计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(8):120−128 [doi: 10.15888/j.cnki.csa.006987]

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

面向不平衡数据的分类算法①

蒋宗礼, 史倩月

(北京工业大学 信息学部, 北京 100124)
通讯作者: 史倩月, E-mail: qianyue_sophia@163.com

摘　要: 不平衡数据在分类时往往会偏向“多数”, 传统过采样生成的样本不能较好的表达原始数据集分布特征. 改进的变分自编码器结合数据预处理方法, 通过少数类样本训练, 使用变分自编码器的生成器生成样本, 用于以均衡训练数据集, 从而解决传统采样导致的不平衡数据引起分类过拟合问题. 我们在 UCI 四个常用的数据集上进行了实验, 结果表明该算法在保证准确率的同时提高了 *F_measure* 和 *G_mean*.

关键词: 不平衡数据; 分类; 变分自编码器; 过采样; 深度学习

引用格式: 蒋宗礼,史倩月.面向不平衡数据的分类算法.计算机系统应用,2019,28(8):120–128. http://www.c-s-a.org.cn/1003-3254/6987.html

Classification Algorithm for Imbalanced Data Set

JIANG Zong-Li, SHI Qian-Yue

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Imbalanced dataset tends to be biased towards “majority” when classifying, and samples generated by traditional over-sampling cannot well express the distribution characteristics of the original dataset. The improved variational autoencoders combine with data preprocessing method, generate samples by the generator of variational autoencoders trained by the minority class samples to balance the training data set, solve the overfitting problem caused by imbalanced dataset of traditional sampling. Experiments are carried out on four commonly used UCI datasets, the results demonstrate that the proposed method shows better classification performance in *F_measure* and *G_mean* with high accuracy.

Key words: imbalanced dataset; classification; variational autoencoders; over-sampling; deep learning

不平衡数据指在数据集中一类或多类的样本数量远远超过其他类的样本数量, 疾病诊断[1]、情感识别[2]、故障诊断[3]等常见数据都是不平衡数据.通常, 令多数类样本为负类, 少数类样本为正类. 传统分类算法使用不平衡数据时的分类结果往往偏向多数类样本, 性能较差, 提高不平衡数据的分类精度成为当前的研究热点.

针对上述问题, Chawla 等人[4]提出 SMOTE (Synthetic Minority Oversampling Technique), 该算法在少数类样本与其邻近点间通过乘以 0 到 1 的随机数线性插入样本. Han H 等人[5]提出 Borderline-SMOTE 方法, 将样本数据点分为安全点、边界点和噪音点, 在分类边界通过 SMOTE 方法生成数据. Barua 等人[6]提出 MWMOTE, 根据少数类样本距离和密度因素赋予对应信息权重, 使用聚类方法生成簇并用 SMOTE 合成少数类样本. 以上方法多数在边界生成数据, 容易造成模糊边界的问题, 并且多以欧几里得距离计算样本的分布, 难以接近真实的数据分布, 可能会产生噪声而误分.

变分自编码器 (Variational Auto Encoder, VAE)[7]是由 Kingma DP 和 Welling M 在 2014 年提出的生成

① 收稿时间: 2019-01-08; 修改时间: 2019-02-03, 2019-02-18; 采用时间: 2019-02-25; csa 在线出版时间: 2019-08-08

模型, 是深度学习方法中的一种无监督模型. 作为热门的生成模型之一, 已有许多学者对其进行研究, 文献[8]提出基于变分自编码器进行异常检测, 文献[9]使用变分自编码器提取语言特征, 文献[10]提出了一种基于变分贝叶斯自编码器的局部放电数据匹配方法.

