ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 984－996
DOI:10. 16337/j. 1004-9037. 2022. 05. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种基于格雷码置乱与分块混沌置乱的医学影像隐私保护分类方案

陈国明[1]，袁泽铎[2]，龙　舜[2]，麦舒桃[3]

(1.广东第二师范学院计算机学院,广州 510303;2.暨南大学信息科学技术学院,广州 510632;3.广州中医药大学第二附属医院,广州 510120)

摘　要：针对传统隐私保护机器学习方案抵抗对抗攻击能力较弱的特点,提出一种基于格雷码置乱和分块混沌置乱的医学影像加密方案(Gray ＋ block chaotic scrambling optimized for medical image encryption,GBCS),并应用于隐私保护的分类挖掘。首先对图像进行位平面切割;然后,对图像不同位平面进行格雷码置乱后再进行分块,在分块的基础上分别进行混沌加密;最后通过深度网络对加密后的图像进行分类学习。通过在公开乳腺癌和青光眼数据集上进行交叉验证仿真实验,对GBCS的隐私保护与分类性能进行量化分析,并从图像直方图、信息熵和对抗攻击能力等指标考虑其安全性。实验结果表明医学图像在GBCS加密前后的性能差距在可接受范围内,方案能更好地平衡性能与隐私保护的矛盾,能有效抵御对抗样本的攻击,验证了本文方法的有效性。

关键词：隐私保护分类;对抗防御;图像分块;图像置乱;混沌

中图分类号：TP389.1　　**文献标志码**：A

A Privacy-Preserving Medical Image Classification Scheme Based on Gray Code Scrambling and Block Chaotic Scrambling

CHEN Guoming[1], YUAN Zeduo[2], LONG Shun[2], MAI Shutao[3]

(1. School of Computer Science, Guangdong University of Education, Guangzhou 510303, China;2. School of Information Science and Technology, Jinan University, Guangzhou 510632, China;3. The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510120, China)

Abstract: This paper proposes a medical image encryption scheme based on Gray code scrambling and block chaotic scrambling Gray＋block chaotic scrambling optimized for medical image encryption(GBCS), which is applied to privacy protection classification. First, the image is sliced by bit-planes. Then, different bit-planes of images are scrambled by the Gray code and then divided into blocks, and chaotic encryption is carried out on these blocks. Finally, the encrypted images are classified by deep learning network. We quantitatively analyze the privacy protection and classification performance of GBCS through cross-validation simulation on public breast cancer and glaucoma datasets, and perform a safety analysis of the method by histogram, information entropy, and anti-attack ability. The experimental results prove the

基金项目:国家重点研发计划(2019YFC0120100);广东省自然科学基金(2018A0303130169,2020A151501212);广东省普通高校重点领域专项(2020ZDZX1023,2021ZDZX1062);工业装备质量大数据工业和信息化部重点实验室开放基金(2021-IEQBD-03);广东省大数据分析与处理重点实验室开放基金(201902)。
收稿日期:2021-10-24;**修订日期**:2022-01-26

effectiveness of our method. The performance gap of medical images before and after GBCS encryption are within an acceptable range. The proposed scheme can better balance the contradiction between performance and privacy protection requirements, and effectively resist the attack of adversarial samples.

Key words: privacy preserving classification; adversarial defense; image blocking; image scrambling; chaotic

引　言

数字图像技术在互联网上广泛使用引起人们对隐私保护安全性和对抗攻击能力的关注。医学影像由于包含患者生理特征和医疗记录等大量敏感信息,若遭到攻击和篡改,容易引起医生误诊和不恰当治疗,导致严重后果,因此尤其需要重视。目前,常用于提高图像隐私保护安全性的途径有图像加密和在图像里隐藏信息等。其中,图像置乱是一种图像加密技术,常用算法包括Arnold变换、幻方算法、Hilbert曲线置乱、Conwaygame算法等。

当前机器学习特别是深度学习技术在数字图像处理领域迅速发展,成为图像处理技术的主流。基于机器学习和深度学习上的隐私保护基本上分为以同态加密方法为主的加密方法和以差分隐私保护为主的扰动方法两类。已有许多学者对这两种方法进行了研究。Talbi等[1]提出了一种端到端的隐私保护数据分类方案,允许多个数据拥有者对加密数据进行增量决策树学习而无需访问此数据的实际内容。Vani等[2]提出一种基于加密数据的K近邻(K-nearest neighbor,KNN)分类算法来确保用户在访问云上数据库时的安全性。Samanthula等[3]提出一种针对云中加密数据的KNN分类算法,该算法可以保护数据、用户的查询操作并且隐藏数据的访问模式。Wu等[4]在语义安全的混合加密云数据库上,基于同态加密设计了一种有效的KNN分类协议,既保护了数据库的安全性同时也保护了用户的查询隐私,并且整个过程对云服务器来说不可见。

放弃图像的统计特性有助于隐私保护,但容易导致深度网络对图像处理的性能显著下降,这无异于放弃深度学习的优势。为平衡图像安全性和深度网络处理性能,Fakhr等[5]提出了一种基于多个随机加密森林和两层压缩感知加密的隐私保护分类算法,在COREL1K和CIFAR10数据集上的实验表明该方案比具有明文特征的最近邻分类器的分类精度更好,而且由于类标签也被加密,云服务提供者无法得知用户数据的分类结果。Maekawa等[6]提出了一种隐私保护支持向量机(Support vector machine,SVM),能在从块置乱后的数据上提取的特征进行运算,实验表明该SVM可用于人脸识别应用。Mahmood等[7]使用混沌系统开发了一种基于随机化和加密增强矩阵运算的对称全同态加密算法。该算法能消除噪声,具有较好随机性、对初始条件的敏感性、具有较高的机密性和隐私性。

对抗样本[8-9]是近年来各种机器学习系统需要抵御的攻击类型。如果攻击者在数据加入不可感知微小扰动,那么分类器的特征受到一定程度的干扰,会导致分类器的错误分类。攻击者通过约束扰动图像和真实图像之间的距离,来实施对分类器的攻击。针对此类攻击行为,三胞胎网络从改进分类器的角度来解决这类问题,还有的攻击方法是通过在损失函数的梯度方向迭代地移动以增加损失,把扰动的图像映射到满足原始图像距离约束的输入子空间。在防御对抗攻击方面,有的方法将对抗样本和原始样本混合同时训练更强鲁棒性的分类模型,有的方法隐藏分类模型的原始梯度,还有使用随机化方法通过向模型引入随机性提高模型的鲁棒性,使其能容忍高噪声。除此以外还有的方法在输入模型判定之前,先对当前对抗样本去噪,剔除扰动的信息,使对模型的攻击失效。本文针对传统隐私保护机器学习方案抵抗对抗攻击的能力较弱的特点,依据图像自身的特征,提出一种基于通过分块局部置乱来保留图像的相邻像素的局部相关性,在置乱时引入的混沌机制,依赖混沌整体有序而局部随机特性加强鲁棒特征,并利用邻域自适应性克服深度学习普遍存在的对抗样本攻击问题。对抗样本实施攻击的一个主要因素是过度线性化,而神经网络模型主要是基于线性模块构建。因此解决深度学习的对抗

