的维度在密码算法识别工作中具有十分关键的作用, 但观察实验现象可知, 并非信息熵越大的指标其识别任何密码算法的效果越

表 15 实验各特征指标数据量

Table 15 The sizes of each feature index

特征指标	数据量(条)
$F_1$	4000
$F_3$	4000
$F_7$	6000
$F_8$	3000
$F_9$	3000
$F_{12}$	3000
$F_{13}$	3000
$F_{14}$	2000
$F_{15}$	2000



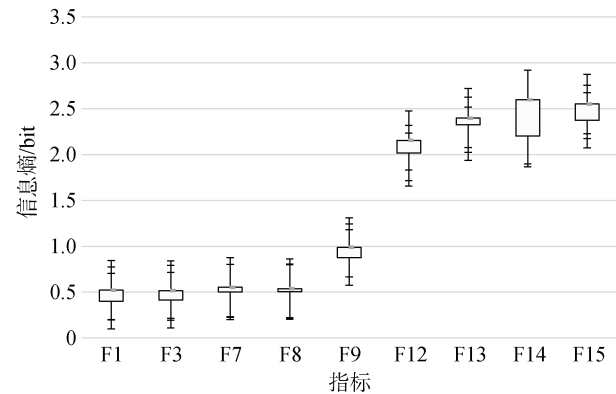


图 15 特征指标信息熵分布

Figure 15 The distributions of feature indices' entropy

好, 同样地, 信息熵相对较低的指标, 其识别效果也并非低于高熵指标。例如游程指标( $F_9$ )在低信息熵的情况下, 在 AES-128、Grain-128 等算法的识别准确度仍然能够超过高信息熵指标(如随机游动指标  $F_{12}$ 、随机游动频数指标  $F_{13}$  等)。对于实验中出现的部分特殊情况, 通过对比几种指标的具体数据形式进一步地分析, 给出更加全面的解释。

观察实验结果, 游程指标( $F_9$ )、近似熵指标( $F_{13}$ )的结果能够达到和序列指标( $F_{12}$ )、随机游动指标( $F_{14}$ )等识别准确率相同的水平, 另一方面, 随机游动指标( $F_{14}$ )、随机游动频数指标( $F_{15}$ )虽然具有较高的信息熵, 但与序列指标和游程指标相比, 其识别准确率并没有一致地大于前两者, 在部分算法(如 PRESENT、Grain-128 等)的识别实验中准确度反而更低。从维数角度看, 实验中游程分布指标的维度与最长游程长度相关, 统计全部密文所有 1 游程长度后, 得到最长游程长度为 20, 故实验中游程指标的维度被设定为 20, 同时, 通过对比序列指标的不同维度的识别效果, 实验结果中选取的最佳维度为 16。表 16 列举除块内频数指标外 6 种特征指标的维度关系。

表 16 多维特征指标维度

Table 16 The dimensions of multidimensional feature indices

特征指标	维度
$F_8$	20
$F_9$	20
$F_{12}$	16
$F_{13}$	18
$F_{14}$	8
$F_{15}$	18

表 16 说明游程指标( $F_{19}$ )和近似熵指标( $F_{13}$ )虽然信息熵较低, 但其维度弥补了这一不足。二者的维度大于其他多维指标, 故而在识别效果上, 这两种指标也并未显现出明显低于其他高信息熵指标的现象。

进一步地, 对于信息熵值较高的指标随机游动指标  $F_{14}$  和随机游动频数指标  $F_{15}$ , 通过观察提取到的特征数据各维度取值情况, 将指标各个维度的取值分布与游程指标( $F_9$ )和近似熵指标( $F_{13}$ )作对比为例, 表示为图 16、图 17。

