Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

基于 Wasserstein GAN 的新一代人工智能小样本数据增强方法——以生物领域癌症分期数据为例

刘宇飞[a,b,d], 周源[b,*], 刘欣[a], 董放[b], 王畅[a], 王子鸿[c]

[a] College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[b] School of Public Policy and Management, Tsinghua University, Beijing 100084, China
[c] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[d] Center for Strategic Studies, Chinese Academy of Engineering, Beijing 100088, China

ARTICLE INFO

Article history:
Received 5 March 2018
Revised 2 June 2018
Accepted 7 November 2018
Available online 11 January 2019

关键词
人工智能
生成式对抗网络
深度神经网络
小样本
癌症

摘要

以大数据为基础的深度学习算法在推动新一代人工智能快速发展中意义重大。然而深度学习的有效利用对标注样本数量的高度依赖，使得深度学习在小样本数据环境下的应用受到制约。本研究提出了一种基于生成对抗网络（generative adversarial network，GAN）和深度神经网络（deep neural network，DNN）分类器的方法。首先，将原始样本划分为训练集样本和测试集样本，采用训练集样本训练 GAN 后生成模拟样本数据，扩增训练集样本规模；然后，使用模拟样本训练 DNN 分类器；最后，使用测试集样本测试分类器，并通过指标验证该方法在小样本多分类问题下的有效性。作为实证案例，将该方法应用于生物领域癌症分期识别，结果表明该方法比传统方法获得更高的识别准确率。同时，该方法是一次将基于原始样本的经典统计机器学习分类方法转变为基于数据增强的深度学习分类方法的尝试。本研究有助于探索以深度学习为代表的新一代人工智能技术在应用范围与应用效果方面的潜力。这将对各领域全面推进新一代人工智能的发展具有重要意义。

1. 引言与背景

自1956年人工智能概念被提出以后[1]，经过60多年的发展，特别是在移动互联网、大数据、超级计算、传感网、脑科学等新理论、新技术的引领下，人工智能发展所处的信息环境和数据基础已经发生了巨大而深刻的变化[2]。国务院2017年7月印发的《新一代人工智能发展规划》标志着人工智能将全面进入新一代人工智能的发展阶段[3]。该规划指出新一代人工智能呈现出深度学习、跨界融合、人机协同、群智开放和自主智能的新特点，其技术基础是大数据驱动算法[4]。潘云鹤院士[2]也提出人工智能的基本方法是数据驱动算法，大数据智能是新一代人工智能的重要发展方向。

在大数据环境下，Hinton和Salakhutdinov[5]发起的深度学习已经成为大数据智能的关键核心技术[6]，并在智能驾驶[7]、智慧城市[8]、语音识别[9]和信息检索[10]等多个领域取得重大突破。以深度学习为核心的大数据智能方法相比于经典的统计机器学习方法具有更大的模型容量，经大规模已标注样本数据充分训练后，可以展现出更加优越的性能[11]。换言之，在现有的深度学习技术框架下，训练数据规模越大、标注质量越高，算法模型的性能就越好。标注数据对深度学习的应用效果有决定性的作

* Corresponding author.
E-mail address: zhou_yuan@tsinghua.edu.cn

2095-8099/
英文原文: Engineering 2019, 5(1): 156 –163
引用本文: Yufei Liu, Yuan Zhou, Xin Liu, Fang Dong, Chang Wang, Zihong Wang. Wasserstein GAN-Based Small-Sample Augmentation for New-Generation Artificial Intelligence: A Case Study of Cancer-Staging Data in Biology. *Engineering*, https://doi.org/10.1016/j.eng.2018.11.018

用，同时影响着新一代人工智能行业的发展态势。然而在许多应用场景中，标注样本获取困难且成本很高。例如，在生物领域进行数据分类时，一个训练样本的获取往往需要长时间且昂贵的实验[12]；在数控机床健康保障中，数控机床出现故障的情况属于个例，异常样本的收集往往需要长期积累，然而我国数控机床仍在普及阶段，使得异常样本的收集更加困难[13]；在战略情报分析中，数据标注需要各领域专家的高度配合，得到大规模标注样本的成本同样极高[14–20]。导致数据获取困难的原因除了高昂的成本外，数据的复杂性和高维性也是关键因素。在这种情况下，当特征空间的维度大于样本数量时，被认为是小样本问题[21]。该问题使得经典统计机器学习算法无法获得优良的泛化性能，同时更进一步制约了以深度学习为核心的大数据智能方法在此类领域中的应用。在实践中，面临小样本问题的领域远多于具有大数据环境的领域，这严重制约了新一代人工智能的发展[22–26]。

为了解决标注样本不足的问题，近年来的相关研究主要集中在过采样技术。过采样技术是对原始数据集进行过采样，从而扩大标注样本规模。早期通过简单复制原始数据集或者在原始数据集上加上人工噪声形成新数据集，可以解决数据不平衡情况下标注样本不足的问题[27]。2002年，Chawla等[28]提出了经典的合成少数类过采样技术（synthetic minority oversampling technique，SMOTE），并生成模拟样本。Han等[29]基于SMOTE提出了边界和距离边界合成少数类过采样技术，但这两种技术仅仅考虑了相邻的样本并过采样处于边界附近的样本。2008年，He等[30]提出了自适应合成抽样方法，根据少数样本类的学习难度对数据集分配权重。2014年，Barua等[31]提出了多数加权少数过采样技术，目的在于产生有效的模拟样本。这种方法通过先找出难以学习的少数类样本，然后根据距离其最近的多数类样本的欧几里得距离分配权重，最后通过聚类方法生成模拟样本。2015年，Xie等[32]提出了基于低维空间局部密度的少数类过采样技术，通过将训练样本映射到低维空间并分配权重，以解决过采样中的维数灾难问题。2017年Douzas和Bacao[33]提出了自组织映射过采样技术，其生成的模拟数据仅对特定的分类器有较好的效果。现有的研究主要通过在少数类数据集中加入过采样生成的模拟样本，集中解决数据不平衡的问题。然而，在实际应用中，数据集在很多领域面临的问题不是数据不平衡的问题，而是每一类样本规模均较小，难以对模型进行有效训练。