样本模型的线性特征所导致的攻击的一个方法是引入非线性,这主要通过混沌结合图像自身的特征来解决。

因此本文提出一种图像置乱加密算法(Gray block chaotic scrambling optimzed for medical image encryption, GBCS),应用于基于隐私保护的分类挖掘。该算法首先对图像进行位平面切分,然后搭配格雷码置乱、分块混沌置乱对图像进行加密。最后通过深度学习网络对加密后的图像进行分类学习以进行量化分析。本文对比了GBCS与格雷码置乱、分块混沌置乱和全局混沌置乱等置乱策略在安全性和分类性能上的表现,在公开数据集进行仿真实验的结果表明,GBCS的8个位平面分类准确率分别为65.89%、64.08%、75.83%、84.44%、89.07%、81.46%、75.17%和68.21%,平均准确率为75.51%,与未经过处理的原始图像相比平均准确率几乎没有差异(相差0.26%),但安全性得到提升,这表明GBCS既能提高医学影像的安全性也保持了一定图像分类性能。在受到对抗样本攻击,例如快速梯度攻击时,GBCS的不同位平面的分类性能说明GBCS能有效抵御对抗样本的攻击。

1 基于置乱的图像加密

数字图像的加密方法大致分为图像空间域像素置乱、变换域加密、混沌加密、秘密分割与秘密共享加密等。其中,图像置乱改变图像的直方图分布来修改图像的像素值,通过将图像的信息次序打乱使原始图像变得杂乱无章难以辨认,从而提高图像的对抗攻击能力,是一种常用的图像加密方式。

GBCS方案主要利用邻域自适应性以及随机性来平衡两方面的安全矛盾:(1)要提高医学图像加密的安全性,降低邻域像素之间的相关性;(2)利用医学图像像素的邻域相关性,防御对抗样本攻击,提高分类的准确性。图像低位平面随机性强而高位平面则结构细节完整而有序,位平面置乱,能够均匀打散噪声攻击对图像的影响。格雷码在相邻两编码之间只有一位数据发生变化,其余位状态不变,具备逻辑相邻性,抗干扰能力强。置乱后图像像素取值改变,能抵消噪声攻击,降低分类错误的几率。在分块置乱中,噪声干扰主要集中在纹理块,通过混沌整体有序而局部随机性,可以平衡对抗攻击对纹理块以及平滑块的影响,通过调节随机性、降低相关性和加强隐私性,同时控制分类精准度。如果结合两组或以上不同混沌序列,可进一步增强加密安全性。在分块的大小方面,分块大小以及块内像素之间的相关性对分类准确性的影响具有重要意义。随机性有利于加密安全,有序性有利于抗攻击分类准确性,自适应地寻求平衡点以求平衡性能与隐私保护需求的矛盾成为该方案的主要设计目标。本文创新性地提出利用GBCS抵御对抗样本的攻击,其基本思想是针对网络模型的局部线性脆弱的特点,引入混沌非线性,并利用其随机性消除干扰,原始图像和受对抗样本攻击的图像之间的差距在GBCS方案的位平面切割分离和离散化处理下,不同位平面受到扰动后的分类贡献不同,有的位平面扰动噪声被抑制,通过发掘能够抑制对抗扰动的位平面进行图像分类,可以抵御对抗样本的攻击。

1.1 GBCS方案

图1是GBCS方案的基本流程。首先采用位平面切分技术对图像进行处理,得到图像的8个位平面,并分别对每个位平面进行格雷码置乱。在此基础上,采用图像分块算法对所有图片进行处理,并对每个子块进行混沌序列置乱,形成加密后的图像数据集。

GBCS首先对分类图像进行位平面切割。位平面是数字图像的重要属性,进行位平面切分后再对图像进行置乱可以通过改变图像的像素值抵御针对单像素置乱的已知明文攻击,高位位平面能携带更多图像原始的统计特性和重要特征,可以更好地对图像进行描述和表示。图像经过位平面分解后再进行各种操作和运算可以扩展密钥空间。

在位平面切分的基础上,采用格雷码置乱、分块混沌置乱和全局混沌置乱3种方法进行图像置乱并进行对比。由于改变了图像的直方图,提高了图像的信息熵同时增强了图像抗攻击能力,这3种方法比其它图像置乱方法具有更高安全性。

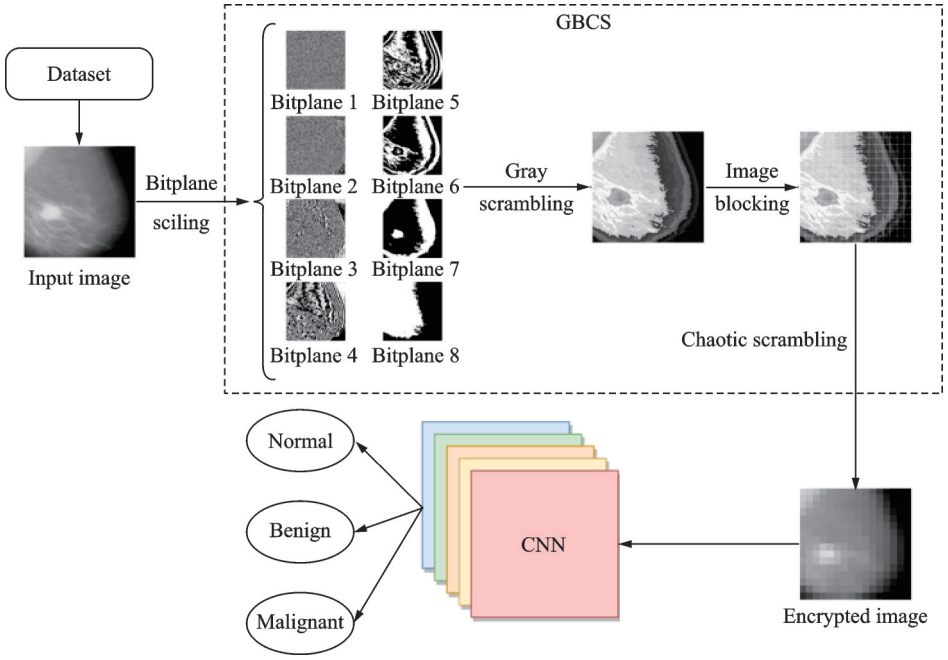

图1 GBCS方案的流程

Fig.1 Pipeline of GBCS

医疗图像的安全对于患者安全和保密非常重要。本文提出的基于GBCS方案的医疗图像加密保护分类方案。该方案中,客户端的医护人员将加密图像发送到云端,另一个客户端的医护人员从云端接收加密后的图像进行解密。在云端服务器构建的卷积神经网络(Convolutional neural network,CNN)等深度学习网络可以被客户端的数据分析人员用来对经过隐私保护后的加密图像进行分类与识别。