现有的过采样预处理方法主要通过计算欧几里得距离、密度等影响因素来学习数据间的分布. 然而, 随着互联网的发展, 数据以大容量、高维度、不平衡的趋势递增, 只根据简单的衡量因素生成的样本无法全面代表大量的高维样本数据. VAE 已广泛应用于计算机视觉、图像处理、自然语言处理等领域. 但其发展时间较短, 还需突破更多的领域, 本文对其做了探索, 结合过采样来解决不平衡数据引起分类误差问题.

不平衡数据分类问题广泛影响着现实生活. 例如, 医疗诊断领域的基因表达样本, 其特征展现出高纬度的特点, 决定疾病的特征占其中的极少数, 同时, 疾病的样本数量远远小于其他的样本数量, 呈现出高度的高维不平衡性使数据分类时忽略极少数的癌变基因. 在银行信用卡欺诈检测中, 欺诈交易占数据的极少数使分类容易误分欺诈交易数据, 而对欺诈数据的误分类造成的代价往往更严重. 网络入侵、情感分类、语音识别等领域都存在明显的数据不平衡特性. 为验证本文提出模型可以有效改善原始不平衡数据对分类产生的偏斜影响, 使用 UCI 数据库四个常用的数据集进行实验, 结果证明使用变分自编码器数据预处理相比其它过采样算法提高了算法的 *F_measure* 和 *G_mean*, 具有重要的现实意义.

1 相关工作

1.1 不平衡数据处理常用方法

不平衡数据训练时多数类样本信息占主导地位, 导致分类结果偏向多数类样本, 主要有数据层面和算法层面的解决办法[11].

数据层面通过重新分布数据以减小不平衡度, 包括欠采样, 过采样和混合采样. 欠采样即去除多数类样本, 如随机欠采样、Tomek Links, 欠采样方法虽然可以使数据达到平衡状态, 但是在减少样本的同时也减少了对分类有重要影响的样本信息, 会影响分类结果. 过采样即增加少数类样本, 如随机过采样、SMOTE、Borderline-SMOTE, 但其容易导致分类过拟合, 且存在模糊边界等问题. 混合采样结合欠采样和过采样方法, 如 SMOTE+Tomek Link 算法[12], 该算法首先使用 SMOTE 生成数据, 然后利用 Tomek Link 方法清理噪声数据.

算法层面典型的解决方法有代价敏感算法[13]和集成学习方法[14]. 代价敏感方法对不同的类赋予不同的错分代价以降低少数类样本的错分率, 集成学习方法集合多个弱分类器并赋予不同的权重来提高分类性能. 算法层面的解决方法主要针对某一类数据集改进, 难以扩展.

1.2 自动编码器

自动编码器 (Auto-Encoder) 由 Rumelhart 在 1986 年提出, 其网络结构如图 1 所示. 其中, 输入层到隐藏层的映射表示为编码器, 隐藏层映射到输出层构成解码器.

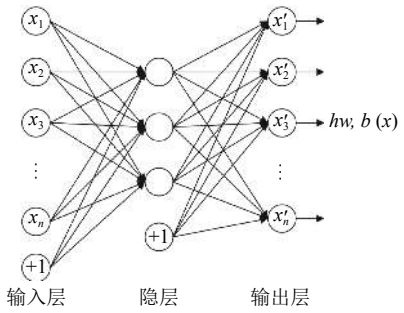


图 1 自动编码器网络结构图

编码过程:

$$z = s(wx + b) \tag{1}$$

解码过程:

$$x' = s(w'x + b') \tag{2}$$

其中, w, w'为权重矩阵, b, b'为偏置项. $s(x)$为激活函数, 通常取线性函数或者 Sigmoid 函数.

$$s(x) = \frac{1}{(1 + \exp(-x))} \tag{3}$$

自动编码器首先对输入向量 x 编码得到编码结果 z, 然后对 z 解码得到重构向量x'. 其学习过程是无监督的, 目标是使输出数据尽可能重现输入数据, 即最小化重构误差.

自动编码器是一种数据压缩算法, 编码阶段将高维数据映射成低维数据, 实现数据的特征提取, 解码阶段则与编码阶段相反, 从而实现对输入数据的复现[15].