除了SMOTE，还可以尝试用一些传统方法来解决小样本问题。1995年，Bishop等[34]提出带着噪声训练可以得到更好的结果，这一想法符合Tikhonov规则。2004年，Zhou和Jiang[35]提出了一种名为基于神经元集群的C4.5决策树方法。这种方法首先对神经网络进行训练，然后使用训练后的神经网络生成训练集样本。2006年，Li和Lin[36]提出了一种基于内核密度估计的虚拟样本生成方法，根据已知小样本训练数据，估计出样本数据服从的概率密度函数，然后根据估计得到的概率密度函数生成训练样本。2009年，Li和Fang[37]利用组发现技术和超球特征方程提出了一种非线性虚拟样本生成技术，用于解决样本不足的问题。这些方法对于解决小样本问题具有推动作用，但是难以有效捕捉小样本的固有特征，导致训练的虚拟样本生成模型仍具有局限性。

随着新一代人工智能及大数据智能的快速发展，近年来提出的以深度神经网络（deep neural network，DNN）为基础的生成对抗网络（generative adversarial network，GAN）为解决此类小样本问题提供了新的方法，同时为小样本数据下应用以深度学习为核心的大数据智能方法提供了可能。GAN是一种强大的样本生成模型[38]，由Goodfellow等[39]于2014年提出。GAN针对标注样本不足的问题，能够生成与真实样本分布相同的模拟样本，扩大标注样本规模[40]。GAN包含生成器和判别器，并均为深度神经网络结构。GAN在生成器与判别器相互博弈的过程中学习真实样本的分布[41]。在GAN的训练过程中，生成器尽可能地拟合真实样本的分布并生成模拟样本，判别器则尽可能准确地分辨出真实样本和模拟样本。

近期研究表明，GAN已在多个领域中发挥重要作用，如图片合成、语言处理、低质量数据的监督式学习。在图片合成方面，Santan和Hotz[42]提出了一种根据真实驾驶场景的数据分布生成模拟图片的方法。Gou等[43]采用GAN生成模拟样本提高了人眼识别的准确率。在自然语言处理方面，Li等[44]采用GAN生成模拟对话数据，Pascual等[45]基于GAN提出了一种优化问答系统框架。这些研究表明，GAN可以生成符合原始样本分布的模拟样本，同时，在多个领域的应用也表明，GAN对领域知识没有很强的依赖性，可以很容易地推广到其他领域。目前，在提高监督式学习数据质量的研究中，GAN主要用于解决数据不平衡问题。Fiore等[46]采用GAN生成小样本类别的模拟样本，并与原始数据混合，训练分类器以提升信用卡欺诈的识别效果。Douzas和Bacao[47]在不同数据集上

采用GAN生成小样本类别的模拟样本，并与多种过采样方法进行对比，结果显示GAN方法更为有效。这些研究表明，GAN在提高数据质量——数据增强方面取得了成功，基于深度学习生成的模拟样本质量优于采用传统过采样方法生成的样本质量。此外，大多数标准数据增强方法都已经集成到扩增器（Augmentor）中，Augmentor是一种公认的具有高级应用编程接口（application programming interface，API）的数据增强工具[48]。然而，GAN目前应用于数据增强方面的研究依然集中在解决数据不平衡问题上，应用于小样本环境下监督式学习的研究还比较缺乏。因此，本文尝试将GAN应用于多类别的小样本数据增强，这样可能会提高基于DNNs的监督式机器学习在各领域的性能。另一方面，随着数据规模的增大，DNN的应用成为可能，凭借样本容量更大的模型，可以为多分类问题提供相比于经典机器学习更好的性能。

基于上述研究，为了解决小样本下的监督式学习问题，本文提出一种结合GAN与DNN的分类器方法。首先，将原始样本划分为训练集样本和测试集样本，使用生成对抗网络对生物领域癌症分期训练样本进行数据增强，生成大量模拟样本。然后，使用生成的模拟样本训练基于深度学习的DNN分类器。最后，使用测试样本对训练得到的分类器进行测试，并结合经典监督式机器学习方法——DNN、SMOTE和GAN，通过对比多个指标，验证该方法在小样本多分类问题下的有效性。该方法是一次将基于原始样本的经典统计机器学习分类方法转变为基于数据增强的深度学习分类方法的尝试。本研究有助于探索以深度学习为代表的新一代人工智能在应用范围与应用效果方面的潜力。同时，本研究首次将GAN与DNN分类器结合应用于改善癌症分期识别精度。期望通过实证分析的结果为癌症的早期诊断分析提供帮助。

2. 方法

2.1. 研究流程

为了探索小样本下监督式学习问题的解决方案，扩大深度学习的适用范围，本文提出一种基于GAN和DNN分类器的小样本多分类方法。提出的方法流程如图1所示：

（1）将原始样本划分为训练集样本和测试集样本，采用训练集样本分别训练GAN模型，优化GAN模型参数；

（2）采用GAN的生成器生成模拟样本，采用GAN的判别器进行过滤；

（3）采用过滤后的模拟样本训练DNN分类器，采用测试集样本测试DNN分类器。

2.2. 生成式对抗网络

本文采用沃瑟斯坦生成式对抗网络（Wasserstein generative adversarial network，WGAN）生成对抗样本[49]。由于原始GAN的训练过程是一种极大极小的博弈游戏，优化目标是达到纳什平衡[40]，所以这一过程存在梯度消失的问题[50]。相比于原始GAN，WGAN采用的是Wasserstein距离而不是Jensen-Shannon（JS）差异来评估真实样本和模拟样本之间的分布差异[51]。同时，由于采