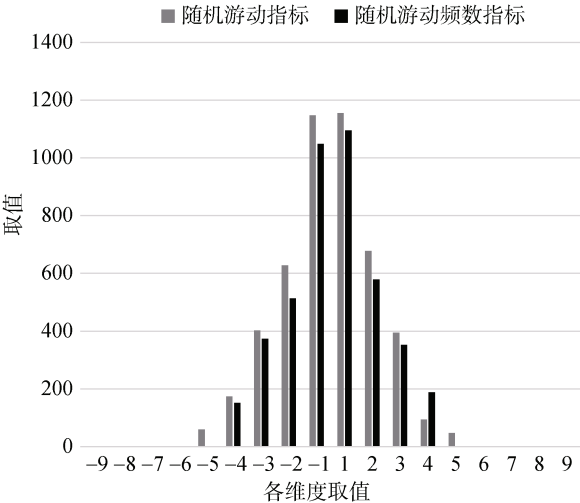


图 16 特征指标取值分布 (1)

Figure 16 The distributions of feature indices' values

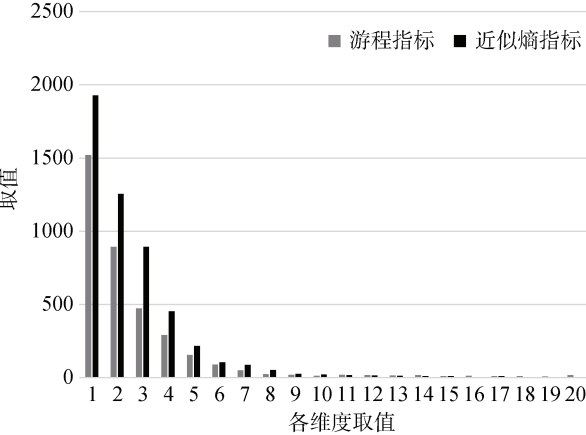


图 17 特征指标取值分布 (2)

Figure 17 The distributions of feature indices' values (2)

由图 16、17 可知, 实验中随机游动指标与随机游动频数指标的各个维度取值分布在近一半的维度中均为 0。也即这些维度的取值在实验中没有发生变化, 对于机器学习模型和神经网络模型而言, 这些固定不变的维度取值实际上并未有效影响和促进模型识别效果的提升。相比较而言, 相对低熵的游程指标( $F_9$ )与近似

熵指标( $F_{13}$ )各个维度的指标分布鲜有出现恒等于 0 的维度,大部分特征维度都起到了作用。换言之,这两种指标的有效维度实际上小于定义中所描述的维度,也小于其他两种低熵的指标。因此,这两种多维指标的有效维度远小于游程指标,故这两种指标对不同密码算法的识别准确度也并非均高于其他维度较低的指标。

综合以上分析,较为细致全面地从指标性质、信息熵与维度取值分布等方面研究了实验现象的原因,在比较了维度和信息熵大小两个方面后,本文认为两种性质均对指标的工作效果产生影响,而维度(有效维度)对指标的识别效果贡献更大,故在实际工作中选取指标时,应当综合考虑指标的性质根据需求作出适当的选择。

## 5.2 机器学习模型原理分析

除特征性质等影响因素外,机器学习模型的原理也是影响实验结果的重要因素之一。从 4 种模型的工作原理出发,分析模型架构、数据量等性质,比较不同模型的工作效果。

随机森林与 Adaboosting 模型是集成学习的两种典型模型,在数据分类与回归预测等方面具有广泛的应用。机器学习经过数十年的发展,以单一学习器为代表的传统机器学习方法不再适用,面对大数据量的高维复杂数据集,多学习器的集成学习模型效果显著<sup>[29]</sup>。通过对两种模型定义及结构的分析对比,随机森林模型可并行化训练,在大规模数据集上具有极高的效率,随机抽样的工作方法使得其泛化能力较强,且对缺失特征的数据不敏感。但随机森林模型无法适应有噪声的特征数据(特征数值被扰动),当特征取值较多时,该模型易受干扰。与之相比,Adaboosting 模型作为提升学习概念的模型,其核心思想是综合弱分类器变为更强的分类器,因此 Adaboosting 模型的稳定性比随机森林更强,实验结果观察到,Adaboosting 模型实验中 9 种指标对 DES 算法、KASUMI 算法的识别准确率值相差不超过 10%,与随机森林模型相比,准确率分布更加集中。另一方面,Adaboosting 算法注重对错误划分数据的修正,因此该算法在同样数据集的条件下,对训练数据的准确度与随机森林相比更加准确,因此,Adaboosting 算法的准确率较随机森林算法总体提升 10% 左右。另一方面,随机森林在高维度指标数据的实验中效果总体优于 Adaboosting 算法,如随机游动指标  $F_{14}$  等识别算法效果总体高于 Adaboosting 算法。这也是其擅长处理多维数据的并行结构所决定的。

