计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):249−256 [doi: 10.15888/j.cnki.csa.008916]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 VAE-CGAN 的牦牛等级评定算法①

李　丹[1], 张玉安[1], 何　杰[1], 陈占琦[1], 宋维芳[2], 宋仁德[3]

[1](青海大学 计算机技术与应用系, 西宁 810016)
[2](门源县畜牧兽医工作站, 海北 812200)
[3](玉树州畜牧兽医工作站, 玉树 815000)
通信作者: 张玉安, E-mail: 2011990029@qhu.edu.cn

摘　要: 在牦牛高效养殖过程中, 牦牛等级评定是牦牛育种工作中的重要环节. 为了在牦牛等级评定研究中, 降低数据集分布不平衡对牦牛等级预测结果的影响, 提出一种基于改进条件生成对抗网络模型的牦牛等级评定模型 VAE-CGAN. 首先, 为获取高质量生成样本, 模型通过引入变分自编码器取代条件生成对抗网络输入中的随机噪声, 降低了随机变量带来的不确定性. 此外, 模型将牦牛标签作为条件信息输入到生成对抗模型中来获取指定类别的生成样本, 生成样本及训练样本则会被用于训练深度神经网络分类器. 实验结果显示, 模型整体预测准确率达到了 97.9%. 而且与生成对抗网络相比较, 在数量较少的特级牦牛等级预测上的精准率、召回率和 $F1$ 值分别提升了 16.7%、16.6% 和 19.4%. 实验结果表明该模型可以实现高精准度和低误分类率的牦牛等级分类.

关键词: 牦牛高效养殖; 牦牛等级预测; 变分自编码器; 条件生成对抗网络; 生成样本; 深度学习; 数据增强

引用格式: 李丹,张玉安,何杰,陈占琦,宋维芳,宋仁德.基于 VAE-CGAN 的牦牛等级评定算法.计算机系统应用,2023,32(1):249–256. http://www.c-s-a.org.cn/1003-3254/8916.html

Grade Evaluation Algorithm of Yak Based on VAE-CGAN

LI Dan[1], ZHANG Yu-An[1], HE Jie[1], CHEN Zhan-Qi[1], SONG Wei-Fang[2], SONG Ren-De[3]

[1](Department of Computer Technology and Applications, Qinghai University, Xining 810016, China)
[2](Menyuan County Animal Husbandry and Veterinary Workstation, Haibei 812200, China)
[3](Yushu Animal Husbandry and Veterinary Workstation, Yushu 815000, China)

Abstract: Yak grade evaluation is an important part of high-efficiency yak breeding. To reduce the influence of imbalanced data set distribution on the prediction results of yak grading in the research, this study proposes a yak grade evaluation model based on an improved conditional generative adversarial network model, called VAE-CGAN. Firstly, to obtain high-quality generated samples, the model reduces the uncertainty from random variables by introducing a variational autoencoder to replace the random noise in the input of the conditional generative adversarial network. In addition, the model inputs the yak label as conditional information into the generative adversarial model to obtain the generated samples of the specified category, and the generated samples and training samples are utilized to train the deep neural network classifier. The experimental results show that the overall prediction accuracy of the model has reached 97.9%. The *Precision*, *Recall*, and $F1$ value on the grade prediction of premium yak have increased by 16.7%, 16.6%, and 19.4% respectively compared with those of the generative adversarial network. The results indicate the model can achieve yak classification with high accuracy and low misclassification rate.

Key words: high-efficiency yak breeding; yak grade prediction; variational autoencoder (VAE); conditional generative adversarial network (CGAN); generated samples; deep learning; data augmentation

① 基金项目: 青海省科技计划 (2020-QY-218); 国家现代农业产业技术体系 (CARS-37)
收稿时间: 2022-05-20; 修改时间: 2022-07-01; 采用时间: 2022-07-22; csa 在线出版时间: 2022-09-14
CNKI 网络首发时间: 2022-11-15

牦牛主要分布在中国西部, 是当地畜牧业经济中不可缺少的重要畜种. 如何进一步提高牦牛的生产性能和繁殖性能, 是牦牛高效养殖中的重要课题[1]. 牦牛选育[2] 是牦牛高效养殖的关键技术之一, 而牦牛等级评定作为牦牛选育工作中的重要环节, 是指根据种公牛的相关性状和指标将种公牛划分为不同的等级, 从而选留体况良好的个体, 有效地提高牦牛的生产效率[3]. 目前, 牦牛等级评定主要是由人工操作, 但是其效率和科技含量较低. 随着科技的发展, 人工智能技术逐渐地应用到养殖领域中, 例如, 牦牛体重预测[4]、牦牛脸识别及牛肉胴体产量等[5,6]. 在采用人工智能技术进行牦牛等级评定过程中, 牦牛等级评定被视为一个多分类问题, 通过将牦牛的等级鉴定信息, 如: 体高、体斜长和外貌等, 输入到分类模型中, 从而将种公牛划分为普通种公牛、优质种公牛和特级种公牛 3 类. 人工智能技术的应用不仅降低了牦牛等级评定过程中的主观性影响和人工误差, 而且能够为之后的牦牛选育研究提供参考.