1.2 格雷码置乱

格雷码置乱是常用的置乱技术,与普通二进制编码相比,格雷码的特点是相邻二进制码之间只有一个码元发生变化,即只有一个比特位不同(这通常被称为"逻辑连接"),同时也能进一步扩展密钥空间。

格雷码是一种普通的二进制通信编码格式,两个相邻二进制码之间只有一位码元不同,从第j位开始,自然二进制码到格雷码的转换方式如下

$$\begin{cases} G(i)=B(i) & i=j-1 \\ G(i)=B(i+1)\oplus B(i) & 0<i<j-1 \end{cases} \tag{1}$$

式中:$G(i)$和$B(i)$分别为转换后的N位格雷码和转换前的自然二进制码的第i位,其中"⊕"表示异或操作。

$$\begin{cases} B(i)=G(i) & i=N-1 \\ B(i)=G(i)\oplus B(i+1) & 0<i<N-1 \end{cases} \tag{2}$$

图像位平面的像素值被格雷码转换之后的值所代替,根据图像的大小计算得到首次加密的格雷码迭代转换次数r_0

$$r_0=\mathrm{mod}((L+W),7)+1 \tag{3}$$

式中L和W分别为图像的长度和宽度。

一幅图像(实验样本图像提取出的感兴趣区域)经格雷码置乱后的变化如图2所示,由于原始影像

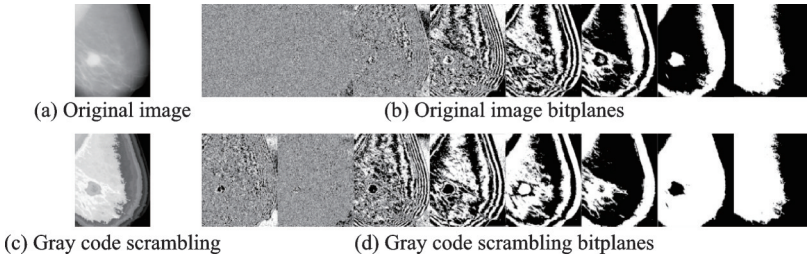

图2 格雷码置乱及其位平面

Fig.2 Gray code scrambling and its associated bitplanes

及其位平面图像的每个位置的像素值与未置乱前相比已发生改变,导致图像产生肉眼可见的明显变化。

格雷码置乱的不足在于它具有周期性:图像在经过周期性置乱后会恢复成原始图像,这降低了抗攻击能力。因此需要研究具有更高抗攻击能力的置乱方法。

1.3 全局混沌置乱

全局混沌置乱是一种全局全方位的置乱策略,它利用了混沌现象(出现在非线性动态系统中的伪随机过程)所具有的非周期性、遍历性、伪随机性以及对初始条件和结构高度敏感等特点以及天然的信息隐藏能力,被广泛用于信息置乱[10]。

混沌系统产生的混沌序列局部保留了随机性,利用这个特征可能使得构造置乱图像的过程中关键的特性信号被保留下来,可用于对图像进行描述和表示。混沌系统所表现出复杂动力学特征大大增强了系统的随机性,同时具有更复杂的相空间,契合图像的加密过程。全局像素的置乱方法降低了图像相邻像素点以及图像各个组成部分之间的相关性,可以取得良好的置乱效果。

图像进行全局混沌置乱加密之后的图像是无序的、均匀的,全局混沌置乱对密钥有很强的敏感性,如果密钥错误,经过解密后的图像与明文图像完全不符,无法获得正确的明文内容。

1.4 分块混沌置乱

注意力机制领域的研究表明图像在深度网络中的性能表现多与图片的某些局部特征相关联[11]。在对图像的分类对抗攻击中对图像的某些局部结构加入细微的噪声扰动可能使深度网络作出截然相反的判断。显然,充分考察原始图像的局部变化具有重要意义。图像分块有助于更好地获取局部图像特征,排除多余信息干扰,从而准确处理细节特征。

有鉴于此,GBCS提出先对格雷码置乱后的图像进行分块处理然后在各个子块上分别进行混沌置乱的方案。该方案的基本步骤如下:

(1)读取要置乱的数字图像,获取其尺寸大小M和N;

(2)按照预先设定好的分块大小将图像划分成各个块;

(3)根据密钥x_0(即初始值),产生一个混沌序列;

(4)通过对混沌序列进行排序来获得有序序列,从而生成下标序列;

(5)融合像素矩阵的R、G、B三个通道的分量矩阵,获得一个像素矩阵;

(6)根据下标序列将每个像素坐标变换到一个新位置,得到一个置乱之后的像素矩阵;

(7)将像素矩阵分离为R、G、B三个通道的分量矩阵,得到一个新的置乱后的像素矩阵,根据新的像素矩阵即可生成置乱后的彩色数字图像。

该置乱算法具有可逆性。解密算法使用和加密算法相同的密钥,但过程相反。运用该算法对图像进行分块混沌置乱的结果如图3所示(其中将图像按照16×16进行分块)。从图3中可以看出:对图像进行

分块后再对块内像素进行置乱的图像置乱过程降低了相邻元素和图像组成成分的相关性,安全性能得到大幅度提高。图3(a~h)表示将原始图像切分成8个位平面后进行分块混沌置乱;图3(i)表示原始图像经过分块混沌置乱之后的图像;图3(j)表示原始医学影像。

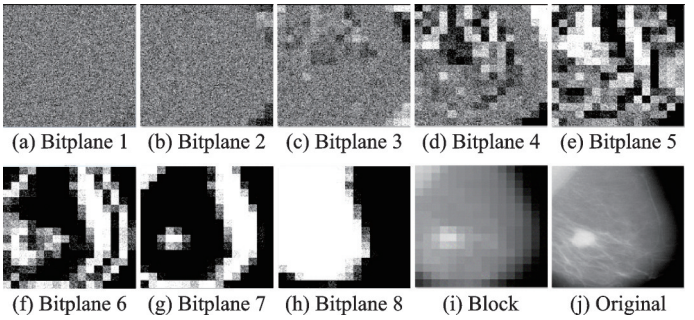

图3 分块混沌置乱以及位平面

Fig.3 Block chaotic scrambling and its associated bitplanes

此外,为了验证GBCS方案的泛化性能,实验中将GBCS与其他置乱方案应用于不同类型的医学图像中,结果如图4所示。图4(a)为乳腺医学影像,图4(b)为青光眼医学影像。从左到右依次是原始图像、格雷码置乱、分块混沌置乱、GBCS(本文所提方案)、全局混沌置乱等不同方案置乱的结果,实验证明GBCS具有较好的适应性,能够应用于不同类型的医学图像上。

