一种能量函数意义下的生成式对抗网络

王功明 [1,2] 乔俊飞 [1,2] 王 磊 [1,2]

摘 要 生成式对抗网络 (Generative adversarial network, GAN) 是目前人工智能领域的一个研究热点, 引起了众多学者的关注. 针对现有 GAN 生成模型效率低下和判别模型的梯度消失问题, 本文提出一种基于重构误差的能量函数意义下的生成式对抗网络模型 (Energy reconstruction error GAN, E-REGAN). 首先, 将自适应深度信念网络 (Adaptive deep belief network, ADBN) 作为生成模型, 来快速学习给定样本数据的概率分布并进一步生成相似的样本数据. 其次, 将自适应深度自编码器 (Adaptive deep autoencoder, ADAE) 的重构误差 (Reconstruction error, RE) 作为一个表征判别模型性能的能量函数, 能量越小表示 GAN 学习优化过程越趋近纳什均衡的平衡点, 否则反之. 同时, 通过反推法给出了 E-REGAN 的稳定性分析. 最后在 MNIST 和 CIFAR-10 标准数据集上的实验结果表明, 相较于现有的类似模型, E-REGAN 在学习速度和数据生成能力两方面均有较大提高.

关键词 生成式对抗网络, 能量函数, 重构误差, 自适应深度信念网络, 自适应深度自编码器, 纳什均衡

引用格式 王功明, 乔俊飞, 王磊. 一种能量函数意义下的生成式对抗网络. 自动化学报, 2018, **44**(5): 793−803

DOI 10.16383/j.aas.2018.c170600

A Generative Adversarial Network Based on Energy Function

WANG Gong-Ming[1,2] QIAO Jun-Fei[1,2] WANG Lei[1,2]

Abstract Generative adversarial network (GAN) has become a hot research in artificial intelligence, and has received much attention from scholars. In view of low efficiency of generative model and gradient disappearance of discriminative model, a GAN based on energy function (E-REGAN) is proposed in this paper, in which reconstruction error (RE) acts as the energy function. Firstly, an adaptive deep belief network (ADBN) is presented as the generative model, which is used to fast learn the probability distribution of given sample data and further generate new data with similar probability distribution. Secondly, the RE in adaptive deep auto-encoder (ADAE) acts as an energy function evaluating the performance of discriminative model; the smaller energy function, the closer to Nash equilibrium the learning optimization process of GAN will be, and vice versa. Meanwhile, the stability analysis of the proposed E-REGAN is given using the inverse inference method. Finally, the simulation results from MNIST and CIFAR-10 benchmark dataset experiments show that, compared with the existing similar models, the proposed E-REGAN achieves significant improvement in learning rate and data generation capability.

Key words Generative adversarial network (GAN), energy function, reconstruction error (RE), adaptive deep belief network (ADBN), adaptive deep auto-encoder (ADAE), Nash equilibrium

Citation Wang Gong-Ming, Qiao Jun-Fei, Wang Lei. A generative adversarial network based on energy function. *Acta Automatica Sinica*, 2018, **44**(5): 793−803

生成式对抗网络 (Generative adversarial network, GAN) 是由 Goodfellow 等[1] 于 2014 年根据对抗竞争思想提出来的一种优化生成模型. GAN 要解决的问题是如何从训练样本中学习到概率分布特征, 并进一步生成新样本数据, 训练样本是图片即生成新图片, 训练样本是文字即输出新文字. GAN 在学习算法上受博弈论中的零和博弈 (参与博弈的各方在严格竞争下, 一方的收益必然意味着另一方的损失, 博弈各方的收益和损失相加总和永远为零, 双方不存在合作的可能) 的启发, 网络由一个生成模型 (Generative model) 和一个判别模型 (Discriminative model) 构成. 生成模型的训练目的是试图生成与训练样本具有一致概率分布的新样本, 并作为判别模型的输入; 判别模型的训练目的是判断生成模型生成的样本是否是真实的训练样本[2−4]. GAN 的优化和训练过程是一个极小极大的博弈问题, 最终的目的是判别模型无法分辨出生成样本的真伪, 即极大化判别模型的判断能力,极小化生成模型输出被

收稿日期 2017-10-31 录用日期 2017-12-23
Manuscript received October 31, 2017; accepted December 23, 2017
国家自然科学基金 (61533002) 资助
Supported by National Natural Science Foundation of China (61533002)
本文责任编委 王坤峰
Recommended by Associate Editor WANG Kun-Feng
1. 北京工业大学信息学部 北京 100124 2. 计算智能与智能系统北京市重点实验室 北京 100124
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124

判断为伪造的概率[5−6].

然而, GAN 的优化和训练过程中也存在一些缺陷. 一方面, 基于传统深度学习方法的生成模型无法快速地学习并生成样本数据, 即学习效率低、算法收敛慢. 另一方面基于梯度下降算法的训练存在梯度消失的问题, 即当真实样本和生成样本之间具有极小重叠甚至没有重叠时, 其目标函数的 Jensen-Shannon 散度是一个常数, 导致优化目标不连续[1, 7−8]. 为了解决以上问题, Arjovsky 等[9] 提出了一种 Wasserstein GAN (W-GAN) 模型, 利用 Earth-Mover 代替 Jensen-Shannon 散度来表征真实样本和生成样本分布之间的差异, 用一个评价函数来对应 GAN 的判别模型, 而且评价函数需要建立在 Lipschitz 连续性假设上. 尽管 W-GAN 避开了优化目标不连续的障碍, 但是其存在的最大问题是模型的收敛性无法保证. Donahue 等[10] 提出一种 Bi-GAN 模型, 将复杂数据映射到隐变量空间, 从而实现特征学习, 并且引入了一个解码器用于将真实数据映射到隐变量空间, 成功地实现了优化问题的等效转移. 尽管能够部分地解决问题, 但是 Bi-GAN 的训练耗时相对较长.