2 改进的不平衡数据分类模型

深度学习通过对输入数据进行多层特征变换可学

习到更复杂的数据特征, 变分自编码器由神经网络学习训练样本的分布, 可以生成与训练样本近似的数据, 本文结合变分自编码器解决传统过采样技术的过拟合问题.

2.1 变分自编码器

变分自编码器基于变分下界和贝叶斯理论, 目标是最大化边缘似然函数的变分下界, 其模型图如图 2 所示.

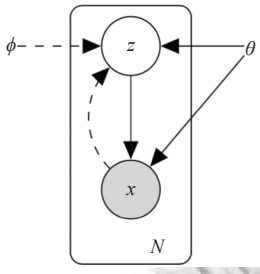


图 2 变分自编码器的图模型

其中, z 为隐变量, x 是希望生成的目标数据. 虚线表示后验分布 $p_\theta(x|z)$ 的近似分布 $q_\phi(z|x)$, 实线表示生成模型 $p_\theta(x|z)p_\theta(z)$, ϕ,θ 是在训练过程中共同学习的网络层参数[16].

变分自编码器目标函数的推导过程如下:

假设 $X=\{x^{(1)},\cdots,x^{(N)}\}$ 是独立同分布的数据集, $x^{(i)}$ 由条件分布 $p_\theta(x|z)$ 生成, z 服从先验分布 $p_\theta(z)$, 数据集 x 的对数似然函数可写为式 (4)

$$\log p_\theta(X)=\log p_\theta\left(x^{(1)},\cdots,x^{(N)}\right)=\sum_{i=1}^{N}\log p_\theta\left(x^{(i)}\right) \tag{4}$$

$x^{(i)}$ 的边缘似然函数为

$$p_\theta\left(x^{(i)}\right)=\int p_\theta\left(x^{(i)}|z\right)p_\theta(z)dz \tag{5}$$

为求解对数似然函数, 引入后验概率 $p_\theta(z|x)$ 和 $p_\theta(z|x)$ 的近似后验概率 $q_\phi(z|x)$. 使用 KL 散度 (Kullback-Leibler Divergence, KLD) 衡量 $q_\phi\left(z|x^{(i)}\right)$ 与 $p_\theta\left(z|x^{(i)}\right)$ 的距离:

$$\begin{aligned}&D_{KL}\left[q_\phi\left(z|x^{(i)}\right)\middle\|p_\theta\left(z|x^{(i)}\right)\right]\\&=E_{q_\phi(z|x)}\left[\log q_\phi\left(z|x^{(i)}\right)-\log p_\theta\left(z|x^{(i)}\right)\right]\end{aligned} \tag{6}$$

代入贝叶斯公式, 并进一步化简, 式 (6) 可得出如下公式:

$$\begin{aligned}\log p_\theta\left(x^{(i)}\right)=&D_{KL}\left[q_\phi\left(z|x^{(i)}\right)\middle\|p_\theta\left(z|x^{(i)}\right)\right]\\&+L\left(\theta,\phi;x^{(i)}\right)\end{aligned} \tag{7}$$

$$\begin{aligned}L\left(\theta,\phi;x^{(i)}\right)=&-D_{KL}\left[q_\phi\left(z|x^{(i)}\right)\|p_\theta(z)\right]\\&+E_{q_\phi(z|x^{(i)})}\left[\log p_\theta\left(x^{(i)}|z\right)\right]\end{aligned} \tag{8}$$

由于 KL 散度非负, 存在不等式 (9):

$$\log p_\theta\left(x^{(i)}\right)\geqslant L\left(\theta,\phi;x^{(i)}\right) \tag{9}$$