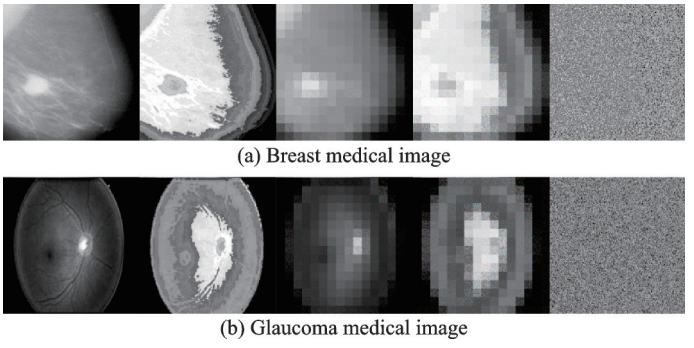

图4 不同置乱方案应用于不同医学图像的结果图

Fig.4 Results of different scrambling schemes applied to different medical images

本文针对传统隐私保护机器学习方案抵抗对抗攻击的能力较弱的特点,设计出GBCS方案,通过深度学习网络对加密后的图像进行分类学习。本文通过设计在公开医学数据集上进行交叉验证仿真实验对GBCS的隐私保护与分类性能进行量化分析,从图像直方图、信息熵、对抗攻击修复能力等指标考虑其安全性,并对比它与其他加密方法的分类表现。为了验证GBCS抵御对抗样本攻击的能力,还设计了原始图像受到快速梯度攻击后,GBCS不同位平面防御对抗攻击的分类性能实验,展现其有效防御添加对抗扰动噪声导致的模型分类错误。

2 实验结果与分析

2.1 实验方法

针对如何更好利用邻域自适应性以及随机性来平衡两方面安全的矛盾,即加强医学图像隐私安全性又同时防御攻击,提高分类的准确性。本文设计如下实验量化地分析上述各个图像加密方案的优劣,本文基于乳腺X线图像分析学会(Mammographic image analysis society, MIAS)乳腺摄影图像数据集对各个方案进行测试,MIAS中每幅原始图像大小为1 024×1 024,利用最小矩阵原则从原始图像中裁剪得到主要的乳腺部分并去除图像中无关的摄像标签后得到本文所用数据集。其中根据病变情况将乳腺钼靶X摄像分为3类:正常(Normal,N)共205幅、良性(Benign,B)共67幅、恶性(Malignant,M)共50幅。另一个青光眼数据集ORIGA,包含650张由专业医生标注的眼底图,分为两个类别:其中正常(Normal)共483幅,青光眼(Glaucoma)共167幅。由于医学影像难以获取,数量较少,容易出现过拟合的问题,在训练过程中使用数据增强技术,通过随机的错切变换、水平翻转、随机缩放等方式扩充数据集。

首先对加密方案进行图像安全性分析,接着研究方案对深度网络分类的性能影响,前者通过图像直方图、信息熵和对抗攻击能力3个指标进行衡量。图像加密通过将直方图进行均衡化和泛化降低从图像直方图中提取特征的可能性,图像直方图越均匀,加密性能越好。信息熵是测量随机变量不确定

性的一种度量，由于图像是一种多维的数据结构，因此本文使用二维信息熵来描述图像灰度层面、空间信息的混乱程度，定义如下

$$\begin{cases} P_{i,j}=f(i,j)/(W\cdot H) \\ E=-\sum_{i=0}^{255}\sum_{j=0}^{255}P_{i,j}\log_2 P_{i,j} \end{cases} \tag{4}$$

式中：W和H分别为图像的宽和高；(i,j)为一个二元组，i表示某个滑动窗口内中心的灰度值，j为该窗口内除了中心像素的灰度均值；$f(i,j)$表示(i,j)二元组在整个图像中出现的次数；E表示信息熵。二维熵将空间信息纳入了计算，同时考量了周围的灰度信息，反映图像像素位置的维度信息和像素领域内灰度分布的综合特征，熵值越大代表图像越混乱。对加密算法而言，熵值越大意味着加密后的图像安全性更高。而方案的抗攻击能力通过对加密后图像进行仿真模拟攻击（包括低通滤波、直方图均衡化、重缩放、旋转等）进行验证。

在对抗攻击能力提升的情况下，图像必须保持原有图像一定数目的统计特征才能使其在保护图像隐私的基础上满足图像分类、识别以及各种数据挖掘的任务需求。本文通过随机地将数据集的90%、10%划分为训练集和验证集进行交叉测试来分析上述各加密方案对深度网络处理的性能的影响，即将训练集通过不同置乱处理得到新数据集，各自在同一深度网络中经过80轮迭代后用验证集测试，以验证集上的Top-1准确率（为分类正确的样本数与样本总数的比值，通常来说准确率越高深度网络的性能越好）进行衡量。基准值是将原本的医学影像仅经过位平面切分后形成的训练样本集和测试样本集的分类准确率。

此外，本文在实验中发现分块的大小对于图像在深度网络中的分类性能表现有显著影响。为此，本文将图像分别按照8×8、16×16、32×32、64×64、128×128、256×256不同大小做分块，分别对每个块进行混沌置乱后再输入同一个深度网络中考察对分类性能的影响，并进行对比分析。为了展现相邻像素的局部相关性抵御对抗样本攻击的有效性，在受到对抗样本攻击后，通过设计GBCS的不同位平面在对抗样本攻击后的分类性能来验证本文的方案抵御对抗样本攻击的有效性。

2.2 安全性分析

图像直方图可以表示数字图像中像素值的分布，标绘了图像中每个亮度值的像素数，GBCS对图像分块之后进行混沌置乱，为了验证不同混沌系统置乱的效果，本文对同一图像（按照16×16分块，以下同）采用不同混沌系统进行置乱后的图像直方图进行分析，实验结果如图5所示。可以看出在GBCS中运用结合Picecwise和Tent混沌系统进行加密之后，图像直方图分布均匀，加密性能较好。运用3CNN