LeCun 等[11] 提出了能量模型的概念, 并将待优化参数的每一种取值对应于一个能量取值, 通过最小化能量来获取对应的参数取值. 这种能量模型概念的提出有助于提高 GAN 的学习效率, 因为该方法无需在损失函数中加入正则化项也能较精确地完成训练目标. 同时, 能量函数模型无需计算复杂的配分函数 (Partition functions). 当传统浅层神经网络用作判别模型时, 能量函数的选择相对简单, 即将实际输出和期望输出的差值作为能量函数. 然而, 当判别模型用深度学习模型表述时, 由于深度学习模型的训练方式大都采用无监督学习[12], 所以能量函数的选取及其稳定性分析比较困难.

针对以上问题, 本文提出一种基于重构误差的能量函数的 GAN 模型 (Energy reconstruction error GAN, E-REGAN). 生成模型由基于深度学习模型的自适应深度信念网络 (Adaptive deep belief network, ADBN) 来实现, 判别模型由自适应深度自编码器 (Adaptive deep auto-encoder, ADAE) 来实现. 其中, ADBN 的输入是真实样本 x 和噪音 z 的组合, 自适应学习率能够加快生成模型和判别模型的学习速度[13−15], ADAE 的重构误差作为能量函数. 在 MNIST 和 CIFAR-10 标准数据集上的实验结果表明, 与现有的几种类似 GAN 模型相比, E-REDBN 模型在学习速度和数据生成能力两方面均有较大提高.

本文结构安排如下: 第 1 节介绍介绍生成式对抗网络; 第 2 节介绍 E-REGAN 模型, 包括学习过程和网络性能分析; 第 3 节给出实验研究; 第 4 节对本文工作进行总结.

1 生成式对抗网络

GAN 学习原理启发于博弈论中的二人零和博弈 (Two-player game). GAN 模型中的博弈双方分别由生成式模型和判别式模型来实现. 生成模型 G 用来学习已有样本的概率分布, 并试图生成与已有样本一致分布的数据, 通常用到的方法的是利用服从某一分布 (高斯分布或均匀分布) 的噪声 z 生成一个类似真实训练数据的样本, 越逼近真实样本越好. 判别模型 D 是一个二分类器, 估计一个样本来自于训练数据 (而非生成数据) 的概率. 如果样本来自于真实的训练数据, D 输出大概率; 否则, D 输出小概率. GAN 的训练过程即不断的调整 G 和 D, 直到 D 不能把生成的样本从真实样本中区分出来为止. 在调整过程中, 需要做到: 1) 优化 G, 使它尽可能地生成让 D 无法区分的样本; 2) 优化 D, 使它尽可能地区分出生成的样本. 当 D 无法区分出生成的样本时, 可以认为 G 达到最优状态.

假定已有的样本数据为 x, 生成的样本数据为 $G(z)$, 那么生成模型 G 和判别模型 D 的损失函数定义如下:

$$F_G(z) = D(G(z)) \tag{1}$$

$$F_D(x,z) = D(x) + \max{}^*(\alpha - D(G(z))) \tag{2}$$

其中, $\max^*(\cdot) = \max(0,\cdot)$, α 是一个正实数.

由式 (1) 和式 (2) 可知, 最小化 $F_G(z)$ 就是最大化 $F_D(x,z)$ 中的第二项, 即 GAN 的优化过程就是一个极小极大化问题.

$$\min_G \max_D \big\{L(G,D) = \mathrm{E}_{x\in P_{\mathrm{data}}(x)}(\log D(x)) + \mathrm{E}_{z\in P_z(z)}(\log(\alpha - D(G(z))))\big\} \tag{3}$$

其中, $P_{\mathrm{data}}(x)$ 和 $P_z(z)$ 分别表示真实样本数据的概率分布和初始噪音数据的概率分布 (可视为一种先验分布), $\mathrm{E}(\cdot)$ 表示计算期望值.

2 E-REGAN 模型

2.1 基于能量的模型

假设一个预测问题: 一个概率模型利用观测到的输入数据 X 来预测输出数据 Y. 要想完成此任务需要学习 X 到 Y 的映射规律, 并求取能够使得 $P(Y/X)$ 最大化的映射权值参数 W. 当给定输入数据 X 和映射权值参数 W 时, 必然会得到一个对应的输出 Y, 定义 $E(W,Y,X)$ 为此时概率模型的能量. 在有监督学习框架下, 这种基于能量模型的学习原理是对于训练集中每一个输入样本 X, 输入输

出组合 (X,Y) 对应的能量当且仅当 Y 是期望输出时, 取得最小值; Y 越偏离期望输出, 能量值越大.

2.2 E-REGAN 结构及其学习过程

E-REGAN 模型由自适应深度信念网络 (ADBN) 和自适应深度自编码器 (ADAE) 构成, ADBN 和 ADAE 分别充当生成模型 G 和判别模型 D. ADBN 和 ADAE 的训练交替进行展开, 先训练生成模型 G, 优化判别模型 D, 看是否满足能量函数指标要求; 然后固定判别器 D, 继续训练生成模型 G, 使得 D 的判别准确率最小化. 当且仅当 $P_{\text{data}}=P_g$ (纳什均衡)[1] 时, 达到全局最优解. E-REGAN 的结构原理如图 1 所示.

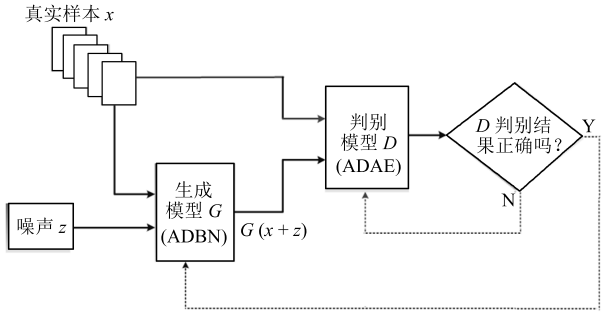

图 1　E-REGAN 结构原理图

Fig. 1　Structure and scheme of E-REGAN

2.2.1 自适应深度信念网络

ADBN 由若干个顺序堆叠的自适应受限玻尔兹曼机 (Adaptive restricted Boltzmann machine, ARBM) 和一个输出层构成, 前一个 ARBM 的输出作为后一个 ARBM 的输入. ARBM 只有两层神经元, 一层为可视层, 由显性神经元组成, 用于输入训练数据; 另一层为隐含层, 由隐性神经元组成, 用于提取训练数据的特征. ARBM 的结构如图 2 所示, 其中可视层有 m 个节点, 隐含层有 n 个节点, W 是连接权值矩阵.