由此得到目标函数的变分下界:

$$L(\theta,\phi;X)=\sum_{i=1}^{N}L(\theta,\phi;x^{(i)}) \tag{10}$$

目标函数 $L\left(\theta,\phi;x^{(i)}\right)$ 的第一项 $-D_{KL}\left[q_\phi\left(z|x^{(i)}\right)\|p_\theta(z)\right]$ 作为编码器, 把输入数据 x 映射到其对应的隐变量 z, 同时作为正则化项约束 $q_\phi\left(z|x^{(i)}\right)$ 接近真实先验分布 $p_\theta(z)$. 令 $p_\theta(z)=N(z;0,I)$, $q_\phi(z|x)=N(z;\mu,\sigma^2)$, 有如下公式:

$$\begin{aligned}D_{KL}\left[q_\phi\left(z|x^{(i)}\right)\|p_\theta(z)\right]=&\frac{1}{2}\sum_{j=1}^{J}\left(1+\log\left(\left(\sigma_j^{(i)}\right)^2\right)\right.\\&\left.-\left(\mu_j^{(i)}\right)^2-\left(\sigma_j^{(i)}\right)^2\right)\end{aligned} \tag{11}$$

其中, j 为 $\sigma^{(i)}$ 的第 j 个元素, $\mu^{(i)},\sigma^{(i)}$ 由编码器计算得出.

目标函数第二项 $E_{q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)]$ 是一个对数似然期望, 在自编码器中即为重构误差损失函数. 其中 $q_\phi(z|x)=N(z;\mu,\sigma^2)$, 根据蒙特卡罗 (Monte Carlo) 方法有如下公式:

$$E_{q_\phi(z|x^{(i)})}\left[\log p_\theta\left(x^{(i)}|z\right)\right]\approx\frac{1}{L}\sum_{i=1}^{L}\log p_\theta\left(x^{(i)}|z^{(i)}\right) \tag{12}$$

式 (12) 从 $N(\mu,\sigma^2)$ 采样 z 后计算 $\log p_\theta(x^{(i)}|z^{(i)})$ 的平均值, 该过程不可微. 变分自编码器通过参数重构化解决式 (12) 无法梯度下降求解的问题, 参数重构引入了噪声随机变量 $\varepsilon\sim N(0,1)$, 令 $z=\mu^{(i)}+\sigma^{(i)}\otimes\varepsilon^{(i)}$, 将采样步骤与模型参数分离.

转换后的目标函数 $L(\theta,\phi;x^{(i)})$ 如式 (13) 所示:

$$\begin{aligned}L\left(\theta,\phi;x^{(i)}\right)\approx&\frac{1}{2}\sum_{j=1}^{J}\left(1+\log\left(\left(\sigma_j^{(i)}\right)^2\right)-\left(\mu_j^{(i)}\right)^2-\left(\sigma_j^{(i)}\right)^2\right)\\&+\frac{1}{L}\sum_{l=1}^{L}\log p_\theta\left(x^{(i)}|z^{(i,l)}\right)\end{aligned} \tag{13}$$

2.2 融合变分自编码器的不平衡数据处理

利用过采样和变分自编码器的优点, 本文提出融合变分自编码器的过采样预处理技术, 首先使用变分自编码器学习少数类样本的分布特征, 然后利用自编码器的生成器生成相似数据以降低原始数据的不平衡

度, 最后将平衡后的数据作为输入数据训练逻辑回归分类器. 整体分为以下 3 个阶段:

第一阶段: 变分自编码器学习少数类样本分布特征.

变分自编码器的结构与自编码器相似, 编码器 Q 将输入数据经过多层非线性特征转换映射为高斯分布, 解码器 P 将由高斯分布采样的隐变量重构为输入数据.

变分自编码器的结构图如图 3 所示.