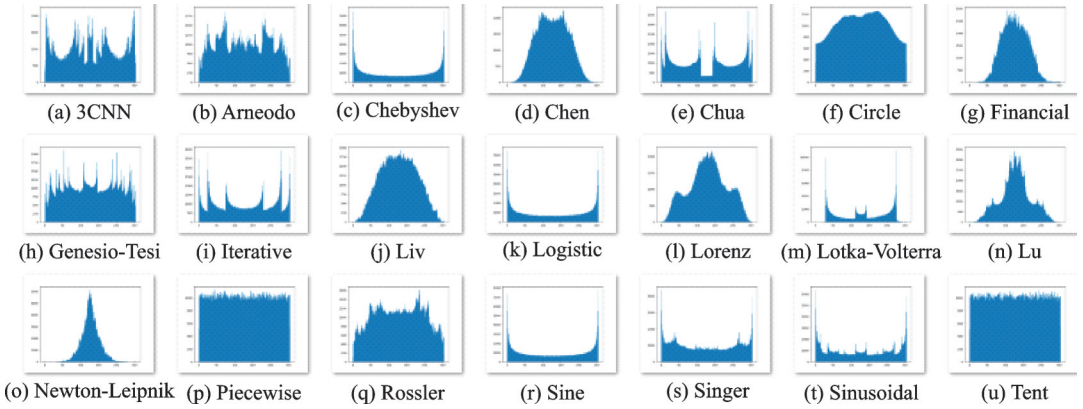

图5 各种不同的混沌系统置乱的直方图对比

Fig.5 Comparison of different chaotic scrambling histograms

等其他混沌系统进行加密后图像直方图分布不均衡,置乱后的加密性能较差,故在GBCS中,本文采用结合Piecewise和Tent作为混沌置乱阶段的加密策略。Piecewise tent混沌映射的迭代公式定义如下

$$x_{n+1}=\begin{cases}4x_n & 0\leqslant x_n\leqslant\dfrac{1}{4}\\ 2-4x_n & \dfrac{1}{4}\leqslant x_n<\dfrac{1}{2}\\ 4x_n-2 & \dfrac{1}{2}\leqslant x_n<\dfrac{3}{4}\\ 4(1-x_n) & \dfrac{3}{4}\leqslant x_n\leqslant 1\end{cases}\tag{5}$$

图6(a)为具有分段线性的Piecewise tent混沌映射,其具备短周期性。图6(b)为对Piecewise tent混沌系统进行扰动,该方案对Piecewise tent分段线性函数进行扰动。通过对混沌系统进行扰动,序列具备良好的随机性能,避免产生的序列退化为0。经过扰动后的序列频数检验结果为0.001 3,平衡度为0.036,序列具有良好的随机性。初始条件发生微小改变后,位变化率为0.46,Piecewise tent具备良好的初值敏感性。

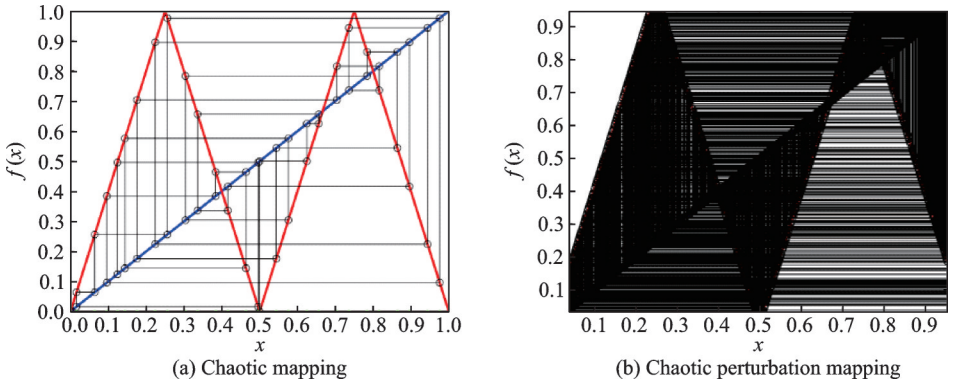

图6 Piecewise tent混沌映射的函数曲线

Fig.6 Function curves of chaotic Piecewise tent

表1列出经过上文所提及的3种置乱策略以及本文所提出的GBCS置乱后的图像信息熵值均比原图要大,可见图像的安全性都有所提高,以全局混沌置乱方案表现最佳。

表1 不同置乱算法的信息熵比较

Table 1 Comparison of information entropy on different scrambling methods

Original	Gray	Block	Chaotic	GBCS
7.314 8	7.405 1	7.410 3	7.926 3	**7.631 1**

注:加粗字体为本文算法结果。

图7展示GBCS应对不同仿真模拟攻击(包括低通滤攻击、直方图均衡化攻击、重缩放攻击、旋转攻击、噪声攻击等)的对抗攻击能力。可见被攻击后的图像均可以较好地恢复(原始信息基本得到保留),这表明GBCS具有较好的对抗攻击能力。

2.3 分类性能对比

本文通过用CNN深度网络处理经过了上述各个置乱加密处理后的图像来比较它们在完成分类任务的性能差异,测试方法如前2.1所述。本文所采用的CNN分类神经网络结构如图8所示,简化的网路包含了3层卷积层,3层池化层以及输入输出层和两层全连接层,激活函数用sotftmax函数。网络参数为Conv2D(32,(3,3))、Conv2D(64,(3,3))以及MaxPooling2D(2,2)等,神经网络超参数Batchsize为20,Epochs为80,Dropout为0.5。损失函数使用交叉熵损失函数,自适应学习率优化算法采用RM-