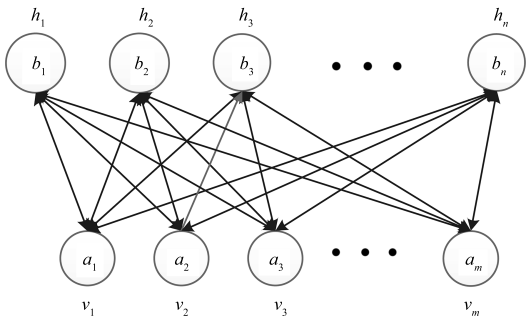

图 2　ARBM 结构图

Fig. 2　Structure of ARBM

给定模型参数 $\boldsymbol{\theta}=\{W,\boldsymbol{a},\boldsymbol{b}\}$, 那么可视层和隐含层的联合概率分布 $P(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})$ 用能量函数 $E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})$ 定义为

$$P(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})=\frac{1}{Z}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} \tag{4}$$

其中, $Z=\sum_{\boldsymbol{v},\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})}$ 是归一化因子, 模型关于 $\boldsymbol{v}$ 的边缘分布为

$$P(\boldsymbol{v};\boldsymbol{\theta})=\frac{1}{Z}\sum_{\boldsymbol{h}}\mathrm{e}^{-E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})} \tag{5}$$

对于一个伯努利 (可视层) 分布–伯努利 (隐含层) 分布的 RBM, 能量函数定义为

$$E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})=-\sum_{i=1}^{m}a_iv_i-\sum_{j=1}^{n}b_jh_j-\sum_{i=1}^{m}\sum_{j=1}^{n}v_iw_{ij}h_j \tag{6}$$

其中, w_{ij} 是 RBM 的连接权值, a_i 和 b_j 分别表示可视层节点和隐含层节点的偏置. 那么条件概率分布可表示为

$$P\left(h_j=1/\boldsymbol{v};\boldsymbol{\theta}\right)=\sigma\left(b_j+\sum_{i=1}^{m}v_iw_{ij}\right) \tag{7}$$

$$P\left(v_i=1/\boldsymbol{v};\boldsymbol{\theta}\right)=\sigma\left(a_i+\sum_{j=1}^{n}w_{ij}h_j\right) \tag{8}$$

其中, $\sigma(\cdot)$ 是一个 Sigmoid 函数.

可视层和隐含层是二值状态, 判断其二值概率取值的标准常通过设定一个阈值来实现[16], 以隐含层为例, 可表示为

$$h_j=\begin{cases}1, & \text{若 } p\left(h_j=1/\boldsymbol{v}\right)\geq\delta\\ 0, & \text{若 } p\left(h_j=1/\boldsymbol{v}\right)<\delta\end{cases} \tag{9}$$

其中, δ 为一个介于 $0.5\sim1$ 的常数.

通过计算对数似然函数 $\log P(\boldsymbol{v};\boldsymbol{\theta})$ 的梯度, 并根据 ARBM 训练过程连续两次迭代后的参数更新方向的异同[13, 15] 设计自适应学习率 η 的方法, 可以得到 ARBM 权值更新公式为

$$w_{ij}\left(\tau+1\right)=w_{ij}\left(\tau\right)+\eta\Delta w_{ij} \tag{10}$$

$$\Delta w_{ij}=E_{\text{data}}(v_ih_j)-E_{\text{model}}(v_ih_j) \tag{11}$$

$$\eta=\begin{cases}u\eta, & \left(\Delta w_{ij}\right)^{(t)}\times\left(\Delta w_{ij}\right)^{(t+1)}>0\\ v\eta, & \left(\Delta w_{ij}\right)^{(t)}\times\left(\Delta w_{ij}\right)^{(t+1)}<0\end{cases} \tag{12}$$

$$\left(\Delta w_{ij}\right)^{(t)}=v_i^{(t)}h_j^{(t)}-v_i^{(t+1)}h_j^{(t+1)} \tag{13}$$

$$(\Delta w_{ij})^{(t+1)} = v_i^{(t+1)} h_j^{(t+1)} - v_i^{(t+2)} h_j^{(t+2)} \quad (14)$$

其中, τ 和 η 分别表示 ARBM 的迭代次数和学习率, $\mathrm{E}_{\mathrm{data}}(v_i h_j)$ 和 $\mathrm{E}_{\mathrm{model}}(v_i h_j)$ 分别表示训练集中观测数据的期望和模型确定的分布上的期望, t 是吉布斯采样步数[17]. 通常情况下, $\mathrm{E}_{\mathrm{model}}(v_i h_j)$ 可由吉布斯采样近似得到[17]. ARBM 的这种训练称为自适应对比散度 (Adaptive contrastive divergence, ACD) 算法[13, 15, 18]. u 和 v 分别表示学习率增大系数和减小系数, 且 $0 < v < 1 < u$. 学习率自适应变化的原理是, 当连续两次迭代后的参数更新方向 (变化量的正负) 相同时, 学习率会加大, 相反则减小.

堆叠的 ARBM 训练结束后, 再利用 BP 算法从输出层开始由上到下对整个 ADBN 的权值进行微调 (Fine-tuning)[18].

2.2.2 自适应深度自编码器

ADAE 由若干个 ARBM 顺序堆叠而成, 是一种用于对数据进行无监督特征提取和原始数据复原的深度模型. ADAE 与 ADBN 的区别在于 ADAE 没有监督信号层[19−20], 其结构如图 3 所示. 可用 ACD 算法快速训练 ADAE 并实现判别模型对真实样本和生成样本的正确判别. 由于 ADAE 是一种无监督学习模型, 所以基于逼近误差的能量函数模型选取方法不再适用. 在此, 开创性地将 ADAE 的重构误差 (Reconstruction error, RE) 作为能量函数模型, RE 定义如下所示:

$$RE = \frac{1}{N_s \times N_d} \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} |v_{ij} - \hat{v}_{ij}| \quad (15)$$

其中, N_s 和 N_d 分别表示样本数据个数和样本数据维数; v_{ij} 和 $\hat{v}_{ij}$ 分别表示原始输入样本集的数据点值和输入样本集的重构数据点值.