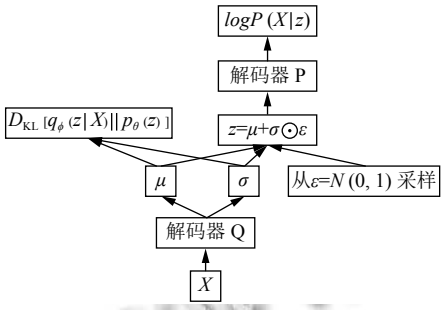


图 3 变分自编码器结构图

本文令编码器 Q 和解码器 P 为含有一个隐藏层的神经网络.

其编码器和解码器的模型如图 4 所示.

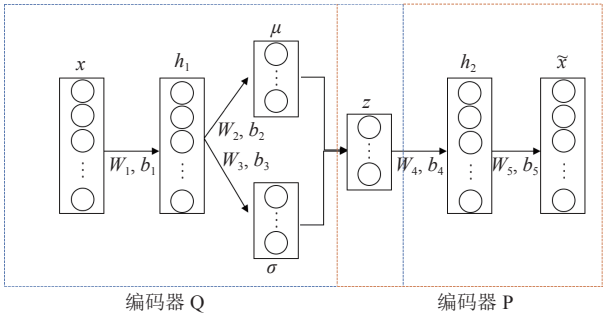


图 4 变分自编码器的编码器和解码器模型图

编码器:

$$\mu = W_2 h_1 + b_2 \tag{14}$$

$$\sigma = W_3 h_1 + b_3 \tag{15}$$

$$h_1 = \tanh(W_1 x + b_1) \tag{16}$$

解码器:

$$\log p_\theta(x|z) = \sum_{i=1}^{D_x} x_i \log y_i + (1 - x_i)\log(1 - y_i) \tag{17}$$

$$y = sigmoid(W_5 h_2 + b_5) \tag{18}$$

$$h_2 = \tanh(W_4 z + b_4) \tag{19}$$

其中, W_1, W_2, W_3, W_4, W_5为变分自编码器的连接权值矩阵, h_1, h_2, h_3, h_4, h_5为自编码器的偏置向量.

对构建的变分自编码器, 采用随机梯度下降算法最小化重构误差进而不断调整自编码器网络的参数 W, b, 第 l 层的W^l, b^l更新公式如下:

$$\Delta W^l = \Delta W^l + \frac{\partial L\left(x^{(i)}; W^l, b^l\right)}{\partial W^l} \tag{20}$$

$$\Delta b^l = \Delta b^l + \frac{\partial L\left(x^{(i)}; W^l, b^l\right)}{\partial b^l} \tag{21}$$

$$W^l = W^l - \eta \Delta W^l \tag{22}$$

$$b^l = b^l - \eta \Delta b^l \tag{23}$$

第二阶段: 采样隐变量并输入到生成器中生成指定数量的样本.

由于变分自编码器假设先验分布为高斯分布并进行了参数重构化, 因此只需要从标准正态分布中采样隐变量, 将其输入到第一阶段训练的生成器中就可以生成相似样本. 生成样本的数量对分类结果有至关重要的作用, 目前还没有统一的方法决定样本采样量, 本文通过观察不平衡率与分类结果折线图找到最优采样量.

第三阶段: 将生成数据与原始数据结合作为输入数据训练逻辑回归分类器.

整体模型结构如图 5 所示.

融合变分自编码器的不平衡数据处理训练算法如下:

算法 1. 融合变分自编码器的不平衡数据分类算法

1) 将样本数据集分为训练集和测试集.
2) 训练集的少数类样本X_{min}作为输入数据输入到变分自编码器中, 根据公式 (15)(16) 计算μ, σ.
3) 从$N(0,1)$采样ε, 根据公式$z=\mu+\sigma\cdot\varepsilon$计算隐变量 z.
4) 隐变量 z 输入到解码层, 根据公式 (19) 计算输出 y.
5) 根据公式 (13) 计算损失函数$L(W,b;X)$, 根据公式 (20)-公式 (23) 更新参数. 若算法不收敛, 重复步骤 2)-步骤 5); 若收敛, 停止训练.
6) 通过变分自编码器的生成器生成 N 个数据.
7) 生成样本数据和原训练集数据结合, 输入到逻辑回归分类器中训练.
8) 测试集作为输入数据, 输入到训练好的分类器中计算评估分数.