与集成学习等机器学习算法相比,神经网络模型代表的深度学习技术在复杂高维数据集等场合效

果显著。决定神经网络模型的工作效果往往是数据集的大小,更多有效的数据集将大大降低神经网络训练的错误,从而调整得到在测试集上表现更好的模型<sup>[30]</sup>。实验中神经网络模型的准确度部分低于随机森林和 Adaboosting 模型,故从数据集角度分析,1 GB 数据量可能无法使神经网络模型的工作效果达到最优化。若在实验数据量的基础上,继续适当扩大数据量,研究是否能够进一步提高神经网络的准确度。为了对比不同数据集下神经网络的综合能力,选择数据量分别为 2 GB、5 GB 和 8 GB 的密文特征数据集进行对比,为方便比较两种神经网络在不同数据量下的综合效果,取输出结果的平均准确率作比较。结果如图 18 所示。

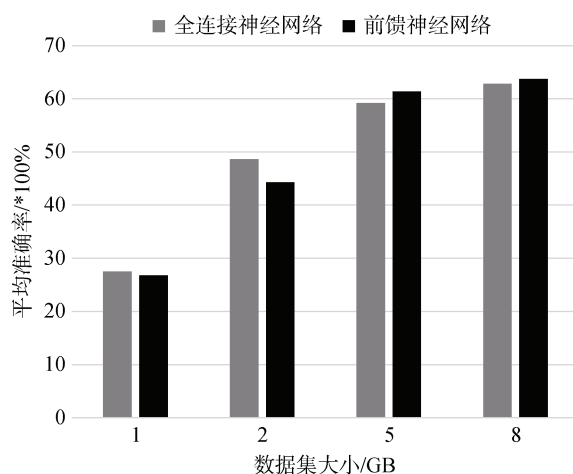


图 18 数据量与准确率的关系  
Figure 18 The relationship between data sizes and accuracy

根据图 18,随着实验数据量从 1 GB 增加到 8 GB,两种模型识别的平均准确率从 25% 的水平提升到 60% 以上,当数据量增加到一定水平时,准确率的改进不再明显。由此可知,数据量是影响神经网络模型工作效果的主要因素之一。考虑到实际应用中,从公开信道能够有效截获的密文量有限,因此分析者应当最大程度地将密文量控制在最佳范围内,使得相应模型能够同时兼顾工作效率和工作效果。

综上所述,机器学习模型及神经网络模型都能理想地完成对随机密钥条件下密码算法的识别。由于各种模型结构的差异,各个模型在识别不同密码算法时的效果有所差异。在较小数据量的条件下,随机森林、Adaboosting 等机器学习模型效果往往优于神经网络模型,而随着实验数据量的增加,神经网络对密码算法的识别能力得到增强。