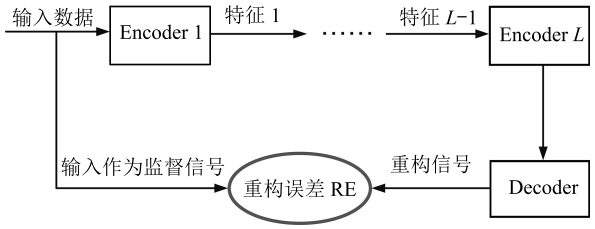

图 3 ADAE 结构原理图

Fig. 3 Structure and scheme of ADAE

用作判别模型时, ADAE 首先对真实样本数据进行特征提取, 经过快速有效地编码, 得到对真实样本数据的抽象特征表述 F 并保存下来; 然后对抽象表述进行复原处理 (解码), 得到真实样本数据的重构信息. 在此过程中, 将重构误差作为一种能量函数, 能量函数越小, 对应的判别模型学习的越充分, 理想情况下能量函数为 0. 但在实际应用中, 考虑到判别模型的训练成本问题, 经常设置一个能量函数指数 λ $(0 < \lambda < 1)$.

训练结束后, 将生成模型的生成数据作为 ADAE 的输入, 得到对生成数据的抽象特征表述 F'. 当 F 和 F' 的绝对误差 $|F - F'|$ 小于或等于能量函数阈值 γ $(0 < \gamma < \lambda)$, 需要稳定判别模型, 继续训练生成模型, 然后再以同样的原理利用判别模型检验生成模型的生成能力. 由于以能量函数阈值和能量函数指数作为评价指标的优化问题不需要求梯度校正信号, 所以只需通过判断是否满足指标要求来做更进一步的迭代优化即可. E-REGAN 的成功之处在于, 只要满足充足的训练迭代步数, ADBN 和 ADAE 将无限接近理想状态[21−23]. 在此, 自适应学习率对加速训练过程起着重要作用.

2.3 E-REGAN 稳定性分析

GAN 的基本思想源自博弈论的二人零和博弈, 由一个生成模型 G 和一个判别模型 D 构成, 通过对抗学习的方式来迭代训练, 直至逼近纳什均衡. 因此, E-REGAN 的稳定性主要取决于 G 和 D 在迭代训练过程中能否达到纳什均衡.

假设 P_G 是生成样本 $G(z)$ 的概率密度分布, 定义纳什均衡函数为

$$f(G, D) = \int_x \int_z F_D(x, z) P_{\mathrm{data}}(x) P_z(z) \,\mathrm{d}x\mathrm{d}z \quad (16)$$

$$g(G, D) = \int_x \int_z F_G(x+z) P_{\mathrm{data}}(x) P_z(z) \,\mathrm{d}x\mathrm{d}z \quad (17)$$

训练判别模型 D 来最小化 f, 训练生成模型 G 来最小化 g, 那么纳什均衡的平衡点即为 G 和 D 的最优组合对 (G^*, D^*), 且满足

$$f(G^*, D^*) \le f(G^*, D), \ \forall D \quad (18)$$

$$g(G^*, D^*) \le f(G, D^*), \ \forall G \quad (19)$$

定理 1. 如果 (G^*, D^*) 是一个纳什均衡的平衡点, 那么当 $P_{G^*} = P_{\mathrm{data}}$ 时, 满足 $f(G^*, D^*) = \alpha$.

证明. 对式 (16) 作展开处理

$$f(G^*, D) = \int_x P_{\mathrm{data}}(x) D(x) \,\mathrm{d}x + \int_z P(z) \max^* (\alpha - D(G(z))) \,\mathrm{d}z = \int_x (P_{\mathrm{data}}(x) D(x) +$$

$$P_{G^*}(x)\max{}^*(\alpha - D(G(z)))\,\mathrm{d}x \tag{20}$$

现给出式 (2) 的一般形式如下所示:

$$\psi(D) = AD + B\max{}^*(\alpha - D) \tag{21}$$

其中, $A, B \geq 0$, $0 \leq D < \infty$. 那么 $\psi(D)$ 的导数为

$$\psi'(D) = \begin{cases} A - B, & 0 \leq D < \alpha \\ A, & \alpha < D < \infty \end{cases} \tag{22}$$

由式 (22) 可知, 当 $A < B$ 时, $\psi(D)$ 在区间 $[0, \alpha)$ 上单调递减, 在区间 (α, ∞) 上单调递增. 由于 $\psi(D)$ 是连续函数, 所以 $\psi(D)$ 的最小值即为 α. 当 $A \geq B$ 时, $\psi(D)$ 在区间 $[0, \infty)$ 上单调递增, 此时 $\psi(D)$ 的最小值为 0, 即 $\psi(D)$ 在 $[0, \infty)$ 上是收敛的.

由于 $\psi(D)$ 在 $[0, \infty)$ 上是收敛的, 所以存在以下两种情况:

1) 如果 $D^*(x) > \alpha$ 且 $\hat{D}(x) = \min(D^*(x), \alpha)$, 那么不难得到 $f(G^*, \hat{D}) < f(G^*, D^*)$, 这与式 (17) 相违背.

2) 用 $\psi(D)$ 的最小值代替 $D^*(x)$ 可得 $f(G^*, \hat{D})$ 的上界[24], 即

$$f\left(G^*, \hat{D}\right) \leq \alpha \tag{23}$$

根据式 (19) 可得

$$\int_x P_{G^*}(x)\,D^*(x)\mathrm{d}x \leq \int_x P_{\mathrm{data}}(x) D^*(x)\,\mathrm{d}x \tag{24}$$

联立式 (20) 和式 (24) 可得

$$\int_x P_{G^*}(x)\,D^*(x)\mathrm{d}x + \int_x P_{G^*}(x)\,\overset{*}{\max}(\alpha - D^*(x))\mathrm{d}x \leq f(G^*, D^*) \tag{25}$$

由于 $D^*(x) \leq \alpha$, 所以有

$$\alpha \leq f(G^*, D^*) \tag{26}$$

通过式 (23) 和式 (26) 可知, $\alpha \leq f(G^*, D^*) \leq \alpha$, 即 $f(G^*, D^*) = \alpha$. □

分析 E-REGAN 模型结构可知, ADBN 具有可靠的稳定性[13, 15], 所以生成模型 G 稳定. 接下来讨论判别模型 D 的稳定性问题, 即重构误差 RE 的收敛性问题.