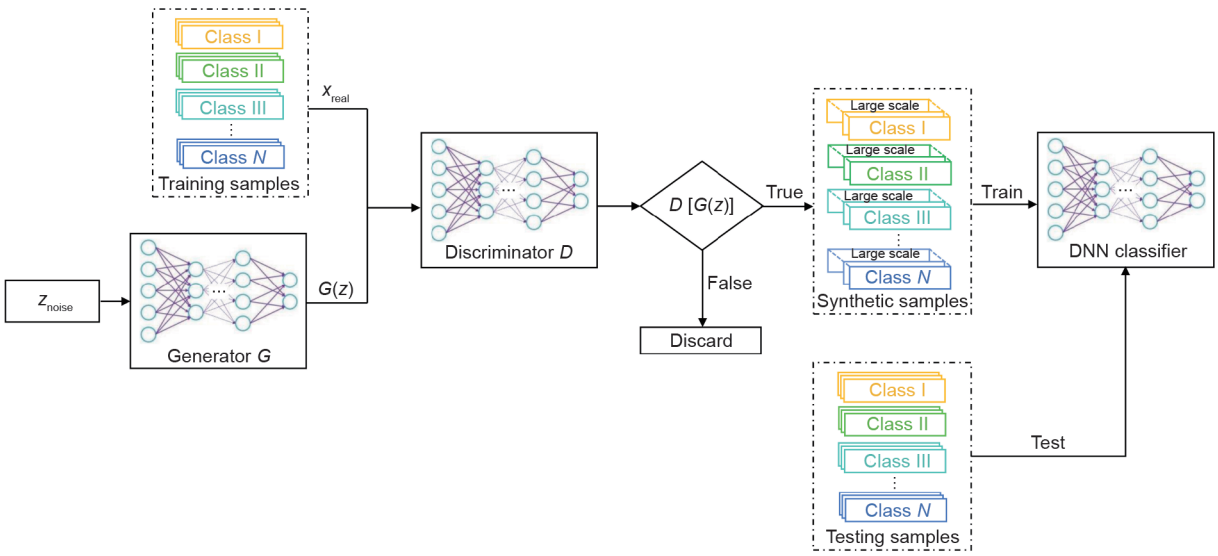

图1. 小样本多分类方法流程。

用Wasserstein距离，WGAN训练速度更快，训练过程更加稳定[52]。

WGAN模拟样本的生成过程包含两个过程：生成器生成原始模拟样本，判别器对原始模拟样本判断过滤最终得到模拟样本。首先，WGAN生成器和判别器经多次训练，在其损失函数均收敛时，生成器开始生成原始模拟样本。然后，根据GAN的对抗思想[53]，生成器尝试生成可以骗过判别器的模拟样本，而判别器尽可能分辨真实样本与模拟样本。因此，当原始模拟样本经判别器判断为真实样本时，则该模拟样本就骗过了判别器，这部分骗过判别器的模拟样本就作为最终的有效模拟样本。在本文提出的方法中，通过采用各个类别的真实样本分别训练相应的WGAN生成相应类别的模拟样本。

2.3. 深度神经网络

本文所选择的DNN是一种基于深度学习具有多层神经网络结构的分类器。DNN分类器能够利用几种计算模型学习到数据的多层次抽象特征，其中高层次的特征由低层次的特征组合构造而成，能够更加有效地表达数据的分布特征，从而相比于经典统计机器学习模型获得更好的学习效果。为了避免出现过度学习的现象，DNN分类器采用WGAN生成的大量模拟样本进行训练，并采用测试样本测试，这样可以有效检验分类器的泛化性能。为了测试DNN分类器的表现，采用基于混淆矩阵（图2）的三个多分类指标评估DNN分类器，即accuracy，*F*-measure，*G*-mean。Accuracy是分类预测正确的比例，*F*-measure是精准率和召回率的调和平均值[54]，*G*-mean是每个类别召回率的几何平均值[55]。Accuracy，*F*-measure和*G*-mean的定义如式（1）至式（3）所示。

$$\text{Accuracy} = \frac{\sum_{i=1}^{L} n_{ii}}{\sum_{i=1,j=1}^{L} n_{ij}} \tag{1}$$

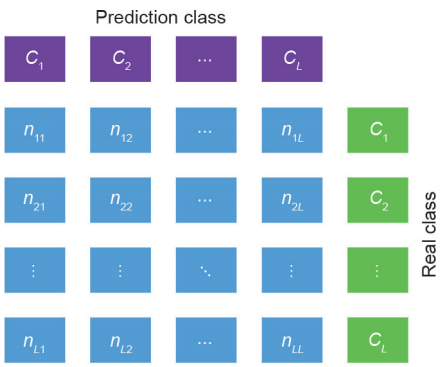

图2. 混淆矩阵示意图。

$$F\text{-measure} = \frac{2}{L}\frac{\sum_{i=1}^{L} R_i \sum_{i=1}^{L} P_i}{\sum_{i=1}^{L} R_i + \sum_{i=1}^{L} P_i} \tag{2}$$

$$G\text{-mean} = \left(\prod_{i=1}^{L} R_i\right)^{1/L} \tag{3}$$

在这些公式中，L为类别数，C_i类样本被正确预测为C_i类的样本数为n_{ii}，被错误预测为C_j类的样本数为n_{ij}。R_i和P_i分别为C_i类召回率和精度，具体计算公式：

$$R_i = \frac{n_{ii}}{\sum_{j=1}^{L} n_{ij}} \tag{4}$$

$$P_i = \frac{n_{ii}}{\sum_{j=1}^{L} n_{ji}} \tag{5}$$

3. 实证分析与讨论

病理数据作为一种患者隐私，具有数据标引难获取以及获取成本高的特点，本方法在此领域的应用具有典型性，将为一系列的类似问题提供解决思路。肝细胞癌（hepatocellular carcinoma，HCC）是一种常见的恶性肿瘤，5年生存率低于15%[56,57]。而肝细胞癌的早期治疗能够有效提升患者的5年生存率，因此，肝细胞癌的早期识别对于该癌症的治疗有重要价值，在近年来的研究中也逐渐受到关注。但是，受限于含有癌症分期信息的样本数量缺乏，相关研究进展缓慢。糖基化是最广泛的蛋白质翻译后修饰形式之一，在不同的生物学过程中起着关键作用[58–60]。许多癌症相关的过程，包括致癌转化[59,61]、肿瘤进展[62]和抗肿瘤免疫[63]都与蛋白质的异常糖基化有关。此外，多种肿瘤标志物都是来源于血清糖组中变化的糖蛋白[64–67]。因此，糖基化数据可以作为癌症分期预测的有效方法。本研究将WGAN结合DNN的方法应用于肝癌样本数据的对抗生成，对小样本下肝癌患者的病理周期分类以及肝癌的早期识别具有重要意义。