然而, 在现实的种牦牛场中, 不同等级的牦牛数量分布是不平衡的, 通常普通种公牛和优质种公牛的数量比较多, 而品质优良的特级种公牛的数量却比较少. 如果采用这种数据分布不平衡的牦牛的等级鉴定信息作为分类模型的输入, 对于数量较少的特级种公牛, 其等级鉴定信息会被视为异常数据, 而导致模型不能有效地学习其特征信息, 所以最终的分类模型不能有效地区分特级种公牛的类别, 这极大地影响牦牛等级评定效果. 针对这种数据不平衡问题, 国内外研究人员提出了诸多解决方法. 这些方法主要分为从数据层面进行优化和从算法层面进行优化两大类. 一类从数据层面进行优化是指通过不同的采样方式优化数据分布从而减少类别不平衡, 例如, 欠采样、过采样和混合采样. 经典的采样算法有 SMOTE[7]、Borderline-SMOTE 方法[8] 和 NCL 方法[9] 等. 然而, 基于采样方法仍然存在很多问题, 例如, 欠采样方法容易造成了信息丢失, 过采样方法容易出现过度拟合. 另一类从算法层面进行优化的研究热点是代价敏感学习[10], 其目标是最小化总期望的误分类代价, 虽然它效果明显且移植性强, 但是难以确定误分类代价和有效地评估分类器的性能.

综上所述, 为了解决因牦牛等级数据分布不平衡而导致的无法区分数量较少的特种牦牛问题. 本文提出了一种基于 VAE-CGAN 的牦牛等级评定算法. VAE-CGAN 是一种基于生成模型的算法, 它将牦牛等级数据作为原始样本输入到生成模型得到与原始样本分布近似的生成样本, 从而解决牦牛等级数据分布不平衡问题, 实现有效的牦牛等级评定[11].

1 VAE-CGAN 模型

VAE-CGAN 模型是由 3 个相互独立的模块组成: (a) 变分自编码器 (variational autoencoder, VAE); (b) 条件生成对抗网络 (conditional generative adversarial network, CGAN); (c) 分类器 (classifier, *C*). VAE-CGAN 模型框架如图 1 所示.

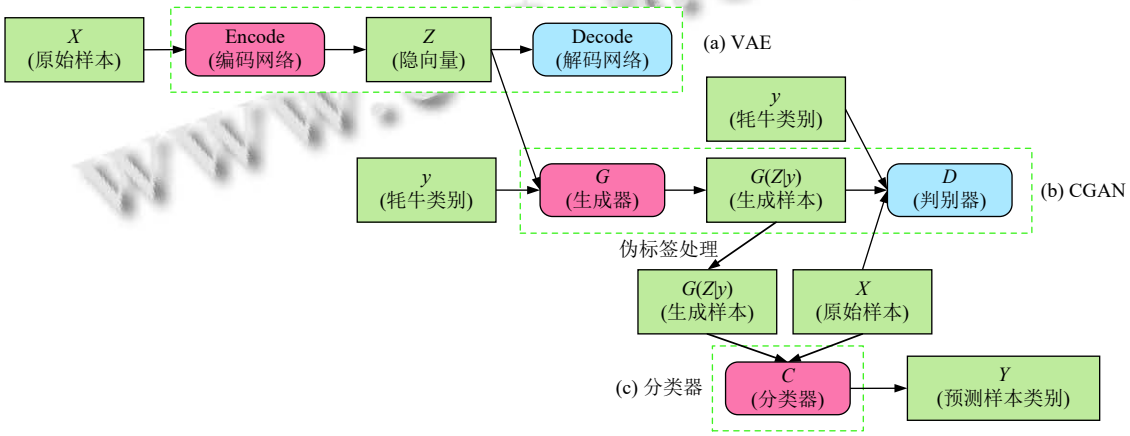

图 1 VAE-CGAN 模型框架

VAE-CGAN 将牦牛等级数据作为原始样本 X 进行输入. 原始样本 X 首先经过 VAE、CGAN 和伪标签处理得到与原始样本 X 分布近似的生成样本 $G(Z|y)$ 来解决牦牛等级数据分布不平衡问题, 然后将原始样本

和生成样本输入到分类器中实现牦牛等级评定. VAE-CGAN 的具体训练过程主要分为 3 步. 首先, 将原始样本 X 输入变分自编码器 VAE, 利用编码网络提取原始样本 X 的潜在特征向量 Z. 其次, 将潜在特征向量 Z 和原始样本的类别信息 y 输入到 CGAN, 通过生成器 G 得到生成样本 $G(Z|y)$. 生成样本 $G(Z|y)$、原始样本 X 和其类别信息 y 作为判别器 D 的输入, 在生成器 G 和判别器 D 的对抗生成过程中获取更高质量的生成样本 $G(Z|y)$. 最后, 使用伪标签技术生成生成样本 $G(Z|y)$ 的伪标签, 并将含有伪标签的生成样本 $G(Z|y)$、原始样本 X 和类别信息 y 输入到分类器 C 中实现牦牛等级评定.