RE 是对 ADAE 的重构输入数据和原始输入数据计算绝对误差的方法, 计算过程中需要提前知道每一个隐含层 (ARBM) 的状态 (中间变量). 所以, 要证明 RE 的收敛性问题必须保证 ADAE 中间变量的有界性. 不失一般性, 将式 (7) 和式 (8) 中激活函数的上下渐近线设为 A_H 和 A_L, 那么对于任何一个 ARBM, s_i^0 和 s_i^t 分别表示可视层的输入状态和经过 t 次采样得到的重构状态, s_j^0 和 s_j^t 分别表示由 s_i^0 得到的隐含层状态和对模型经过 t 次采样得到的隐含层状态. 那么经过 t 次采样后有

$$s_i^0 \in [A_L, A_H] \tag{27}$$

$$s_j^0 = A_L + (A_H - A_L)\sigma\left(b_j + \sum_{i=1}^{m} s_i^0 w_{ij}\right) \tag{28}$$

$$s_i^t = A_L + (A_H - A_L)\sigma\left(a_i + \sum_{j=1}^{n} w_{ij} s_j^{t-1}\right) \tag{29}$$

$$s_j^t = A_L + (A_H - A_L)\sigma\left(b_j + \sum_{i=1}^{m} s_i^t w_{ij}\right) \tag{30}$$

由式 (27) ∼ (30) 可以看出, 在每个 ARBM 的吉布斯采样过程中, 网络输出与采样过程的中间状态有关.

定理 2. 假设 s_j^0, s_i^t 和 s_j^t 分别是 ARBM 的输入状态、中间状态和输出状态, 那么 ADAE 中间变量有界的充分必要条件是所有状态满足 s_j^0, s_i^t, $s_j^t \in [A_L, A_H]$.

证明. 由于组成 ADAE 的 ARBM 是顺序叠加的, 所以当 $s_j^0, s_i^t \in [A_L, A_H]$ 时, 由式 (27) ∼ (30) 可知, 最后一个 ARBM 的隐含层在经过 t 次吉布斯采样后, 状态范围必定为 $[A_L, A_H]$, 即 $s_j^t \in [A_L, A_H]$. 所以满足整个 ADAE 在训练过程中输入输出有界性, 网络稳定. 充分性得证.

若 ADAE 稳定, 则每个 ARBM 的可视层和隐含层状态均满足输入输出有界性. 由于 Sigmoid 函数具有单调递增性, 且随着二值神经元中取 1 的神经元个数不断增加. 可得不等式

$$s_j^t > s_i^t \tag{31}$$

$$s_i^t > s_j^0 \tag{32}$$

即中间状态满足

$$s_j^0, s_i^t, s_j^t \in [A_L, A_H] \tag{33}$$

必要性得证. □

通过以上分析及对两个定理的证明可知, E-REGAN 具有可靠的稳定性.

3　实验研究

为了验证所提 E-REGAN 模型的对抗学习能力, 分别在 MNIST 数据集和 CIFAR-10 数据集上进行测试. 为了排除无关因素对实验结果的影响,

客观反映 E-REGAN 的性能, 仿真实验的编译软件和计算机运行环境设置如下: 编译软件为 MATLAB 8.2 版本, 计算机处理器为 Intel(R) Core(TM) i7-4790, 主频为 3.6 GHz, RAM 为 8 GB.

3.1 MNIST 数据集

MNIST 数据集包含 60 000 张训练数字图像和 10 000 张测试数字图像, 每一个数字均用多种手写体显示. 每张图像均是一个 0 ∼ 9 的手写数字, 大小为 28 × 28 的像素规格. 随着模式识别和数据挖掘技术的不断发展, 很多理论方法应用到该数据库中. 该数据库被视为一种理想的、标准的测试新方法的经典对象[25−26]. 取 100 张图像数据和随机噪音的组合作为训练样本对生成模型进行训练, 首先对 E-REGAN 的生成模型 ADBN 进行训练, 然后对判别模型 ADAE 进行编码解码训练. 其中 ADBN 结构为 781-80-80-781, ADAE 结构为 781-80-80-80-80. E-REGAN 生成模型 ADBN 的固有学习参数设置如表 1 所示.

表 1 MNIST 数据集测试中 ADBN 的固有参数

Table 1 Fixed parameters of ADBN on MNIST dataset

η_0	τ	t	u	v	λ	γ
0.1	200	2	1.5	0.7	0.02	0.01

η_0 表示学习率的初始值

图 4 是生成模型 ADBN 最后一个阶段的训练均方根误差 (Root mean square error, RMSE). 从图 4 可以看出, 尽管上一阶段 ADBN 的生成样本依然被判别模型从真实样本中识别出来, 但是 ADBN 上一阶段的训练 RMSE 已经很小, 收敛到 0.005. 最后一个阶段训练 RMSE 收敛到 0.000376, 训练速度非常快, 在前 100 次迭代中已接近收敛. 图 5 是判别模型的能量函数变化曲线. 其中, SS-E-REGAN (Single sample E-REGAN) 是指生成模型的输入只有真实样本 x 的 E-REGAN, SN-E-REGAN (Single noise E-REGAN) 是指生成模型的输入只有噪音 z 的 E-REGAN. 从图 5 可以看出, 缺少噪音输入的生成模型对应的 E-REGAN 鲁棒性差, 缺少真实样本输入的生成模型对应的 E-REGAN 精度不高; 相比之下, 同时将噪音和真实样本作为生成模型输入的 E-REGAN 具有较强的鲁棒性, 对抗学习性能优越. 同时, ADAE 中的自适应学习率也加速了能量函数的收敛速度.