3.1. 数据获取

在本研究中，作为实验数据的血清样本由同济医院（华中科技大学同济医学院）提供。首先通过肽*N*-糖苷酶F（PNGase F）切断糖链与天冬酰氨间的糖肽键，从人血清糖蛋白中特异性释放*N*-聚糖[68]。然后进行全甲基化修饰，并通过MALDI 4800（AB SCIEX, Concord, Canada）进行质谱检测，得到样本中*N*-聚糖质荷比（*m/z*）在质谱上峰的分布及相对含量，如图3所示。最后，使用Data Explorer

4.5对得到的质谱数据进行处理，并生成包含*m/z*值与质谱强度（参照美国信息交换标准代码对照光谱）的.txt文件。癌症分期按照肿瘤TNM（tumor node metastasis）分期系统进行划分。通过上述生物学过程，获得60个肝癌样本（TNM I期21个、TNM II期24个、TNM III期15个）和作为对照组的18个健康样本。其中每个肝癌样本包含42个特征，每个样本根据其峰值分布及相对强度可表示为一个42维的特征向量（图3）。将所有肝癌分期原始样本分别按照60%和40%的比例划分为训练集样本和测试集样本，划分结果如表1所示。

3.2. 结果分析

根据本文所提出的方法，首先，分别采用TNM I期、TNM II期、TNM III期、对照组的训练集样本训练WGAN；之后，使用训练后的WGAN生成模拟样本。通过多次实验，选择以下表现最好的WGAN模型结构参数：生成器具有1层隐藏层，隐藏层为32个修正线性单元（rectified linear unit，ReLU），生成器的输出层为42个sigmoid单元，噪声向量*z*的维度设置为15；判别器同样具有1层隐藏层，隐藏层为64个ReLUs，判别器的输出层为1个无激活函数单元。每种类别训练样本对应的WGAN结构参数相同。WGAN的开发环境为TensorFlow1.1，在GPU环境下训练。WGAN训练过程包含3×10^5个迭代周期，在每个迭代周期中，判别器首先迭代100次，之后生成器迭代1次。

在生成模拟样本之后，使用模拟样本对DNN分类器进行训练，并使用肝癌分期测试样本对DNN分类器进行测试。DNN分类器的类型为多层感知机（multi-layer perceptron，MLP）分类器。经过多次实验选择以下结构参数：DNN分类器的输入层数量为42，等同于肝癌样本数据的特征数量，DNN分类器具有3个隐藏层，每层均包含32个ReLUs，输出层为softmax函数，损失函数为交叉熵。DNN分类器的开发环境同样为TensorFlow1.1，并在GPU环境下训练，迭代次数为3000次。

为了测试WGAN生成模拟样本数量对DNN模型训练的影响，在100个模拟样本以内时，每生成20个样本，使用当前样本对DNN进行训练，并计算模型的accuracy，*F*-measure，*G*-mean这三项指标，在100个样本以上时，每100个样本测试一次这三项指标。指标变化如图4所示。

随着样本数量的增加，准确率逐渐提高。生成100个模拟样本后，accuracy为51.61%。当训练样本数量达到1000时，accuracy达到64.52%。随着样本数量的增加，准确率继续逐渐上升，当样本数量达到2000时，accuracy达到67.74%。但是在这之后，样本数量的增加并没有带来准确率的继续提升，准确率一直在67%上下波动。直到样本数量达到4000后，准确率稳定在70%左右。另外，*F*-measure在样本数量增加的过程中，变化趋势与准确率的变化趋势基本吻合，表明各类别的预测准确率与总体情况保持一致，各时期的HCC样本都能够被有效地预测，误诊率很低。*G*-mean在样本数量增加的过程中，虽然一直略低于准确率，但总体趋势保持一直，说明在分类器的预测过程中，一直保持着较低的漏诊率。根据accuracy，*F*-measure，*G*-mean三项指标，可以认为当模拟样本数量达到4000时，DNN就可以得到针对小样本下HCC样本分