与传统 CGAN 采用随机变量作为输入来产生生成样本相比, VAE-CGAN 采用由 VAE 训练生成的含有牦牛等级和特征信息的隐向量 Z 作为输入[12], 降低了使用随机向量作为输入所带来的不确定性, 可以更好地收集牦牛的全局特征信息. 此外, VAE-CGAN 将牦牛类别信息作为条件 y 输入到 CGAN 模块中进行生成对抗训练, 使含有牦牛等级信息和特征信息的生成样本更接近于真实样本[13]. 除此之外, VAE-CGAN 采用伪标签技术处理后的带有标签信息生成样本, 实现了生成样本由无监督学习到监督学习的转换[14]. VAE-CGAN 最终采用生成样本 $G(Z|y)$ 与原始样本 X 作为训练集一起训练分类器, 实现了有效牦牛等级评定. 下面将详细介绍 VAE-CGAN 模型的 VAE 模块、CGAN 模块、分类器模块的设计与实现.

1.1 变分自编码器 VAE

VAE[15] 是一种学习样本分布的方法, 它采用估计分布近似逼近样本的真实分布, 从而生成与样本分布类似的重构样本. 在 VAE-CGAN 中引入 VAE, 通过最小化原始样本与重构样本之间的重构损失和原始样本与生成样本之间分布的差异, 从而训练编码网络和解码网络来获取包含原始样本抽象特征和重要信息的隐向量 Z. VAE 的网络结构如图 2 所示.

VAE 的训练主要分为 3 个阶段.

(1) 编码阶段

首先将牦牛等级评定数据 $X=[X_1,\cdots,X_i,\cdots,X_n]$ 作为样本输入到模型, 其中 $X_i\in R^{1\times d}$, $X\in R^{n\times d}$, $d=13$ 表示每条牦牛等级信息含有 13 个特征. 通过在编码网络中完成线性映射及非线性激活, 实现样本的编码过程. 计算公式如式 (1):

$$\mu,\sigma=g(W_1\cdot X+b_1) \tag{1}$$

其中, W_1 为编码网络的权重矩阵, b_1 为节点偏置, g 是激活函数, μ 为均差, σ 为标准差.

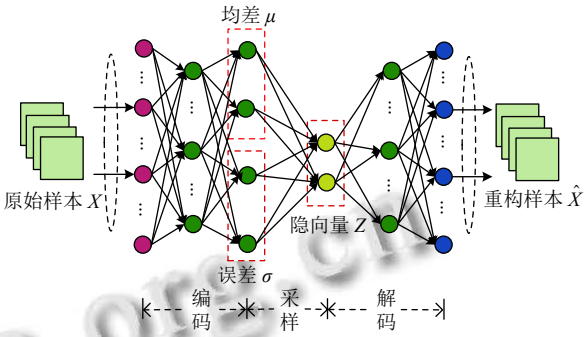

图 2 VAE

(2) 采样阶段

通过引入重参数化技术将隐向量 Z 的随机采样转换为线性运算, 克服了训练过程因随机采样导致无法利用反向传播优化模型参数问题, 使得模型参数参与梯度计算. 计算式如式 (2) 所示:

$$Z=\mu+\sigma\cdot\varepsilon,\ \varepsilon\sim N(0,1) \tag{2}$$

其中, Z 是 VAE 模型的隐向量, ε 为高斯噪声.

(3) 解码阶段

其过程与编码阶段类似, 旨在将生成的隐向量通过解码过程, 获取重构后的样本.

VAE 的训练目标分别是最小化样本重构的损失和最小化学习到的隐向量分布与正态分布之间的距离. 由于无法直接估计原始数据 X 的分布, 模型将 VAE 中隐向量 Z 的先验分布看作标准正态分布, 并且采用 KL 散度衡量先验分布 $N(Z;0,1)$ 与隐向量分布 $N\left(Z;\mu,\sigma^2\right)$ 的距离. VAE 模型最终的目标函数由重构误差和 KL 散度两部分共同组成. 如式 (3) 所示:

$$L_{\mathrm{VAE}}=\|X-\hat{X}\|^2+D_{\mathrm{KL}}\left(N\left(Z;\mu,\sigma^2\right)\|N(Z;0,1)\right) \tag{3}$$

其中, X 为原始样本, $\hat{X}$ 为重构样本, N 为正态分布函数.