因此,选择泛化能力强,擅长处理多维复杂数

据的机器学习模型, 提取多维高信息熵的特征指标并提供合适大小数据量的数据集进行识别工作是提高随机密钥条件下密码算法识别工作效果的有效方案。

## 6 总结

本文研究了随机密钥条件下唯密文的密码算法识别问题。通过计算 15 种指标在随机情况下的信息熵值, 并根据定义对指标进行维度标准下的分类。以此建立了基于信息熵值与特征维度的特征指标筛选方法, 得出结论认为具有高信息熵值的多维指标工作效果更加显著。随后选择 7 种不同标准密码算法, 建立包括神经网络模型在内的 4 种机器学习模型对其进行识别实验, 并根据筛选方法选择 6 种多维高熵的指标和 3 种低维指标进行特征提取。本文的实验结果与其他相关工作相比, 对不同算法的识别精确率提高约 45%, 且相应使用数据量减少 40% 左右, 进一步提高密码算法识别工作的效率。通过实验对比和分析, 验证了筛选方法的正确性, 从众多的指标中筛选出效果更好的指标, 提高了工作效率。从特征值分布、机器学习原理和数据量等角度对实验现象给出详细全面的分析, 补充解释了实验结果中的其他情况, 同时分析影响识别效果的其他细节因素, 使得这项工作更加完善。

基于机器学习的密码算法识别是实际工作中进行密码破译的关键技术之一, 为有效快速恢复密钥, 获取明文提供帮助, 在人工智能与密码学的交叉发展中占据关键的地位。在本文工作的基础上, 将扩展相关工作到更复杂的情况下, 对其他密码组件对象等展开识别和区分工作, 进一步提高工作实际应用价值等。

**致谢** 本论文相关工作得到国家密码科学技术重点实验室基金支持, 向实验室相关科研人员和领导表示感谢!

## 参考文献

- [1] von Solms R, van Niekerk J. From Information Security to Cyber Security[J]. *Computers & Security*, 2013, 38: 97-102.
- [2] Alani M M. Applications of Machine Learning in Cryptography: A Survey[C]. *The 3rd International Conference on Cryptography, Security and Privacy*, 2019: 23-27.
- [3] Ramzan Z. On Using Neural Networks to Break Cryptosystems[J]. *Manuscript*, 1998.
- [4] Dileep A D, Sekhar C C. Identification of Block Ciphers Using Support Vector Machines[C]. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 2006: 2696-2701.
- [5] Manjula R, Anitha R. Identification of Encryption Algorithm Using Decision Tree[M]. *Communications in Computer and Information Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 237-246.
- [6] Chou J W, Lin S D, Cheng C M. On the Effectiveness of Using State-of-the-Art Machine Learning Techniques to Launch Cryptographic Distinguishing Attacks[C]. *The 5th ACM workshop on Security and artificial intelligence*, 2012: 105-110.
- [7] Mishra S, Bhattacharjya A. Pattern Analysis of Cipher Text: A Combined Approach[C]. *2013 International Conference on Recent Trends in Information Technology*, 2013: 393-398.
- [8] Barbosa F, Vidal A, Mello F. Machine Learning for Cryptographic Algorithm Identification[J]. *Journal of Information Security and Cryptography (Enigma)*, 2016, 3(1): 3.
- [9] Zhao Z C. Research on Cryptosystem Recognition based on Machine Learning[D]. Zhengzhou: Information Engineering University, 2018.  
(赵志诚. 基于机器学习的密码体制识别研究[D]. 郑州: 战略支援部队信息工程大学, 2018.)
- [10] Li H C. Research on Cryptographic Algorithm Recognition Based on Ciphertext Features[D]. Xi'an: Xidian University, 2018.  
(李洪超. 基于密文特征的密码算法识别研究[D]. 西安: 西安电子科技大学, 2018.)
- [11] Wu Y, Wang T, Xing M, et al. Block Ciphers Identification Scheme Based on the Distribution Character of Randomness Test Values of Ciphertext[J]. *Journal on Communications*, 2015, 36(4): 150-159.  
(吴杨, 王韬, 邢萌, 等. 基于密文随机性度量值分布特征的分组密码算法识别方案[J]. *通信学报*, 2015, 36(4): 150-159.)
- [12] De Souza W A R, Tomlinson A. A Distinguishing Attack with a Neural Network[C]. *2013 IEEE 13th International Conference on Data Mining Workshops*, 2013: 154-161.
- [13] Pamidiparthi S, Velampalli S. Cryptographic Algorithm Identification Using Deep Learning Techniques[M]. *Advances in Intelligent Systems and Computing*. Singapore: Springer Singapore, 2020: 785-793.
- [14] Cao L R. Research on cryptographic algorithm recognition based on deep learning[D]. Chengdu: University of Electronic Science and Technology of China, 2021.  
(曹莉茹. 基于深度学习的密码算法识别研究[D]. 成都: 电子科技大学, 2021.)
- [15] Zaman J, Ghosh R. Review on fifteen Statistical Tests proposed by NIST[J]. *Journal of Theoretical Physics and Cryptography*, 2012, 1: 18-31.
- [16] Biau G, Scornet E. A Random Forest Guided Tour[EB/OL]. 2015: 1511.05741. <https://arxiv.org/abs/1511.05741v1>.
- [17] Cao Y, Miao Q G, Liu J C, et al. Advance and Prospects of AdaBoost Algorithm[J]. *Acta Automatica Sinica*, 2014, 39(6): 745-758.
- [18] Li J H, Li B, Xu J Z, et al. Fully Connected Network-Based Intra Prediction for Image Coding[J]. *IEEE Transactions on Image Processing*, 2018, 27(7): 3236-3247.
- [19] Koçak Y, Üstündağ Şiray G. New Activation Functions for Single Layer Feedforward Neural Network[J]. *Expert Systems with Applications*, 2021, 164: 113977.
- [20] Han S J, Oh H S, Park J. The Improved Data Encryption Standard