图 6 是 E-REGAN 生成的手写数字样本, 图 7 是 SS-E-REGAN 生成的样本图像, 图 8 是 SN-E-REGAN 生成样本图像, 图 9 是利用基于梯度校正信号且无能量函数模型的 GAN (Gradient-GAN, g-GAN) 的生成样本图像, 图 10 是 Loss-sensitive GAN (LS-GAN)[27] 的生成样本图像. 可以看出, 在经过充分的对抗学习过程后, E-REGAN 生成的样

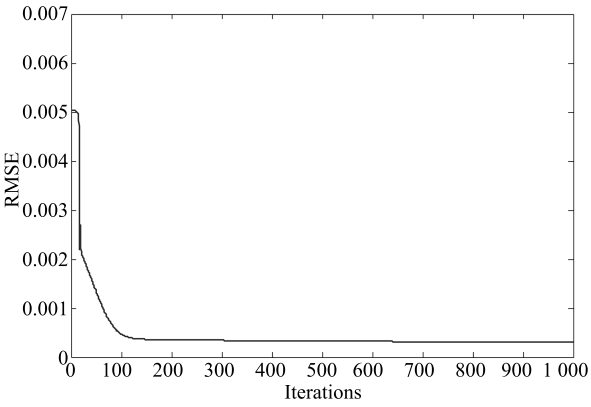

图 4 生成模型 ADBN 的训练 RMSE

Fig. 4 RMSE curve of generative model ADBN

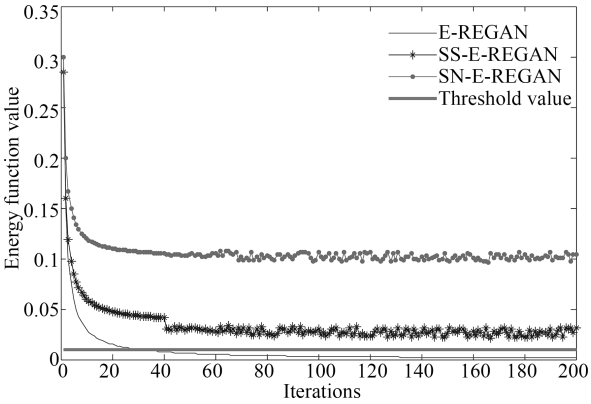

图 5 E-REGAN 的能量函数变化曲线

Fig. 5 Energy function curves of E-REGAN

图 6 E-REGAN 生成的样本图像

Fig. 6 Sample images generated by E-REGAN

图 7　SS-E-REGAN 生成的样本图像

Fig. 7　Sample images generated by SS-E-REGAN

图 8　SN-E-REGAN 生成的样本图像

Fig. 8　Sample images generated by SN-E-REGAN

图 9　g-GAN 生成的样本图像

Fig. 9　Sample images generated by g-GAN

图 10　LS-GAN 生成的样本图像

Fig. 10　Sample images generated by LS-GAN

本最清晰, 与真实的手写数字图像几乎完全一致, 至少用肉眼无法辨别出真伪.

为了更客观地反映 E-REGAN 优越的学习性能和对抗生成能力, 将 E-REGAN 与其他类似模型进行比较, 结果来自 20 次独立实验, 如表 2 所示. 其中, LS-GAN 是 Loss-sensitive GAN[27]; LR-GAN 是 Layered-recursive GAN[28]; Bayesian GAN[29] 是一种基于贝叶斯准则的 GAN. 在对比实验中, 利用分类正确率作为衡量 E-REGAN 在 MNIST 数据集上生成样本 (100 个图像) 的优劣指标 (所有生成样本图像由同一个 RBM 分类器来分类).

表 2　MNIST 数据集实验结果对比

Table 2　Result comparison on MNIST dataset

方法	能量函数 (RE)		分类正确率	平均运行
	均值	方差	(%)	时间 (s)
E-REGAN	**0.0037**	**0.0790**	**92**	**58.62**
SS-E-REGAN	0.0405	2.0618	84	56.94
SN-E-REGAN	0.1873	2.7724	82	60.31
标准 GAN	–	–	79	87.23
LS-GAN[27]	–	–	87	74.61
LR-GAN[28]	–	–	90	71.36
Bayesian GAN[29]	–	–	85	77.48

粗体表示最优值.

由表 2 可知, E-REGAN 具有最好的对抗学习能力和样本生成能力以及较好的鲁棒性能, 同时具有较快的网络学习速度.

3.2 CIFAR-10 数据集

CIFAR-10 数据集包含 10 类 60 000 个 32×32 的彩色图像, 其中有 50 000 个训练图像和 10 000 个测试图像. 该数据集分为 5 个训练块和 1 个测试块, 每个块有 10 000 个图像. 随着深度学习技术的不断发展, 基于深度神经网络的识别方法不断涌现, CIFAR-10 数据库目前已成为最具说服力的测试新方法的数据集之一[30−34]. 取 100 张图像数据和随机噪音的组合作为训练样本对生成模型进行训练, 首先对 E-REGAN 的生成模型 ADBN 进行训练, 然后对 E-REGAN 的判别模型 ADAE 进行编码解码训练. 其中 ADBN 结构: 1 024-100-100-100-1 024, ADAE 结构: 1 024-100-100-100-100-100. E-REGAN 生成模型 ADBN 的固有学习参数设置如表 3 所示.

表 3 CIFAR-10 数据集测试中 ADBN 的固有参数
Table 3 Fixed parameters of ADBN on CIFAR-10 dataset

η_0	τ	t	u	v	λ	γ
0.1	300	2	1.7	0.5	0.05	0.02

η_0 表示学习率的初始值.

图 11 是生成模型 ADBN 在最后一个训练阶段的 RMSE 变化曲线. 可以看出, 尽管上一个对抗迭代过程中 ADBN 的生成样本被判别模型从真实样本中识别出来, 但是 ADBN 的训练 RMSE 已经很小, 收敛到 0.0052. 最后一个对抗迭代过程训练 RMSE 收敛到 0.00032, 训练速度非常快, 在前 16 次迭代中已接近收敛. 图 12 是判别模型的能量函数变化曲线, 可以看出 E-REGAN 的能量函数值在前 50 迭代中已经接近收敛, 且满足能量函数阈值要求, 此时生成模型的性能达到相对理想的状态. 图 13 是 E-REGAN 生成的 CIFAR-10 数据样本, 图 14 是 Loss-sensitive GAN (LS-GAN)[27] 生成的 CIFAR-10 数据样本, 图 15 是 Layered-recursive GAN (LR-GAN)[28] 生成的 CIFAR-10 数据样本, 图 16 是 Bayesian GAN[29] 生成的 CIFAR-10 数据样本. 可以看出, 在经过充分的对抗学习过程后, E-REGAN 生成的样本图像与真实的 CIFAR-10 数据集中的图像最接近, 至少用肉眼无法辨别出真伪.