1.2 条件生成对抗网络 CGAN

CGAN[16] 是一种在 GAN[17] 基础上进行改进的生成式网络, 它由生成器 G 和判别器 D 组成, 其中生成器主要被用于生成与原始样本相似的生成样本, 而判别器主要被用于分辨输入数据的真实性和数据的类别. CGAN 在训练过程中分别对生成器网络和判别器网络中的参数进行更新和优化. 在 VAE-CGAN 中引进 CGAN 是为了生成充足且符合条件 y 的生成样本 $G(Z|y)$, 从

而解决牦牛等级数据的分布不平衡问题. CGAN 的网络结构如图 3 所示.

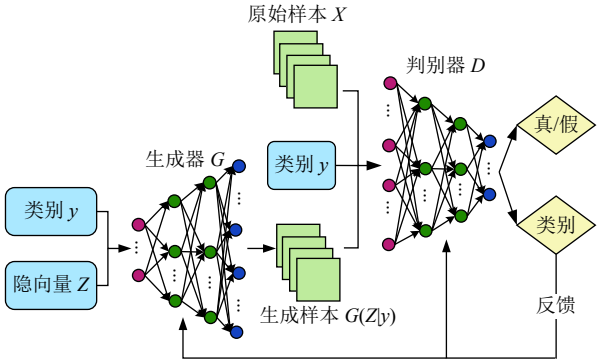

图 3 CGAN

CGAN 的训练主要分为正向传播和反向传播两个阶段.

(1) 正向传播阶段

正向传播阶段, 首先, VAE-CGAN 在通过 VAE 模块的编码网络获取含有抽象特征的隐向量 Z 后, 将隐向量 Z 和牦牛类别信息 y 作为生成器模块的输入, 输入数据通过生成网络生成指定类别的生成样本 $G(Z|y)$. 然后, 判别器 D 将生成样本 $G(Z|y)$、原始样本 X 和条件 y 作为输入, 通过判别网络判断生成样本是否符合条件 y 及生成样本的真实性, 从而输出某一类别真假概率.

(2) 反向传播阶段

反向传播阶段, 通过计算模型交叉熵损失, 最小化损失函数对模型参数进行优化. 优化过程分为两步, 首先, 优化生成器 G, 其目标是使 $G(Z|y)$ 越真实越好, 即 $D(G(Z|y))$越大越好. 然后, 优化判别器 D, 其目标是能够准确地分辨 $(X|y)$ 和 $G(Z|y)$, 所以, $D(Z|y)$ 应该变大, $D(G(Z|y))$应该变小. 在整个优化训练过程中, G 和 D 的优化交替迭代进行[18]. CGAN 中生成器 G 和判别器 D 的损失函数如式 (4)、式 (5) 所示:

$$L_G = E_{Z\sim P_Z}\log(1-D(G(Z|y))) \tag{4}$$

$$L_D = -E_{X\sim P_r}\log D(X|y) - E_{Z\sim P_Z}\log(1-D(G(Z|y))) \tag{5}$$

CGAN 最终的目标函数如式 (6) 所示:

$$\max_D V(D,G) = E_{X\sim P_r}\log D(X|y) + E_{Z\sim P_Z}\log(1-D(G(Z|y))) \tag{6}$$

1.3 分类器 C

分类器 C 是一个 4 层的全连接神经网络, 其输入层含有 13 个神经元, 在两个隐藏层对输入样本进行线性运算和 ReLU 函数非线性激活处理, 在输出层经过 Softmax 函数处理, 最终输出分类类别的概率. 分类器 C 的网络结构如图 4 所示.

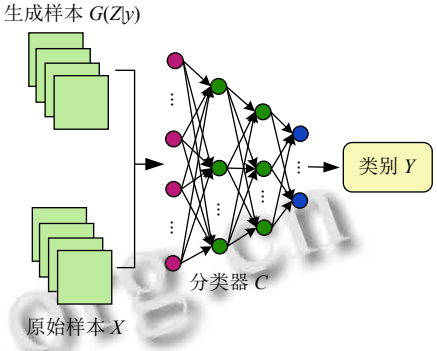

图 4 分类器

在 CGAN 模块得到生成样本 $G(Z|y)$ 后, 首先, 使用伪标签技术处理生成样本, 从而得到含有伪标签的生成样本 $G(Z|y)$. 然后, 将含有伪标签生成样本 $G(Z|y)$ 和原始样本 X 共同输入到分类器 C 中实现牦牛等级评定. 分类器 C 使用交叉熵损失函数为作为目标函数[19], 其目标函数如式 (7) 所示:

$$L_{\text{class}} = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{7}$$

其中, M 表示类别的数量, y_{ic} 等于 1 时, 代表 i 的类别为 c, 预测正确. y_{ic} 等于 0 时, 代表 i 的类别不是 c, 预测错误. p_{ic} 表示预测样本 i 属于类别 c 的概率.

1.4 牦牛等级评定算法 VAE-CGAN

根据上述 VAE-CGAN 算法 3 个主要模块的训练过程, 给出 VAE-CGAN 算法实现的伪代码如算法 1.