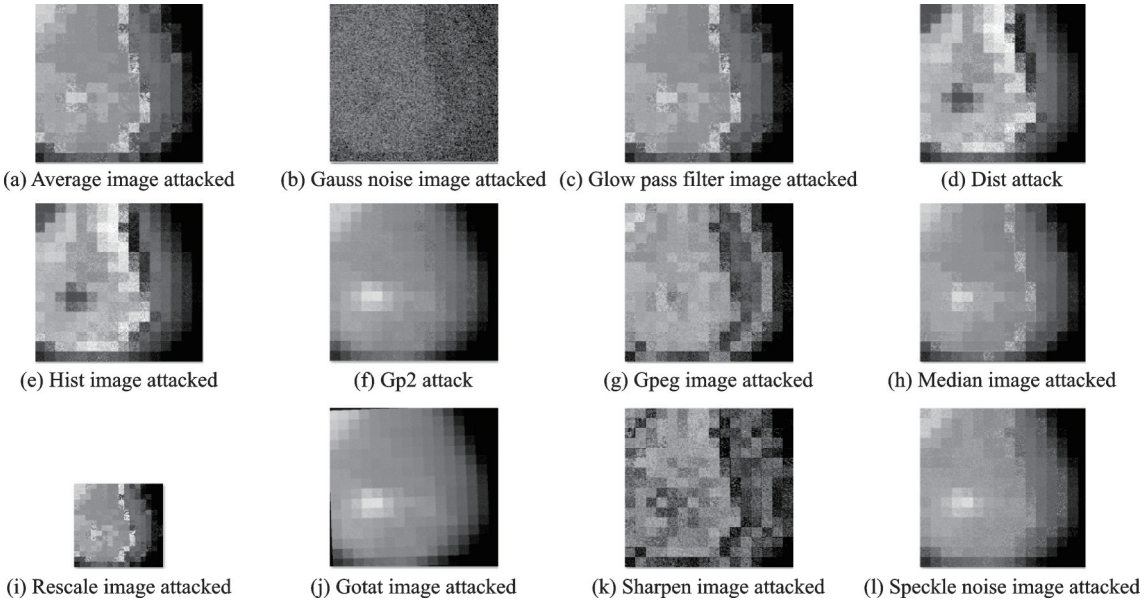

图7 GBCS加密后应对不同仿真模拟攻击

Fig.7 Anti attack analysis in various ways

SProp。乳腺数据集原始图像8个位平面的准确率分别为65.89%、61.59%，67.55%、74.17%，89.08%、92.05%、87.81%和63.91%，平均准确率为75.25%。将经过格雷码置乱(Gray)、经过分块混沌置乱(Block)、经过全局混沌置乱(Chaotic)以及经过GBCS等4种不同置乱策略处理后的测试图像在8个位平面上的分类表现进行对比，结果如表2所示。表2显示经过格雷码置乱与GBCS之后的乳腺图像分类表现与原始图像最为相似，8个位平面平均准确率相比原始图下降差距在一个百分比内。虽然格雷码置乱保留了与原始图像相似的分类性能表现，但如前文讨论所述，进行格雷码置乱带来的安全性提升并不高。全局混沌置乱与原始图像分类性能表现差异最大，8个位平面的准确率相比原始图平均下降11.50%，表明图像经过全局混沌置乱之后，虽然安全性提升明显(直方图分布均匀，二维信息熵提高，肉眼观察相比原始图像也有巨大的差异)，但分类性能损失较大。如何在保留图像统计特征与提高图像安全性中取得一个良好平衡是亟待解决的问题。从表2中可以看出，经过GBCS加密之后的8个位平面的平均准确率与原始图像相差0.26%，处在一个可以接受的范围内，但安全性显著提升。与具有安全性更高和置乱效果更好的全局混沌置乱相比，GBCS的方式保留了较高的分类性能，可以更好地应用于当前流行的各种深度网络中，是一种在分类性能和安全性中取得较好平衡的方案。

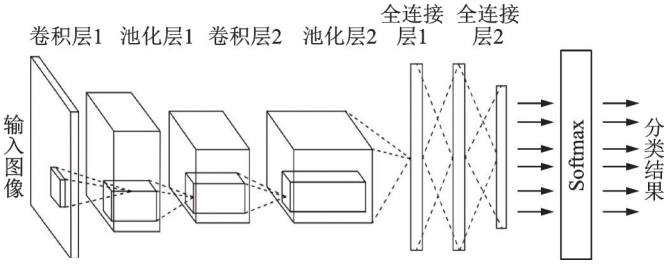

图8 神经网络结构

Fig.8 Neural network structure

表3展示了另一个青光眼医学图像数据集上的实验结果。青光眼数据集原始图像8个位平面的准确率分别为74.77%、74.31%、75.38%、67.08%、81.38%、84.92%、83.54%、80%，平均准确率为77.67%。将经过格雷码置乱、分块混沌置乱、全局混沌置乱以及GBCS等4种不同置乱策略处理后的

表2 基于不同置乱方法加密的位平面CNN分类结果比较(乳腺数据集)

Table 2 Comparison of CNN classification results based on different scramble method encryptions(MIAS)

不同置乱策略	精确度				损失			
	Gray	Block	Chaotic	GBCS	Gray	Block	Chaotic	GBCS
Bitplane 0	0.612 6	0.642 4	0.637 3	0.658 9	0.864 5	0.892 1	0.929 7	0.879 6
Bitplane 1	0.609 4	0.640 8	0.631 3	0.640 8	0.881 4	0.896 5	0.922	0.844 3
Bitplane 2	0.721 9	0.633 8	0.628 1	0.758 3	0.710 9	0.852 0	0.919 8	0.603 7
Bitplane 3	0.871 9	0.678 8	0.615 9	0.844 4	0.308 8	0.755 5	0.944 6	0.426 3
Bitplane 4	0.884 1	0.721 9	0.645 7	0.890 7	0.301 9	0.601 6	0.885 5	0.263 0
Bitplane 5	0.781 3	0.831 1	0.643 8	0.814 6	0.614 1	0.460 7	0.902 1	0.442 1
Bitplane 6	0.838 0	0.775	0.665 6	0.751 7	0.470 2	0.516 5	0.874 5	0.584 3
Bitplane 7	0.646 9	0.748 3	0.629 1	0.682 1	0.866 2	0.641 9	0.919 5	0.762 1
位平面离差	−0.006	−0.043	−0.115	**0.002 6**	0.023 9	0.098	0.308 9	**−0.002**

注:加粗字体表示本实验中的最优值;位平面离差表示不同置乱策略下8个位平面实验结果的平均值与未加密图像(即基准值)8个位平面结果平均值的差异。表3同。

表3 基于不同置乱方法加密的位平面CNN分类结果比较(青光眼数据集)

Table 3 Comparison of CNN classification results based on different scramble method encryption(ORIGA)

不同置乱策略	精确度				损失			
	Gray	Block	Chaotic	GBCS	Gray	Block	Chaotic	GBCS
Bitplane 0	0.750 0	0.750 0	0.742 3	0.743 8	0.602 1	0.568 8	0.566 0	0.570 5
Bitplane 1	0.749 8	0.750 0	0.713 0	0.742 3	0.587 8	0.561 9	0.564 5	0.719 6
Bitplane 2	0.743 8	0.750 0	0.743 8	0.750 0	0.591 2	0.571 1	0.562 5	0.568 0
Bitplane 3	0.765 6	0.819 4	0.731 5	0.843 8	0.667 0	0.493 8	0.566 1	0.267 0
Bitplane 4	0.741 8	0.750 0	0.743 8	0.734 4	0.765 4	0.807 0	0.562 8	0.691 5
Bitplane 5	0.734 4	0.750 0	0.750 0	0.750 0	0.749 4	0.564 8	0.563 6	0.535 7
Bitplane 6	0.656 3	0.740 9	0.750 0	0.765 6	0.779 9	0.578 9	0.575 1	0.559 9
Bitplane 7	0.730 5	0.718 8	0.745 4	0.718 8	0.887 1	0.703 6	0.569 7	0.604 3
位平面离差	−0.042 7	−0.023 1	−0.036 75	**−0.021 41**	**0.058 6**	−0.163 9	−0.203 9	−0.205 6