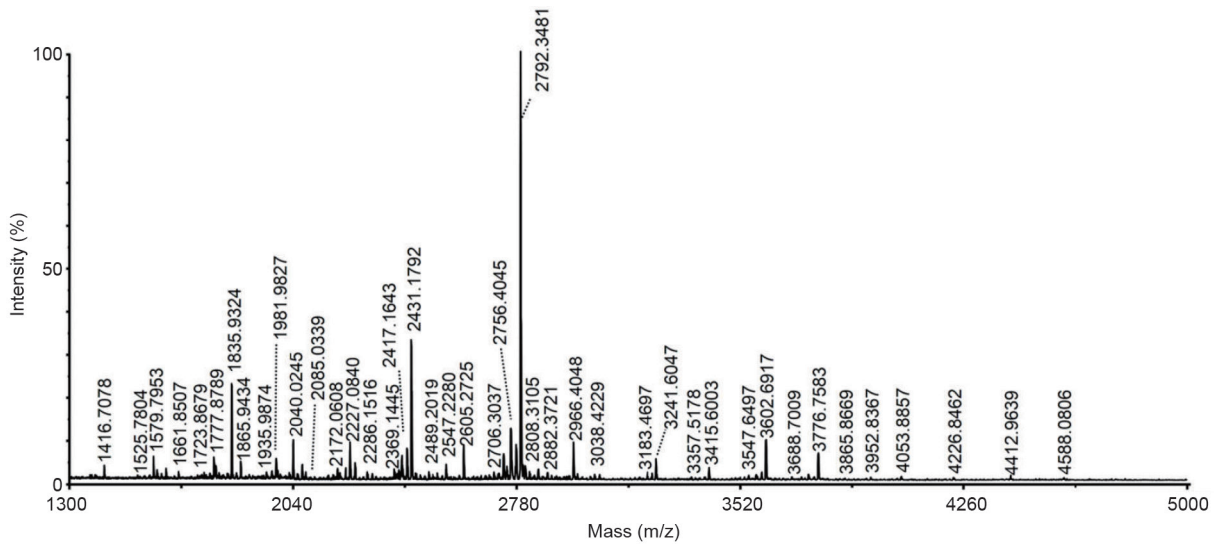

图3. 基于MALDI-MS的肝细胞癌样本*N*-聚糖分布与含量示意图。

期识别的有效训练模型。

具体而言，当生成模拟样本数量为4000时，原始样本的测试准确率为70.97%（31个原始样本，22个预测正确），*F*-measure为70.07%，*G*-mean为68.39%。这时，DNN对癌症每一个时期的预测情况如表2所示。具体而言，当生成模拟样本数量为4000时，原始样本的测试准确率为70.97%（31个原始样本，22个预测正确），*F*-measure为70.07%，*G*-mean为68.39%。这时，DNN对癌症每一个时期的预测情况如表2所示。作为对照组的健康测试样本全部预测正确，8个TNM I期的测试样本有5个预测正确，有2个被预测为TNM II期，有1个被预测为TNM III期。10个TNM II期的测试样本有7个预测正确，有2个被错误的预测为健康样本，这以结果可能会带来比较大的漏诊风险，需要在未来的预测中对TNM II期样本保持关注，另外，还有1个样本被错误的预测为TNM III期。6个TNM III期样本只有3个被预测正确，另外3个样本被预测为TNM I期，对TNM III期样本的预测准确率只有50%，这也拉低了DNN分类器整体的预测准确率，在未来的研究中需要尝试增加TNM III期的原始训练样本数量，进一步提高DNN模型对TNM III期的特异性学习能力。根据生物学，因为TNM I期和II期具有类似的临床表征，可以将其归为一类，共称为癌症的早期。从对于肝细胞癌早期的预测效果来看，共18个癌症早期测试样本，有14个预测正确，预测准确率达到了77.78%，超过了分类器总体的预测准确率，这对于肝癌的早期识别与治疗有重要意义，因为目前的研究发现，对肝癌的早期治疗能够显著提高患者的存活率[69]。Holzinger等[70]最近的一项研究表示如果机器学习能够为数字化病理学提供帮助，将会显著改变病理学家的医学研究流程。因此，对癌症分期的准确预测为进一步研究肝癌的致病机理提供了可能性。

表1 训练集样本和测试集样本划分结果

	HCC original samples		Total
	Training set	Test set	
Healthy	11	7	18
TNM stage I	13	8	21
TNM stage II	14	10	24
TNM stage III	9	6	15
Total	47	31	78

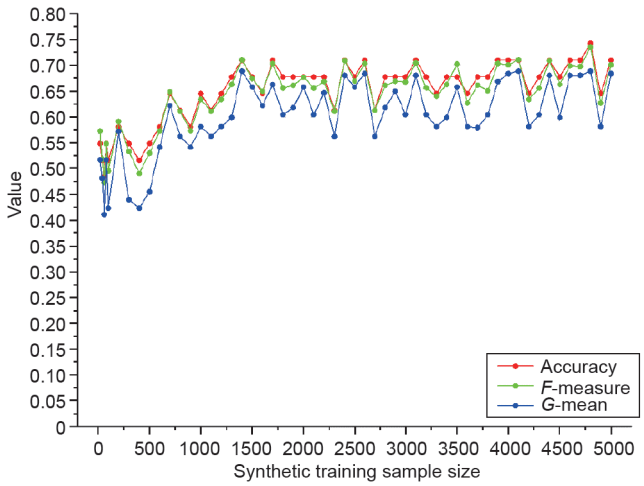

图4. 评价指标与生成样本数量的关系。

3.3. 方法评价

为了验证提出框架应用于小样本肝癌数据分期的有效性，分别在数据生成和分类模型算法上采用了数据过采样算法和经典统计机器学习模型与其进行对比。由于随机森林（random forest，RF）作为一种集成学习模型相比于其他统计机器学习模型具有更高的分类准确性和更好的泛化性能[71]，同时朴素贝叶斯（naive Bayes，NB）算法原理简单，具有稳定的分类性能[72]，因此采用RF和NB作为经典统计机器学习分类模型。对于原始小样本肝癌数据，由于深度学习对样本需求量大，分类器仅在经典统计机器学习算法上测试。在原始小样本肝癌数据应用于经典统计机器学习算法时，采用训练样本训练分类器、测试样本测试分类器。在采用数据过采样方法生成模拟样本时，首先采用SMOTE方法对原始样本数据各类别分别进行过采样，之后采用SMOTE算法生成的模拟样本对分类模型进

表2 HCC分期识别的混淆矩阵

	Predicted				Total
	Healthy	TNM stage I	TNM stage II	TNM stage III	
Real healthy	7	0	0	0	7
Real TNM stage I	0	5	2	1	8
Real TNM stage II	2	0	7	1	10
Real TNM stage III	0	3	0	3	6
Total	9	8	9	5	31

行训练并采用原始样本测试模型。在数据过采样方法实验中，多分类模型分别采用经典统计机器学习分类模型RF、NB以及深度学习分类模型DNN。在生成对抗网络方法中，采用WGAN生成大量模拟数据对分类模型进行训练并采用原始样本对模型进行测试，多分类模型同样采用RF、NB和DNN。

由表3可见，使用训练样本分别对RF、NB进行训练，采用测试样本测试，accuracy分别为54.84%和32.26%，*F*-measure分别为56.73%和12.20%，*G*-mean分别为45.50%和0，三项指标均较低，大量肝癌样本没有被识别出来，尤其是NB的各项指标远低于RF，说明在这一数据集中NB相比于RF对样本数量更加敏感。这一结果表明在小样本数据的环境下，难以对经典机器学习模型进行有效训练。