算法 1. VAE-CGAN 算法

输入: 牦牛体况数据集 X, 牦牛等级类别 y, 迭代次数 $iter$;
输出: 牦牛等级预测结果 Y;

1) 初始化模型 VAE 模块、CGAN 模块及分类器 C;
2) 采用 1/10 的训练集训练分类器 C;
3) for t in 1:$iter$ do:
4) 以 X 作为输入, VAE 网络将其编码为向量 Z;
5) 根据式 (3) 更新 VAE 网络参数;
6) 以 Z 和 y 作为 G 的输入, 以 $G(Z|y)$、Z 和 y 作为 D 的输入, 利用 CGAN 进行数据生成;
7) 根据式 (5) 更新 G 的网络参数, 根据式 (6) 更新 D 的网络参数;
8) 使用分类器 C 对无标签数据 $G(Z|y)$ 进行预测, 得出预测概率 P, 通过预测概率 P 筛选高置信度样本;
9) 将高置信度样本和训练集作为 C 的输入, 分类网络输出牦牛等级类别 Y;

10) 根据式 (7) 更新分类网络参数;
11) end for

2 实验与结果分析

本文实验的训练与测试是基于深度学习框架 PyTorch 实现的, 软件编程环境为 Python 3.7. 计算机的操作系统为 Windows 10, CPU 为 Intel core i5-9400, 内存为 16.0 GB. 模型在 VAE、CGAN 及分类器模块的迭代次数分别设置为 5 000 次、500 次及 4 000 次, 使用的优化器为 Adam, 学习率均设置为 0.000 5.

2.1 数据集及数据预处理

本文的实验数据集来自于青海省海西蒙古族藏族自治州的天峻县种牛场. 数据集中共包含 4 954 条 1–2 岁种公牛的等级鉴定数据, 每条鉴定数据包含 1 个牦牛等级信息 (total degree, TD) 和 13 个牦牛特征信息. 数据集如表 1 所示, 对于牦牛等级信息, 如表 1 的最后 1 列, 分别表示: 0 普通种公牛、1 优质种公牛和 2 特级种公牛, 且 3 个类别之间的数量比为 127:118:1. 对于牦牛特征信息, 如表 1 的前 13 列, 分别表示: 体高 (height, H)[20] 为鬐甲最高点与地面之间的垂直距离; 体斜长 (body length, BL) 为肩端至坐骨端的距离; 管围 (circumference of cannon bone, CCB) 为管骨最细处的周长; 胸围 (chest girth, CG) 为肩胛骨后角垂直体轴绕胸一周的周长; 体重 (weight, W) 为空腹状态下牦牛的体重; 外貌 (appearance, A) 的评分范围为 0–30 分; 体躯 (body, B) 的评分范围为 0–30 分; 生殖器 (genital, G) 的评分范围为 0-10 分; 肢蹄 (food, F) 的评分范围为 0–20 分; 被毛 (hair, H) 的评分范围为 0–20 分; 整体评分 (total, T) 的评分范围为 0–100 分; 体高等级 (height degree, HD) 分别由 0、1、和 2 表示; 体重等级 (weight degree, WD) 分别由 0、1、和 2 表示. 其中, 0、1 和 2 分别表示普通、优质和特级 3 类.

表 1 原始数据

H (cm)	BL (cm)	CCB (cm)	CG (cm)	W (kg)	A (分)	B (分)	G (分)	F (分)	H (分)	T (分)	HD	WD	TD
108	114	15	140	156	22	24	8	14	14	82	0	0	0
110	116	15	142	164	23	24	9	13	15	84	1	0	1
112	121	19	169	248	25	23	8	12	18	86	1	2	2

由于在牦牛等级鉴定数据采集过程中会出现数据缺失和重复等问题, 因此, 在训练模型之前需要进行数据的预处理操作, 它主要包含 3 个步骤: 数据清洗、数据归一化和数据转化. 首先, 采用众数填充缺失值和删除重复行等数据清洗操作. 然后, 由于数据集中的各个特征的取值范围差距较大, 所以需要通过 min-max 标准化方法进行数据归一化操作, 这极大地降低数据不规范对算法准确率的影响. 最后, 由于模型将数据集中的特征信息作为输入, 总体等级信息作为输出, 所以对于包含了模型输入和输出数据集, 需要通过数据转化操作将其划分为输入数据和输出数据两个部分. 数据预处理后的数据格式见表 2.

表 2 处理后的数据

H (cm)	BL (cm)	CCB (cm)	CG (cm)	W (kg)	A (分)	B (分)	G (分)	F (分)	H (分)	T (分)	HD	WD	TD
0.33	0.18	0.37	0.24	0.19	0.53	0.62	0.67	0.20	0.30	0.64	0	0	0
0.43	0.38	0.37	0.25	0.22	0.60	0.62	1.00	0.30	0.35	0.70	0.5	0	1
0.52	0.75	0.87	0.43	0.57	0.73	0.50	0.67	0.40	0.50	0.77	0.5	1	2

2.2 评估指标

在不平衡数据集中, 数量少的类别特征难以学习, 一般预测精度较低. 所以, 评价指标不仅要客观的评价整体的预测精准度, 还要兼顾少数类别的预测精准度. 本文算法在牦牛等级分类效果上评估指标主要有: 准确率 (*Accuracy*)、召回率 (*Recall*)、*F*1 值、精确率 (*Precision*)[21]. 根据真实标签和预测结果可得分类结果的混淆矩阵, 由此计算上述评价指标, 混淆矩阵见表 3.