- (DES) Algorithm[C]. *Proceedings of ISSSTA'95 International Symposium on Spread Spectrum Techniques and Applications*, 1996: 1310-1314.
- [21] Akkar M L, Giraud C. An Implementation of DES and AES, Secure Against Some Attacks[M]. *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001: 309-318.
- [22] Kitsos P, Galanis M D, Koufopavlou O. High-Speed Hardware Implementations of the KASUMI Block Cipher[C]. *2004 IEEE International Symposium on Circuits and Systems*, 2004: II-549.
- [23] Poschmann A. Lightweight cryptography[J]. *Dalam Unievrstiy Bochum*, Bochum, 2009.
- [24] Hell M, Johansson T, Maximov A, et al. A Stream Cipher Proposal: Grain-128[C]. *2006 IEEE International Symposium on Information Theory*, 2006: 1614-1618.
- [25] Shand M, Vuillemin J. Fast Implementations of RSA Cryptography[C]. *Proceedings of IEEE 11th Symposium on Computer Arithmetic*, 1993: 252-259.
- [26] Elgamal T. A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms[J]. *IEEE Transactions on Information Theory*, 1985, 31(4): 469-472.
- [27] Gabri   M, Manoel A, Luneau C, et al. Entropy and Mutual Information in Models of Deep Neural Networks[EB/OL]. 2018: 1805.09785. <https://arxiv.org/abs/1805.09785v2>.
- [28] Azhagusundari B, Thanamani A S. Feature Selection Based on Information Gain[J]. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2013, 2(2): 18-21.
- [29] Polikar R. Ensemble Learning[M]. *Ensemble Machine Learning*. New York, NY: Springer New York, 2012: 1-34.
- [30] Schmidhuber J. Deep Learning in Neural Networks: An Overview[J]. *Neural Networks*, 2015, 61: 85-117.



夏锐琪 于 2020 年在南京大学数学专业获得学士学位。现在信息工程大学数学专业攻读数学学位。研究领域为人工智能与密码学分析。研究兴趣包括: 分组密码的设计与分析。Email: xrq\_97929@163.com



李曼曼 于 2021 年在信息工程大学密码学专业获得博士学位。现任密码研究教研室讲师。研究领域为分组密码安全性分析。研究兴趣包括: 网络空间安全, 分组密码的设计与分析。Email: limanman15@163.com



陈少真 于 2003 年在山东大学数学专业获得博士学位。现任密码研究教研室教授。研究领域为分组密码安全性分析。研究兴趣包括: 信息安全, 分组密码设计与分析。Email: chenshaozhen@vip.sina.com