使用SMOTE算法生成4000个过采样样本，使用这些样本分别对RF、NB进行训练，相比原始样本，各项指标均有所提升，这表明通过SMOTE进行过采样训练经典机器学习模型，可以在一定程度上提升癌症样本的识别准确率并降低模型漏诊率；使用SMOTE算法生成4000个过采样样本对DNN进行训练，相比于使用这些样本分别对RF、NB进行训练，各项指标均有较大提升，accuracy达到64.52%，*F*-measure达到66.05%，*G*-mean达到63.32%，这表明使用大数据量训练深度学习模型相比于经典机器学习模型可以取得更好的效果。

使用由WGAN生成的4000个模拟样本对RF和NB进行训练，相比于使用SMOTE生成的过采样样本对RF和NB进行训练，RF的这三项指标均大幅降低，然而NB的这三项指标均大幅增加，这一结果说明WGAN生成的模拟样本在不同类别经典机器学习模型的表现并不一致，这表明WGAN与经典机器学习模型结合应用并不一定能取得良好的结果。使用WGAN生成的模拟样本对DNN进行训练，相比于SMOTE过采样方法则会进一步提高模型有效性，accuracy由64.52%上升到70.97%，*F*-measure由66.05%上升到70.07%，*G*-mean由63.32%上升到68.39%，这表明WGAN结合DNN的深度学习方法可以有效地解决小样本下的HCC分期识别问题。这也标志着深度学习方法在小样本多分类问题上的有效应用。

3.4. 讨论

根据上述结果，深度学习结合生成对抗网络应用于改善癌症分期识别精度取得很好的效果，各项指标均超过了传统方法。这一结果对癌症研究有重要意义。多数癌症都面临样本量偏小，尤其是具有准确分期信息的样本。这导致癌症早期诊断与治疗的研究进展缓慢，进一步影响了对癌症致病机理的探寻。基于GAN的数据增强方法很可能促进这些问题的解决。本文方法的设计初衷并不仅是为了解决肝癌的分期预测问题，更是为了解决小样本下的有监督学习问题，因此，特意选择了样本数据量偏小、在传统统计机器学习方法中效果不理想的基于血清样本的癌症分期研究。同时，基于深度学习的特性，该方法并不依赖精准的癌症研究领域知识，因此，在保证有效性的同时，大大降低了该方法拓展到其他应用领域的障碍。对因样本问题而难以实现深度学习的领域，本方法的推广为其智能化提供了巨大的潜力[26,73–75]。

表3 不同分类策略的性能对比

	Accuracy	*F*-measure	*G*-mean
RF	0.5484	0.5673	0.4550
NB	0.3226	0.1220	0
RF with SMOTE	0.5806	0.6254	0.4811
NB with SMOTE	0.5484	0.5600	0.5233
DNN with SMOTE	0.6452	0.6605	0.6332
RF with WGAN	0.2903	0.2931	0.2627
NB with WGAN	0.6129	0.6260	0.6043
DNN with WGAN	0.7097	0.7007	0.6839

4. 结论

本文构建了一套深度神经网络结合生成式对抗网络的方法，并首次应用于小样本下癌症分期识别，相比于传统分类方法，不仅改善了有效性，同时充分地保留了样本的特征，可以更好地帮助癌症的早期识别，这对于癌症的诊断具有重要的意义。更重要的是，该研究提出的这种小样本下的有监督深度学习方法，为其他应用领域中的小样本问题提供了具有潜力的高效解决方法，从而提高各个领域的智能化程度。这对各领域全面推进新一代人工智能的发展具有重要意义。这也是我们未来研究的重点，将在更多领域的数据集上应用这一方法，不断提高该方法的适用性[76–78]。

致谢

本研究得到国家自然科学基金项目（91646102, L1724034, L16240452, L1524015, 20905027）、教育部人文社会科学项目（16JDGC011）、中国工程科技知识中

心建设项目（CKCEST-2018-1-13）、中英产学合作项目（UK-CIAPP\260）、清华大学绿色经济与可持续发展研究中心子项目（20153000181）和清华大学自主科研项目（2016THZW）的支持。