表 3 分类结果的混淆矩阵

	预测正例	预测反例
真实正例	*TP*	*FN*
真实反例	*FP*	*TN*

Accuracy 表示数据集预测正确的总样本与预测总样本之间的比例, 计算公式如式 (8):

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{8}$$

Recall 表示在正例中预测正确的样本与正例总样本之间的比例, 计算公式如式 (9):

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

*F*1值是 *Accuracy* 和 *Recall* 的加权平均, 更公平地评价算法的性能. 计算公式如式 (10) 所示:

$$F1 = 2\times\frac{Precision\times Recall}{Precision+Recall} \tag{10}$$

Precision 表示在预测正例中预测正确的样本与预测总样本之间的比例, 计算公式如式 (11):

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

2.3 实验分析

本节首先展示 VAE-CGAN、GAN 和 VAE 模型的训练过程, 确定模型的训练参数; 然后, 比较不同分类器在 VAE-CGAN 上的分类效果, 选定 VAE-CGAN 模型的分类器; 最后, 基于选定分类器, 对 VAE-CGAN、GAN 和 VAE 模型在牦牛等级数据集上等级分类试验结果进行对比分析. 此外, 实验采用 70% 数据集作为训练集, 30% 数据集作为测试集.

2.3.1 模型训练

VAE-CGAN、GAN 和 VAE 模型训练过程的损失函数曲线分别如图 5、图 6 和图 7 所示. 由图可知, 对于 VAE-CGAN 和 GAN 模型, 随着迭代次数的增加, 生成器的生成样本质量越接近真实样本, 所以生成器的损失值曲线呈下降趋势, 而判别器不能有效的分辨生成样本$G(Z|y)$和原始样本X, 所以判别器的损失值曲线呈上升趋势. 当 VAE-CGAN、GAN 和 VAE 模型的训练次数分别达到 100、100 和 800 时, 模型的损失曲线趋于稳定, 损失误差收敛, 模型训练完成.

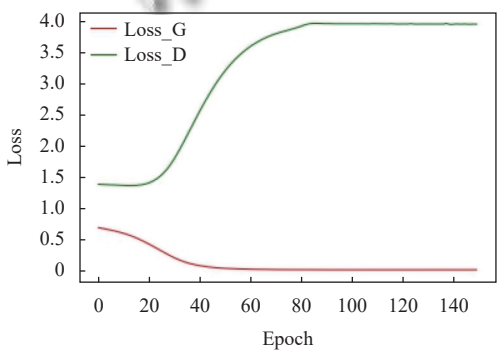

图 5 VAE-CGAN 损失函数曲线

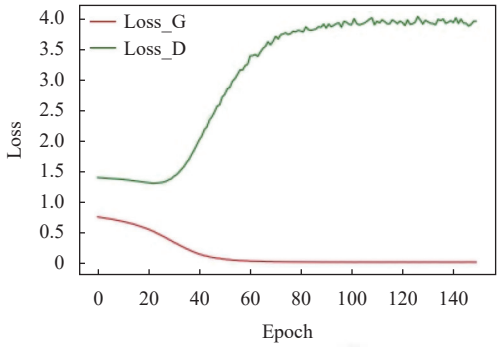

图 6 GAN 损失函数曲线

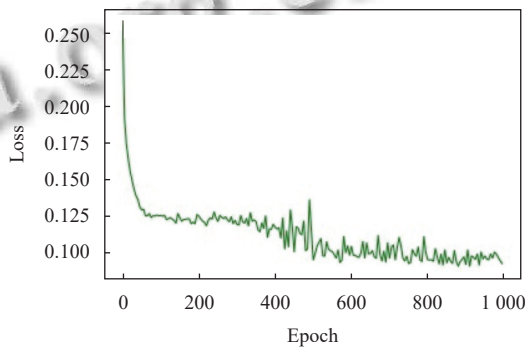

图 7 VAE 损失函数曲线

2.3.2 分类器选取

为了选取一个分类效果好的分类器, 分别采用 Logic[22]、SVM[23] 和 ANN 三种分类器, 在 VAE-CGAN 模型进行分类实验对比. 首先, 获得模型参数并保存 VAE 和 CGAN 网络中的参数. 然后, 将真实样本输入到 VAE 和 CGAN 网络中获取生成样本, 利用伪标签技术对生成样本进行处理. 最后, 将生成样本与真实样本分别训练 Logic、SVM 和 ANN 分类器, 计算出模型的评价指标. 实验结果见表 4.