测试图像在8个位平面上的分类表现进行对比,结果如表3所示,经过分块混沌置乱与GBCS之后的青光眼图像分类表现与原始图像最为相似,8个位平面平均准确率相比原始图相差约2%。但分块混沌置乱带来的安全性并不高。经过格雷码置乱与原始图像分类性能表现差异最大,8个位平面的准确率相比原始图平均下降4.27%。用4种不同置乱策略处理青光眼数据集,其8个位平面上的分类表现相差不是很大,GBCS的方式保留了较高的分类性能。经过GBCS加密之后的8个位平面的平均准确率与原始图像相差2.14%。

2.4 分块大小对分类性能的影响

上述实验结果表明1.1节中提出的GBCS策略是一种较好的折衷方案。考虑到方案在同一张图片不同分块大小应用时对分类性能所带来的影响,应选择适当的分块大小来获得最佳性能。为此,按照GBCS中的流程,在经过位平面处理对图像进行格雷码置乱后,对图像进行8×8、16×16、32×32、64×64、128×128、256×256分块,并对每个分块做基于混沌序列的加密,加密后的图像在同一个深度网络

中迭代80次之后的分类表现如表4所示。从表4的实验结果可见，对乳腺图像实施分块混沌置乱时，先将其按32×32的大小分块，再对每个子块混沌置乱的方法所取得的测试集准确率最高（达到84.38%），图像分类性能最好，而对其他8×8、16×16等大小分块进行混沌置乱后图像的分类性能表现提升并不明显。

从表5的实验结果可见，对于青光眼图像进行分块混沌置乱时，先将其按16×16的大小分块，再对每个子块混沌置乱的方法所取得的测试集准确率最高（达到75.2%），图像分类性能最好。

表6的实验结果展现该方案在不同结构深度网络的分类结果。从表6的实验结果可见，对于乳腺图像进行分块混沌置乱时，先将其按256×256的大小分块，再对每个子块混沌置乱的方法所取得的测试集准确率在Alexnet深度网络最高（达到65.62%）；将其按8×8的大小分块再进行混沌置乱所取得的测试集准确率在Googlenet深度网络最高（达到64.34%）；将其按128×128的大小分块再进行混沌置乱所取得的测试集准确率在Inceptionv3深度网络最高（达到64.58%）；将其按32×32的大小分块再进行混沌置乱所取得的测试集准确率在Mobilenet深度网络最高（达到68.75%）。

表4 不同分块大小的分类性能对比（乳腺数据集）

Table 4 Experimental results of different block sizes (MIAS)

Block size	Train accuracy	Test accuracy	Train loss	Test loss
8×8	0.627 5	0.634 4	1.033	0.900 5
16×16	0.632 8	0.635	0.931 6	0.897 9
32×32	**0.820 3**	**0.843 8**	**0.514 1**	**0.386 1**
64×64	0.643 8	0.640 6	0.872 1	0.935 4
128×128	0.637 3	0.637 5	0.875 3	1.113
256×256	0.637 3	0.637 5	0.918 5	0.905 2

注：加粗字体表示分块实验中最优值，表5、6同。

表5 不同分块大小的分类性能对比（青光眼数据集）

Table 5 Experimental results of different block sizes (ORIGA)

Block size	Train accuracy	Test accuracy	Train loss	Test loss
8×8	0.704 1	0.750 0	0.594 2	0.561 8
16×16	0.724 0	**0.752 0**	0.582 3	0.570 0
32×32	**0.742 5**	0.750 0	0.581 7	**0.559 5**
64×64	0.706 2	0.750 0	**0.573 4**	0.559 6
128×128	0.724 2	0.744 2	0.579 0	0.564 3
256×256	0.734 4	0.748 1	0.598 5	0.564 9

表6 不同分块大小的分类性能对比（不同网络）

Table 6 Experimental results of different block sizes (different networks)

Block size	Alexnet	Googlenet	Inceptionv3	Mobilenet
8×8	0.635 4	**0.643 4**	0.620 2	0.645 8
16×16	0.614 6	0.620 2	0.635 4	0.625 0
32×32	0.635 4	0.635 7	0.614 6	**0.687 5**
64×64	0.625 0	0.635 7	0.635 4	0.656 3
128×128	0.635 4	0.612 4	**0.645 8**	0.583 3
256×256	**0.656 2**	0.627 9	0.635 4	0.656 2

2.5 防御对抗攻击的分类性能

对乳腺原始图像实施快速梯度符号方法（Fast gradient sign method，FGSM）生成图像的对抗样本，该攻击利用CNN网络中梯度变化的最大方向施加扰动噪声，使得模型分类错误以实施攻击。图9是乳腺原始图像与对原始图像实施不同强度FGSM对抗攻击生成的对抗样本图像。第1列从上到下依次为原始图像、对应原始图像的第8、第7、第6、第5位平面；第2列为对原始图像用强度系数EPS＝12/255的FGSM攻击后得到的对抗样本图像及其对应的位平面；第3列为对原始图像用强度系数EPS＝16/255的FGSM攻击后得到的对抗样本图像及其对应的位平面；第4列为对原始图像用强度系数EPS＝24/255的FGSM攻击后得到的对抗样本图像及其对应的位平面；第5列为对原始图像用强度系数EPS＝32/255的FGSM攻击后得到的对抗样本图像及其对应的位平面。

图9展现出乳腺原始图像与实施FGSM对抗攻击后的对抗样本图像及其相对应的位平面变化的情况。原始图像和对抗样本图像在GBCS方案的位平面切割分离下，其不同位平面抗扰动噪声的能力不

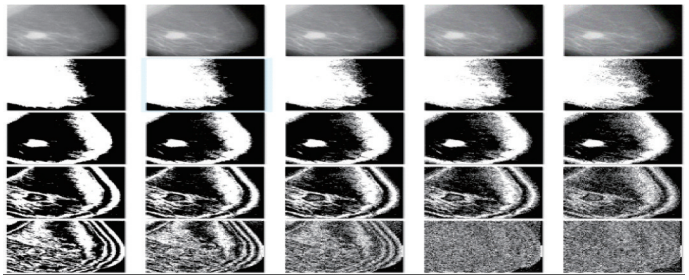

图9 FGSM对原始乳房图像的对抗性攻击

Fig.9 FGSM adversarial attack on original breast image

同,有的位平面扰动被抑制。对比对原始图像实施一定强度的FGSM攻击后得到对抗样本分类器的准确率与对其实施GBCS策略后各个不同位平面的分类准确率,从表7的乳腺数据集实验结果可见,在受到强度EPS=32/255的FGSM攻击时,准确率在第7位平面最高达到89.23%;在受到强度EPS=24/255的FGSM攻击时,准确率在第7位平面最高达到89.56%;在受到强度EPS=16/255的FGSM攻击时,准确率在第7位平面最高达到90.91%;在受到强度EPS=12/255时,准确率在第5位平面最高达到90.94%。实验结果表明,GBCS位平面切割能准确识别对抗样本真实类别,对攻击表现出良好分类性能。