Compliance with ethics guidelines

Yufei Liu, Yuan Zhou, Xin Liu, Fang Dong, Chang Wang, and Zihong Wang declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Crevier D. AI: the tumultuous history of the search for artificial intelligence. New York: Basic Books, Inc.; 1993.
[2] Pan Y. Heading toward Artificial Intelligence 2.0. Engineering 2016;2 (4):409–13.
[3] State Council of the People's Republic of China. Development Plan for a Next-Generation Artificial Intelligence [Internet]. Beijing: www.gov.cn. [cited 2018 Mar 5]. Available from: http://english.gov.cn/policies/latest_releases/2017/07/20/content_281475742458322.htm.
[4] State Council Information Office of the People's Republic of China. The policy interpretation of Development Planning for a Next-Generation Artificial Intelligence [Internet]. Beijing: www.scio.gov.cn. [cited 2018 Mar 5]. Available from: http://www.scio.gov.cn/34473/34515/Document/1559231/ 1559231.htm. Chinese.
[5] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–7.
[6] Zhuang Y, Chen C, Pan Y. Challenges and opportunities: from big data to knowledge in AI 2.0. Front Inf Technol Electronic Eng 2017;18(1):3–14.
[7] Al-Qizwini M, Barjasteh I, Al-Qassab H, Radha H. Deep learning algorithm for autonomous driving using GoogLeNet. In: Proceedings of the 2017 IEEE Intelligent Vehicles Symposium; 2017 Jun 11–14; Los Angeles, CA, USA. New York: IEEE; 2017. p. 89–96.
[8] Wang L, Sng D. Deep learning algorithms with applications to video analytics for a smart city: a survey. 2016: arXiv:1511.06434.
[9] Mohamed A, Dahl G, Hinton G. Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process 2012;20(1):14–22.
[10] Jones N. Artificial-intelligence institute launches free science search engine [Internet]. Heidelberg: Springer Nature. c2018 [cited 2018 Mar 5]. Available from: https://www.nature.com/news/artificial-intelligence-institute-launchesfree- science-search-engine-1.18703.
[11] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: The MIT Press; 2016.
[12] Zhuang F, Luo P, Qing H, Shi Z. Survey on transfer learning research. J Software 2015;26:26–39. Chinese.
[13] Chen J, Yang J, Zhou H, Xiang H, Zhu Z, Li Y, et al. CPS modeling of CNC machine tool work processes using an instruction-domain based approach. Engineering 2015;1(2):247–60.
[14] Urban F, Zhou Y, Nordensvard J, Narain A. Firm-level technology transfer and technology cooperation for wind energy between Europe, China and India: from north–south to south–north cooperation? Energy Sustainable Dev 2015;28:29–40.
[15] Zhou Y, Zhang H, Ding M, Su J. How public demonstration project affects the emergence of a new industry: an empirical study on electric vehicle demonstration project in China. In: Proceedings of the 2013 Suzhou Silicon Valley–Beijing International Innovation Conference; 2013 Jul 8–9. New York: IEEE; 2013. p. 234–9.
[16] Zhou Y, Minshall T. Building global products and competing in innovation: the role of Chinese university spin–outs and required innovation capabilities. Int J Technol Manage 2014;64(2):180–209.
[17] Xu G, Wu Y, Minshall T, Zhou Y. Exploring innovation ecosystems across science, technology, and business: a case of 3D printing in China. Technol Forecast Social Change 2017;136:180–221.
[18] Li X, Zhou Y, Xue L, Huang L. Roadmapping for industrial emergence and innovation gaps to catch-up: a patent analysis of OLED industry in China. Int J Technol Manage 2016;7(1–3):105–43.
[19] Li X, Zhou Y, Xue L, Huang L. Integrating bibliometrics and roadmapping methods: a case of dye-sensitized solar cell technology-based industry in China. Technol Forecast Social Change 2015;97:205–22.
[20] Zhou Y, Pan M, Urban F. Comparing the international knowledge flow of China's wind and solar photovoltaic (PV) industries: patent analysis and implications for sustainable development. Sustainability 2018;10(6):1883.
[21] Theodoridis S, Koutroumbas K. Pattern recognition. 3rd ed. Orlando: Academic Press; 2006.
[22] Nordensvard J, Zhou Y, Zhang X. Innovation core, innovation semi-periphery and technology transfer: the case of wind energy patents. Energy Policy 2018;120:213–27.
[23] Pan M, Zhou Y, Zhou DK. Comparing the innovation strategies of Chinese and European wind turbine firms through a patent lens. Environ Innovation Societal Transitions. Epub 2017 Dec 27.
[24] Zhou Y, Pan M, Zhou DK, Xue L. Stakeholder risk and trust perceptions in the diffusion of green manufacturing technologies: evidence from China. J Environ Dev 2017;27(1):46–73.
[25] Zhou Y, Li X, Lema R, Urban F. Comparing the knowledge bases of wind turbine firms in Asia and Europe: patent trajectories, networks, and globalisation. Sci Public Policy 2016;43(2):476–91.
[26] Chen L, Xu J, Zhou Y. Regulating the environmental behavior of manufacturing SMEs: interfirm alliance as a facilitator. J Cleaner Prod 2017;165:393–404.
[27] DeRouin E, Brown J. Neural network training on unequally represented classes. New York: ASME Press; 1991.
[28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16(1):321–57.
[29] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, editors. Advances in intelligent computing. Berlin: Springer; 2005. p. 878–87.
[30] He H, Bai Y, Garcia EA, Shutao L. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the 2008 IEEE International Joint Conference on Neural Networks; 2008 Jun 1–8; Hong Kong, China. New York: IEEE; 2008. p. 1322–8.
[31] Barua S, Islam MM, Yao X, Kazuyuki M. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng 2014;26(2):405–25.
[32] Xie Z, Jiang L, Ye T, Li X. A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning. In: Renz M, Shahabi C, Zhou X, Cheema M, editors. Database systems for advanced applications. Cham: Springer; 2015. p. 3–18.
[33] Douzas G, Bacao F. Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. Expert Syst Appl 2017;82:40–52.
[34] Bishop CM. Training with noise is equivalent to Tiknonov regularization. Neural Comput 1995;7(1):108–16.
[35] Zhou Z, Jiang Y. Nec4.5: neural ensemble based C4.5. IEEE Trans Knowl Data Eng 2004;16(6):770–3.
[36] Li DC, Lin YS. Using virtual sample generation to build up management knowledge in the early manufacturing stages. Eur J Operat Res 2006;175 (1):413–34.
[37] Li D, Fang Y. A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere. Exp Syst Appl 2009;36 (1):844–51.
[38] Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang F. Generative adversarial networks: introduction and outlook. IEEE/CAA J Autom Sin 2017;4(4):588–98.
[39] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KG, editors. Advances in neural information processing systems. La Jolla: Neural Information Processing Systems Foundation, Inc.; 2014. p. 2672–80.
[40] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. IEEE Signal Process Mag 2018;35(1):53–65.
[41] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015:arXiv:1512.03131.
[42] Santana E, Hotz G. Learning a driving simulator. 2016:arXiv:1608.01230.
[43] Gou C, Wu Y, Wang K, Wang F, Ji Q. Learning-by-synthesis for accurate eye detection. In: Proceedings of the 23rd International Conference on Pattern Recognition; 2016 Dec 4–8; Cancun, Mexico. New York: IEEE; 2017. p. 3362–7.
[44] Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D. Adversarial learning for neural dialogue generation. 2017:arXiv:1701.06547.
[45] Pascual S, Bonafonte A, Serrà J. SEGAN: speech enhancement generative adversarial network. 2017:arXiv:1703.09452.
[46] Fiore U, Santis AD, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Inf Sci 2019;479:448–55.
[47] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst Appl 2017;91:464–71.
[48] Bloice MD, Stocker C, Holzinger A. Augmentor: an image augmentation library for machine learning. 2017:arXiv:1708.04680
[49] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017:arXiv:1701.07875.
[50] Ratliff LJ, Burden SA, Sastry SS. Characterization and computation of local Nash equilibria in continuous games. In: Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing; 2013 Oct 2–4; Monticello, IL, USA. New York: IEEE; 2006. p. 917–24.
[51] Danihelka I, Lakshminarayanan B, Uria B, Wierstra D, Dayan P. Comparison of maximum likelihood and GAN-based training of real NVPs. 2017: arXiv:1705.05263.
[52] Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, et al. Low dose CT Image denoising using a generative adversarial network with Wasserstein distance and