表 4 不同分类器分类结果

算法	类别	*Precision* (%)	*Recall* (%)	*F*1 (%)	*Accuracy* (%)
Logic	0	92.0	98.2	95.0	94.4
	1	97.1	90.5	93.7	
	2	0	0	0	
SVM	0	94.4	98.9	96.6	96.0
	1	98.0	93.4	95.6	
	2	50.0	16.7	25.0	
ANN	0	98.3	98.2	98.2	97.9
	1	97.6	98.1	97.8	
	2	66.7	33.3	44.4	

由表 4 可知, Logic 的评价指标最低, ANN 在 3 个类别和 4 个评价指标上, 都取得了优越的效果. 尤其在特级牦牛的预测中, ANN 的准确率 *Accuracy* 比 Logic

提高了 3.5%、比 SVM 提高了 1.9%, *F*1 值比 Logic 提高了 44.4%, 比 SVM 提高了 19.4%. 这体现了 ANN 作为反馈式神经网络, 能够高效地利用特征信息, 和较强学习能力. 因此在实验选取 ANN 作为 VAE-CGAN 模型分类器.

2.3.3 数据生成对比实验

通常, 生成模型生成的样本质量越好, 则牦牛等级评定分类效果越好. 所以, 实验选取目前主流的生成模型 GAN 和 VAE 与 VAE-CGAN 进行对比, 通过比较不同生成模型在测试集上等级评定分类效果, 来评估 VAE-CGAN 的优越性. 牦牛等级评定分类结果见表 5.

表 5 分类对比结果

算法	类别	*Precision* (%)	*Recall* (%)	*F*1 (%)	*Accuracy* (%)
VAE	0	86.4	98.9	92.2	90.9
	1	97.6	82.6	89.5	
	2	0	0	0	
GAN	0	98.4	96.6	97.5	96.0
	1	96.0	96.0	96.0	
	2	14.3	50.0	22.2	
VAE-CGAN	0	98.3	98.2	98.2	97.9
	1	97.6	98.1	97.8	
	2	66.7	33.3	44.4	

由表 5 可知, VAE-CGAN 在 3 个类别和 4 个评价都取得了优越的效果. 与 VAE 模型相比较, VAE-CGAN 和 GAN 在 *F*1 和 *Accuracy* 指标上显著优于 VAE 模型, 即, 模型在单个类别和全部类别上均实现了优越分类性能, 说明引入生成对抗网络可以有效地提高生成样本质量. 与 GAN 模型相比, VAE-CGAN 各个类别上的 *F*1 值都高于 GAN, 尤其在特级牦牛等级预测上的 *F*1 值比 GAN 提高 22.2%, 说明与 GAN 使用随机噪声作为输入相比, VAE-CGAN 采用含有特征信息的隐向量 Z 和牦牛类别信息 y 作为输入, 可以更好地保留牦牛的全局特征信息. 综上所述, VAE-CGAN 可以区分数量较少的特级牦牛, 在牦牛等级评定实验中取得了优越的性能.

2.3.4 参数敏感度分析

为了评估了 VAE- CGAN 上不同训练集占比对牦牛等级评定准确率的影响, 分别选取 20%–90% 的数据集作为训练集进行实验. 实验结果如图 8 所示.

由图 8 可知, 随着训练集占比的增加, 对于普通牦牛 (0 类) 和优质牦牛 (1 类), *F*1 曲线呈不断上升趋势, 当训练集为 70% 时趋于稳定, 达到 95% 左右. 对于特级牦牛 (2 类), 采用低于 50% 的训练集不能区分特级牦牛, 将训练集由 50% 增加到 70% 时, 特级牦牛 *F*1 值逐渐增加并在 70% 处达到最大值, 继续增大训练集, *F*1 值开始降低. 因此, 实验采用 70% 的数据集作为训练集对 VAE- CGAN 进行训练.

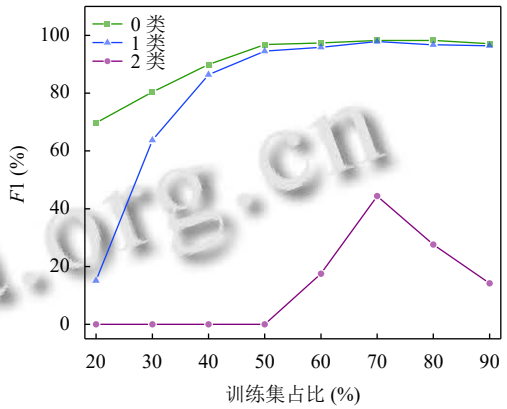

图 8 训练集对 *F*1 的影响

3 结论与发展

为了降低牦牛等级数据分布不平衡对牦牛等级评定的影响, 本文提出基于 VAE-CGAN 的牦牛等级评定算法. 实验结果表明, 与现有的一些生成模型 VAE 和 GAN 相比, VAE-CGAN 不仅达到 97.9% 等级评定分类准确率, 而且在区分特级牦牛性能上得到显著的提升. VAE-CGAN 能够有效地解决数据分布不平衡问题. 在未来研究工作中, 可以从以下 3 个方面进一步改进: (1) 在模型中加入注意力模块, 对网络参数的特征权重进行优化. (2) 加入参数寻优算法, 使模型能够自主寻找最优参数. (3) 优化训练时间, 在提高模型准确率的同时缩减训练时间.