为了更为客观地反映 E-REGAN 优越的学习性能和对抗生成能力, 将 E-REGAN 与其他类似模型同时在 CIFAR-10 数据集上进行试验验证并作比较. 相应结果来自 20 次独立实验, 如表 4 所示. 在对比实验中, 利用测试误差 (Test-error) 作为衡量 E-REGAN 在 CIFAR-10 数据集上生成样本的优劣指标. 测试误差 Test-error 定义如下:

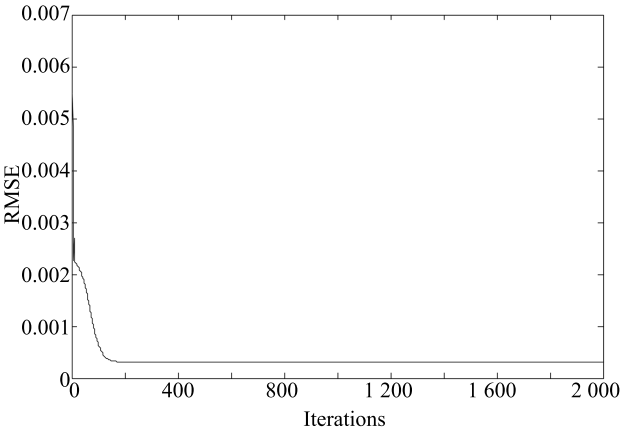

图 11 生成模型 ADBN 的训练 RMSE
Fig. 11 RMSE curve of generative model ADBN

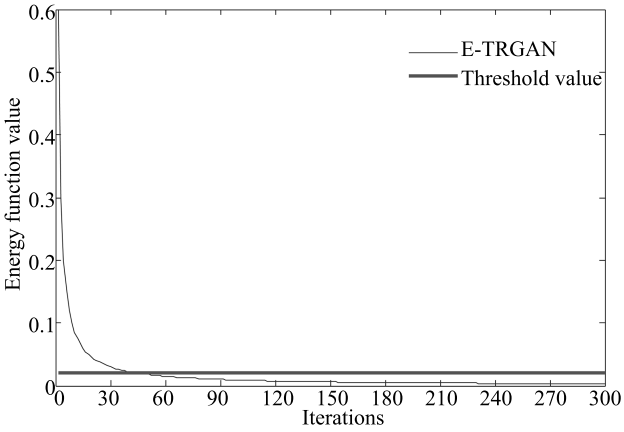

图 12 E-REGAN 的能量函数变化曲线
Fig. 12 Energy function curves of E-REGAN

图 13 E-REGAN 生成的样本图像
Fig. 13 Sample images generated by E-REGAN

表 4　CIFAR-10 数据集实验结果对比

Table 4　Result comparison on CIFAR-10 dataset

方法	能量函数		测试误差		平均运行
	均值	方差	均值	方差	时间 (s)
E-REGAN	**0.0048**	**0.0831**	**0.0160**	**0.0831**	**65.38**
SS-E-REGAN	0.0473	2.2406	0.0431	2.2406	65.75
SN-E-REGAN	0.2097	2.8119	0.0633	2.8119	67.92
标准 GAN	–	–	0.0802	1.9227	90.68
LS-GAN[27]	–	–	0.0358	0.1076	78.24
LR-GAN[28]	–	–	0.0263	0.1547	84.36
Bayesian GAN[29]	–	–	0.0386	0.2037	86.19

粗体表示最优值.

图 14　LS-GAN 生成的样本图像

Fig. 14　Sample images generated by LS-GAN

$$\text{Test-error} = \frac{1}{I}\sum_{i}^{I}|p_i - \hat{p}_i| \tag{34}$$

其中, p_i 和 $\hat{p}_i$ 分别为真实图像和生成图像经过向量化和归一化后的元素, I 为图像向量化后的维数.

从表 4 可以看出, E-REGAN 具有较好的样本生成能力和更强的鲁棒性, 同时, 具有最快的网络学习速度.

4　结束语

针对现有生成式对抗网络 GAN 生成模型学习效率低下和判别模型的学习过程易出现梯度消失的两个缺点, 本文提出了一种能量函数意义下的生成式对抗网络 (E-REGAN). 将自适应深度信念网络 (ADBN) 作为生成模型来加速生成式学习,自适应深度自编码器 (ADAE) 作为判别模型来加速判别式学习. 噪音和真实样本同时作为自适应深度信念网络的输入信号, 判别模型采用无监督学习, 自适应深度自编码器的重构误差作为能量函数. 能量函数意义下的判别基准无需梯度校正信号使得生成模型和判别模型在对抗学习过程中实现快速精确的交替优化. 与其他几种模型相比, 基于能量函数的 E-RAGAN 能够在快速的交替对抗优化过程中生成大量无限接近于真实样本的数据, 且网络鲁棒性较强. 本文仍有不足之处, 由于深度自编码器的引入, 自适应学习率

图 15　LR-GAN 生成的样本图像

Fig. 15　Sample images generated by LR-GAN

图 16　Bayesian GAN 生成的样本图像

Fig. 16　Sample images generated by Bayesian GAN

在面对深层次的特征提取时, 加速效果已不再明显. 在今后的工作中, 如何从网络结构的角度提高特征提取的效率将是优先研究方向.

References

1 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2672−2680

2 Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. arXiv preprint arXiv: 1511.05644, 2015.

3 Mao X D, Li Q, Xie H R, Lau R Y K, Wang Z, Smolley S P. Least squares generative adversarial networks. arXiv preprint ArXiv: 1611.04076, 2016.

4 Durugkar I, Gemp I, Mahadevan S. Generative multi-adversarial networks. arXiv preprint arXiv: 1611.01673, 2016.

5 Huang X, Li Y X, Poursaeed O, Hopcroft J, Belongie1 S. Stacked generative adversarial networks. arXiv preprint arXiv: 1612.04357, 2016.