表7 不同位平面抵御FGSM攻击的分类准确率对比

Table 7 Comparison of classification accuracy of different bitplanes against FGSM attacks

攻击强度	32/255	24/255	16/255	12/255
FGSM攻击	0.636 4	0.643 1	0.636 4	0.636 4
Bitplane 0	0.649 8	0.633 0	0.646 5	0.643 1
Bitplane 1	0.643 1	0.626 3	0.629 6	0.629 6
Bitplane 2	0.693 6	0.639 7	0.643 1	0.626 3
Bitplane 3	0.643 1	0.643 2	0.683 5	0.895 6
Bitplane 4	0.767 7	0.831 6	0.885 5	**0.909 4**
Bitplane 5	0.804 7	0.858 6	0.821 5	0.905 7
Bitplane 6	**0.892 3**	**0.895 6**	**0.909 1**	0.909 2
Bitplane 7	0.690 2	0.713 8	0.686 9	0.690 2

注:加粗字体表示抗攻击实验中最优值。

3 结束语

本文提出基于格雷码置乱和分块混沌置乱的医学影像加密方案GBCS,并应用于基于隐私保护的分类挖掘。利用邻域自适应以及随机性克服深度学习普遍存在的对抗样本攻击问题。先对原始图像进行位平面切分,既能抵御针对单像素置乱的同时也方便对图像进行加密和扰动。在此基础上,利用图像分块算法对原始病理图像进行分块处理,各个子块体现不同方向的局部特性。在各个子块上进行混沌序列置乱。最后通过深度学习网络对加密后的图像进行分类学习。本文从图像直方图、信息熵、对抗攻击能力等指标考虑其安全性,在公开的MIAS乳腺癌和青光眼ORIGA数据集上进行交叉仿真验证其分类性能。分类实验结果表明,GBCS中8个位平面的平均分类准确率相比原始图像仅相差0.26%和2.14%,仿真攻击结果表明图像被攻击后仍可较好恢复。综上可见,医学图像在GBCS加密前后的性能差距在可接受范围内,更好地平衡了性能与隐私保护需求的矛盾。在GBCS切割后的位平面在受到对抗样本攻击时仍然能保持良好分类性能,说明它能有效防御对抗样本的攻击。基于置乱的加密方案的量子门电路实现方案已经在文献[12-13]进行讨论,它具有鲁棒性、可实现性和高效率等优点,今后有希望在量子计算机进行部署。

但GBCS方案也存在以下不足:医院里的医学影像数据集种类样本数量规模较小,也容易导致过拟合问题。今后的任务之一是把改进的GBCS应用于更多不同种类的医学图像,以求进一步扩展其应用范围。

参考文献：

[1] TALBI R, BOUCHENAK S, CHEN L Y. Towards dynamic end-to-end privacy preserving data classification[C]//Proceedings of the 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W).[S.l.]: IEEE, 2018: 73-74.

[2] VANI E, VEENA S, ARAVINDAR J D. Query processing using privacy preserving K-NN classification over encrypted data [C]//Proceedings of 2016 International Conference on Information Communication and Embedded Systems (ICICES). Chennai, India: IEEE, 2016: 1-5.

[3] SAMANTHULA B K, ELEMEHDWI Y, JIANG W. K-nearest neighbor classification over semantically secure encrypted relational data[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(5): 1261-1273.

[4] WU W, LIU J, RONG H, et al. Efficient k-nearest neighbor classification over semantically secure hybrid encrypted cloud database[J]. IEEE Access, 2018, 6: 41771-41784.

[5] FAKHR M W. Multiple encrypted random forests using compressed sensing for private classification[C]//Proceedings of 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT). Sakhier, Bahrain: IEEE, 2018: 1-7.

[6] MAEKAWA T, KAWAMURA A, KINOSHITA Y K, et al. Privacy-preserving SVM computing in the encrypted domain [C]//Proceedings of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Honolulu, HI, USA: IEEE, 2018: 897-902.

[7] MAHMOOD Z H, IBRAHEM M K. A noise free homomorphic encryption based on chaotic system[C]//Proceedings of Information Technology To Enhance e-learning and Other Application. Baghdad, Iraq: IEEE, 2020: 132-137.

[8] 李明慧,江沛佩,王骞,等.针对深度学习模型的对抗性攻击与防御[J].计算机研究与发展,2021,58(5): 909-926.
LI Minghui, JIANG Peipei, WANG Qian, et al. Adversarial attacks and defenses for deep learning models[J]. Journal of Computer Research and Development, 2021, 58(5): 909-926.

[9] 周纯毅,陈大卫,王尚,等.分布式深度学习隐私与安全攻击研究进展与挑战[J].计算机研究与发展,2021,58(5): 927-943.
ZHOU Chunyi, CHEN Dawei, WANG Shang, et al. Research and challenge of distributed deep learning privacy and security attack[J]. Journal of Computer Research and Development, 2021, 58(5): 927-943.

[10] STROGATZ S H. Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering[M]. [S.l.]: CRC Press, 2018.

[11] CHAUDHARI S, MITHAL V, POLATKAN G, et al. An attentive survey of attention models[EB/OL]. (2020-12-15). https://arxiv.org/pdf/ 1904.02874.pdf.

[12] ZHOU R, SUN Y, FAN P. Quantum image gray-code and bit-plane scrambling[J]. Quantum Information Processing, 2015, 14(5): 1717-1734.

[13] ABDELLATIF A A, ABDELATTY B. Robust encryption of quantum medical images[J]. IEEE Access, 2018, 6: 1073-1081.

作者简介：

陈国明(1977-)，男，博士，副教授，研究方向：数据挖掘、机器学习与信息安全等，E-mail: isscgm@163.com。

麦舒桃(1979-)，女，博士，副主任医师，研究方向：中西医结合治疗危急重症等，E-mail: bakumaomao@126.com。

袁泽铎(1997-)，**通信作者**，男，硕士，研究方向：机器学习、计算机视觉等，E-mail: yuanzdstu2020@jnu.edu.cn。

龙舜(1971-)，男，博士，副教授，研究方向：计算机图像与视频处理、人工智能等，E-mail: tlongshun@jnu.edu.cn。

（编辑：陈珺）