参考文献

1 袁凯鑫, 王昕. 青海高原大通牦牛的育种进展. 中国牛业科学, 2019, 45(1): 28–32. [doi: 10.3969/j.issn.1001-9111.2019.01.008]

2 张江. 牦牛高效养殖关键技术. 畜牧兽医科技信息, 2020, (4): 99. [doi: 10.3969/J.ISSN.1671-6027.2020.04.089]

3 骆正杰, 马进寿, 保广才, 等. 青海省牦牛种业发展现状、存在问题及应对策略. 中国畜牧杂志, 2021, 57(2): 231–234. [doi: 10.19556/j.0258-7033.20200429-02]

4 Yan Q, Ding LM, Wei HY, *et al*. Body weight estimation of yaks using body measurements from image analysis.

Measurement, 2019, 140: 76–80. [doi: 10.1016/j.measurement.2019.03.021]
5 陈争涛, 黄灿, 杨波, 等. 基于迁移学习的并行卷积神经网络牦牛脸识别算法. 计算机应用, 2021, 41(5): 1332–1336.
6 Wakholi C, Kim J, Nabwire S, *et al*. Deep learning feature extraction for image-based beef carcass yield estimation. Biosystems Engineering, 2022, 218: 78–93. [doi: 10.1016/j.biosystemseng.2022.04.008]
7 Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on K-means and SMOTE. Information Sciences, 2018, 465: 1–20. [doi: 10.1016/j.ins.2018.06.056]
8 Bennin KE, Keung J, Phannachitta P, *et al*. MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. IEEE Transactions on Software Engineering, 2018, 44(6): 534–550. [doi: 10.1109/TSE.2017.2731766]
9 Laurikkala J. Improving identification of difficult small classes by balancing class distribution. Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe. Cascais: Springer, 2001. 63–66.
10 Roy NKS, Rossi B. Cost-sensitive strategies for data imbalance in bug severity classification: Experimental results. Proceedings of the 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA). Vienna: IEEE, 2017. 426–429.
11 叶德豪, 王琼. 面向不平衡数据的生成对抗网络研究. 工业控制计算机, 2021, 34(5): 95–96. [doi: 10.3969/j.issn.1001-182X.2021.05.039]
12 Gao R, Hou XS, Qin J, *et al*. Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning. IEEE Transactions on Image Processing, 2020, 29: 3665–3680. [doi: 10.1109/TIP.2020.2964429]
13 Dideriksen BU, Derosche K, Tan ZH. iVAE-GAN: Identifiable VAE-GAN models for latent representation learning. IEEE Access, 2022, 10: 48405–48418. [doi: 10.1109/ACCESS.2022.3172333]
14 Haque A. EC-GAN: Low-sample classification using semi-supervised algorithms and GANs. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Online: AAAI, 2021. 15797–15798.
15 Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114, 2013.
16 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784, 2014.
17 邹秀芳, 朱定局. 生成对抗网络研究综述. 计算机系统应用, 2019, 28(11): 1–9. [doi: 10.15888/j.cnki.csa.007156]
18 于龙泽, 肖白, 孙立国. 风光出力场景生成的条件深度卷积生成对抗网络方法. 东北电力大学学报, 2021, 41(6): 90–99. [doi: 10.19718/j.issn.1005-2992.2021-06-0090-10]
19 李梦磊, 刘新, 赵梦凡, 等. 基于语句结构信息的方面级情感分类. 计算机系统应用, 2020, 29(11): 114–120. [doi: 10.15888/j.cnki.csa.007681]
20 刘丽元, 臧长江, 周靖航, 等. 新疆昌吉地区荷斯坦奶牛生长发育规律分析. 中国畜牧兽医, 2015, 42(8): 2036–2041. [doi: 10.16431/j.cnki.1671-7236.2015.08.017]
21 陈巧红, 王磊, 孙麒, 等. 卷积神经网络的短文本分类方法. 计算机系统应用, 2019, 28(5): 137–142. [doi: 10.15888/j.cnki.csa.006887]
22 Shah K, Patel H, Sanghvi D, *et al*. A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 2020, 5(1): 12. [doi: 10.1007/s41133-020-00032-0]
23 尚晖. 基于改进 SVM 的互联网用户分类. 计算机系统应用, 2021, 30(4): 266–270. [doi: 10.15888/j.cnki.csa.007914]

(校对责编: 牛欣悦)