perceptual loss. IEEE Trans Med Imaging 2017;37(6):1348–57.
[53] Mcdaniel P, Papernot N, Celik ZB. Machine learning in adversarial settings. IEEE Secur Privacy 2016;14(3):68–72.
[54] Sousa LR, Miranda T, Sousa RL, Tinoco J. The use of data mining techniques in rockburst risk assessment. Engineering 2017;3(4):552–8.
[55] Sun Y, Kamel MS, Wong AKC, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 2007;40(12):3358–78.
[56] Farazi AP, DePinho RA. Hepatocellular carcinoma pathogenesis: from genes to environment. Nat Rev Cancer 2006;6(9):674–87.
[57] Arzumanyan A, Reis HMGPV, Feitelson MA. Pathogenic mechanisms in HBVand HCV-associated hepatocellular carcinoma. Nat Rev Cancer 2013;13 (2):123–35.
[58] Mechref Y, Hu Y, Garcia A, Zhou S, Desantos-Garcia JL, Hussein A. Defining putative glycan cancer biomarkers by MS. Bioanalysis 2012;4(20):2457–69.
[59] Tang Z, Varghese RS, Bekesova S, Loffredo CA, Hamid MA, Kyselova Z, et al. Identification of N-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. J Proteome Res 2010;9(1):104–12.
[60] Kronewitter SR, De Leoz MLA, Strum JS, An HJ, Dimapasoc LM, Guerrero A, et al. The glycolyzer: automated glycan annotation software for high performance mass spectrometry and its application to ovarian cancer glycan biomarker discovery. Proteomics 2012;12(15–16):2523–38.
[61] Pierce M, Buckhaults P, Chen L, Fregien N. Regulation of Nacetylglucosaminyltransferase V and Asn-linked oligosaccharide b(1,6) branching by a growth factor signaling pathway and effects on cell adhesion and metastatic potential. Glycoconjugate J 1997;14(5):623–30.
[62] Lau KS, Dennis JW. N-Glycans in cancer progression. Glycobiology 2008;18 (10):750–60.
[63] Saldova R, Royle L, Radcliffe CM, Abd Hamid UM, Evans R, Arnold JN, et al. Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG. Glycobiology 2007;17(12):1344–56.
[64] Noda K, Miyoshi E, Gu J, Gao CX, Nakahara S, Kitada T, et al. Relationship between elevated FX expression and increased production of GDP-L-fucose, a common donor substrate for fucosylation in human hepatocellular carcinoma and hepatoma cell lines. Cancer Res 2003;63(19):6282–9.
[65] Basu PS, Majhi R, Batabyal SK. Lectin and serum-PSA interaction as a screening test for prostate cancer. Clin Biochem 2003;36(5):373–6.
[66] Arnold JN, Saldova R, Hamid UMA, Rudd PM. Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. Proteomics 2008;8(16):3284–93.
[67] Adamczyk B, Tharmalingam T, Rudd PM. Glycans as cancer biomarkers. Biochim Biophys Acta Gen Subj 2012;1820(9):1347–53.
[68] Deguchi K, Keira T, Yamada K, Ito H, Takegawa Y, Nakagawa H, et al. Twodimensional hydrophilic interaction chromatography coupling anionexchange and hydrophilic interaction columns for separation of 2- pyridylamino derivatives of neutral and sialylated N-glycans. J Chromatography A 2008;1189(1–2):169–74.
[69] Siemerink E, Mulder NH, Brouwers AH, Hospers GA. Early prediction of response to sorafenib treatment in patients with hepatocellular carcinoma (HCC) with 18F-fluorodeoxyglucose-positron emission tomography (18F-FDGPET). J Clin Oncol 2008;26(21):1–15.
[70] Holzinger A, Malle B, Kieseberg P, Roth PM, Müller H, Reihs R. Machine learning and knowledge extraction in digital pathology needs an integrative approach lecture notes in computer science. In: Holzinger A, Goebel R, Ferri M, Palade V, editors. Towards integrative machine learning and knowledge extraction. Cham: Springer; 2017.
[71] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[72] Mitchell T. Machine learning. Zeng H, Zhang Y, translator. Beijing: China Machine Press; 2003. Chinese.
[73] Liu P, Zhou Y, Zhou DK, Xue L. Energy Performance Contract models for the diffusion of green-manufacturing technologies in China: a stakeholder analysis from SMEs' perspective. Energy Policy 2017;106:59–67.
[74] Kong D, Feng Q, Zhou Y, Xue L. Local implementation for green-manufacturing technology diffusion policy in China: from the user firms' perspectives. J Cleaner Prod 2016;129:113–24.
[75] Zhou Y, Xu G, Minshall T, Liu P. How do public demonstration projects promote green-manufacturing technologies? A case study from China. Sustainable Dev 2015;23(4):217–31.
[76] Kong D, Zhou Y, Liu Y, Xue L. Using the data mining method to assess the innovation gap: a case of industrial robotics in a catching-up country. Technol Forecasting Social Change 2017;119:80–97.
[77] Li M, Zhou Y. Visualizing the knowledge profile on self-powered technology. Nano Energy 2018;51:250–9.
[78] Wang B, Liu Y, Zhou Y, Wen Z. Emerging nanogenerator technology in China: a review and forecast using integrating bibliometrics, patent analysis and technology roadmapping methods. Nano Energy 2018;46:322–30.