6 Saito M, Matsumoto E, Saito S. Temporal generative adversarial nets with singular value clipping. In: Proceedings of the 2017 IEEE Conference on Computer Vision. Venice, Italy: ICCV, 2017. 2849−2858

7 Che T, Li Y R, Zhang R X, Hjelm R D, Li W J, Song Y Q, et al. Maximum-likelihood augmented discrete generative adversarial networks. arXiv preprint arXiv: 1702.07983, 2017.

8 Wang Kun-Feng, Gou Chao, Duan Yan-Jie, Lin Yi-Lun, Zheng Xin-Hu, Wang Fei-Yue. Generative adversarial networks: the state of the art and beyond. *Acta Automatica Sinica*, 2017, **43**(3): 321−332
(王坤峰, 苟超, 段艳杰, 林懿伦, 郑心湖, 王飞跃. 生成式对抗网络 GAN 的研究进展与展望. 自动化学报, 2017, **43**(3): 321−332)

9 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv preprint arXiv: 1701.07875, 2017.

10 Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. arXiv preprint arXiv: 1605.09782, 2016.

11 LeCun Y, Huang F. Loss functions for discriminative training of energy-based models. In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. Barbados: AIS, 2005. 206−213

12 Qiao Jun-Fei, Pan Guang-Yuan, Han Hong-Gui. Design and application of continuous deep belief network. *Acta Automatica Sinica*, 2015, **41**(12): 2138−2146
(乔俊飞, 潘广源, 韩红桂. 一种连续型深度信念网的设计与应用. 自动化学报, 2015, **41**(12): 2138−2146)

13 Qiao Jun-Fei, Wang Gong-Ming, Li Xiao-Li, Han Hong-Gui, Chai Wei. Design and application of deep belief network with adaptive learning rate. *Acta Automatica Sinica*, 2017, **43**(8): 1339−1349
(乔俊飞, 王功明, 李晓理, 韩红桂, 柴伟. 基于自适应学习率的深度信念网设计与应用. 自动化学报, 2017, **43**(8): 1339−1349)

14 Lopes N, Ribeiro B. Towards adaptive learning with improved convergence of deep belief networks on graphics processing units. *Pattern Recognition*, 2014, **47**(1): 114−127

15 Wang Gong-Ming, Li Wen-Jing, Qiao Jun-Fei. Prediction of effluent total phosphorus using PLSR-based adaptive deep belief network. *CIESC Journal*, 2017, **68**(5): 1987−1997
(王功明, 李文静, 乔俊飞. 基于 PLSR 自适应深度信念网络的出水总磷预测. 化工学报, 2017, **68**(5): 1987−1997)

16 Hinton G E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002, **14**(8): 1771−1800

17 Le Roux N, Bengio Y. Representational power of restricted boltzmann machines and deep belief networks. *Neural Computation*, 2008, **20**(6): 1631−1649

18 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

19 Alain G, Bengio Y. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 2014, **15**(1): 3563−3593

20 Chan P P K, Lin Z, Hu X, Tsang E C C, Yeung D S. Sensitivity based robust learning for stacked autoencoder against evasion attack. *Neurocomputing*, 2017, **267**: 572−580

21 Huang G B, Chen L, Siew C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks*, 2006, **17**(4): 879−892

22 Leung F H F, Lam H K, Ling S H, Tam P K S. Tuning of the structure and parameters of a neural network using an improved genetic algorithm. *IEEE Transactions on Neural networks*, 2003, **14**(1): 79−88

23 de la Rosa E, Yu W. Randomized algorithms for nonlinear system identification with deep learning modification. *Information Sciences*, 2016, **364−365**: 197−212

24 Zhao J B, Mathieu M, LeCun Y. Energy-based generative adversarial network. arXiv preprint arXiv: 1609.03126, 2016.

25 Larochelle H, Bengio Y, Louradour J, Lamblin P. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 2009, **10**: 1−40

26 Wang Y, Wang X G, Liu W Y. Unsupervised local deep feature for image recognition. *Information Sciences*, 2016, **351**: 67−75

27 Qi G J. Loss-sensitive generative adversarial networks on lipschitz densities. arXiv preprint arXiv: 1701.06264, 2017.

28 Yang J W, Kannan A, Batra D, Parikh D. LR-GAN: layered recursive generative adversarial networks for image generation. arXiv preprint arXiv: 1703.01560, 2017.

29 Saatchi Y, Wilson A. Bayesian GAN. arXiv preprint arXiv: 1705.09558, 2017.

30 Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv: 1207.0580, 2012.

31 Xu B, Wang N Y, Chen T Q, Li M. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv: 1505.00853, 2015.

32 Goroshin R, Bruna J, Tompson J, Eigen D, LeCun Y. Unsupervised learning of spatiotemporally coherent metrics. In: Proceedings of the 2015 IEEE Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4086−4093

33 Metz L, Poole B, Pfau D, Sohl-Dickstein J. Unrolled generative adversarial networks. arXiv preprint arXiv: 1611.02163, 2016.

34 Springenberg J T. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv: 1511.06390, 2015.

王功明 北京工业大学信息学部博士研究生. 主要研究方向为深度学习, 神经网络结构设计与优化.
E-mail: xiaowangqsd@163.com
(**WANG Gong-Ming** Ph. D. candidate at the Faculty of Information Technology, Beijing University of Technology. His research interest covers deep learning, structure design and optimization of neural networks.)

乔俊飞 北京工业大学信息学部教授. 主要研究方向为污水处理过程智能控制, 神经网络结构设计与优化. 本文通信作者. E-mail: junfeq@bjut.edu.cn
(**QIAO Jun-Fei** Professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers intelligent control of wastewater treatment process, structure design and optimization of neural networks. Corresponding author of this paper.)

王 磊 北京工业大学信息学部博士研究生. 主要研究方向为神经网络结构设计与优化.
E-mail: jade_wanglei@163.com
(**WANG Lei** Ph. D. candidate at the Faculty of Information Technology, Beijing University of Technology. His research interest covers structure design and optimization of neural networks.)