3 实验分析

3.1 数据集描述

本次实验所用数据集为 UCI 4 个常用的数据集[17,18], 具体描述如表 1 所示.

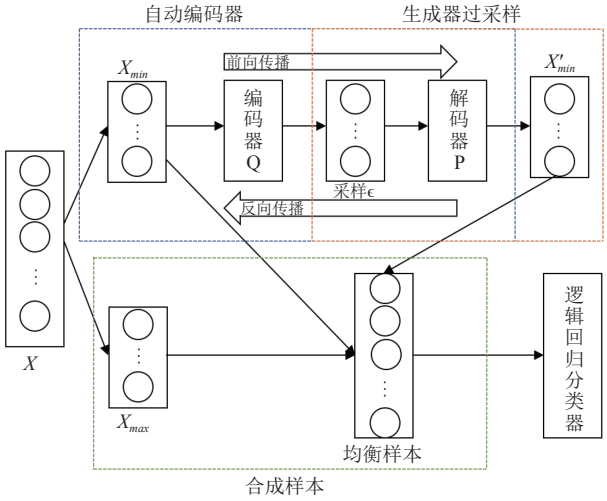


图 5 不平衡数据分类结构图

表 1 数据集信息

数据集	样本数	特征个数	少数类	多数类	不平衡率
Bank	45 211	16	5289	39 922	0.132
Credit	30 000	23	6636	23 364	0.284
Abalone7	4177	8	391	3786	0.103
Yeast1	1484	8	429	1055	0.407

其中, bank 指 UCI Bank Marketing 银行营销数据集, 该数据集通过客户信息以及对客户的电话联系判断客户是否将认购定期存款. credit 指 UIC default of credit card clients 数据集, 该数据集目的是预测用户是否会违约拖欠付款. Abalone7 是 UCI abalone 数据集, 该数据集通过物理量法预测鲍的年龄, 本文令年龄 7 岁为正类, 其它年龄为负类. yeast1 是 UCI Yeast 数据集, 其目标是预测蛋白质的细胞定位点, 本文令类 AUC 为正类, 其它类为负类.

3.2 评价指标

传统方法使用准确率 (正确分类样本个数/总样本个数) 评估分类结果, 该评估指标可以准确评价平衡数据集的分类, 但是衡量不平衡数据集时忽略了少数类样本的分类精度[19]. 例如, 样本数据集中少数类样本占比为 10%, 多数类样本占比为 90%, 若把所有样本分类为多数类样本, 准确率为 90%, 但是少数类样本分类精度为 0.

根据混淆矩阵 (如表 2 所示), 有以下评价指标:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{24}$$

$$precision = \frac{TP}{TP+FP} \tag{25}$$

$$recall = \frac{TP}{TP+FN} \tag{26}$$

$$F\_Measure = \frac{2 \bullet recall \bullet precision}{recall + precision} \tag{27}$$

$$G\_mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{28}$$

查准率 (*precision*) 表示被正确分类的正类样本占所有预测为正类样本的数据比例. 查全率 (*recall*) 表示被正确分类的正类样本占所有正类样本的比例. *F_Measure* 综合考虑了 *precision* 和 *recall*, 是两个衡量指标的调和平均, 可以评价分类器的整体性能, 当两者都较大时, *F_Measure* 才会较大. *G_mean* 是少数类样本分类精度和多数类样本分类精度的几何平均值, 可评价分类器对于每一类的分类性能. *F_Measure* 和 *G_mean* 更适合评价不平衡数据的分类性能, 本文选择 *F_Measure*, *G_mean* 和准确率 *accuracy* 作为评价指标.

表 2 混淆矩阵

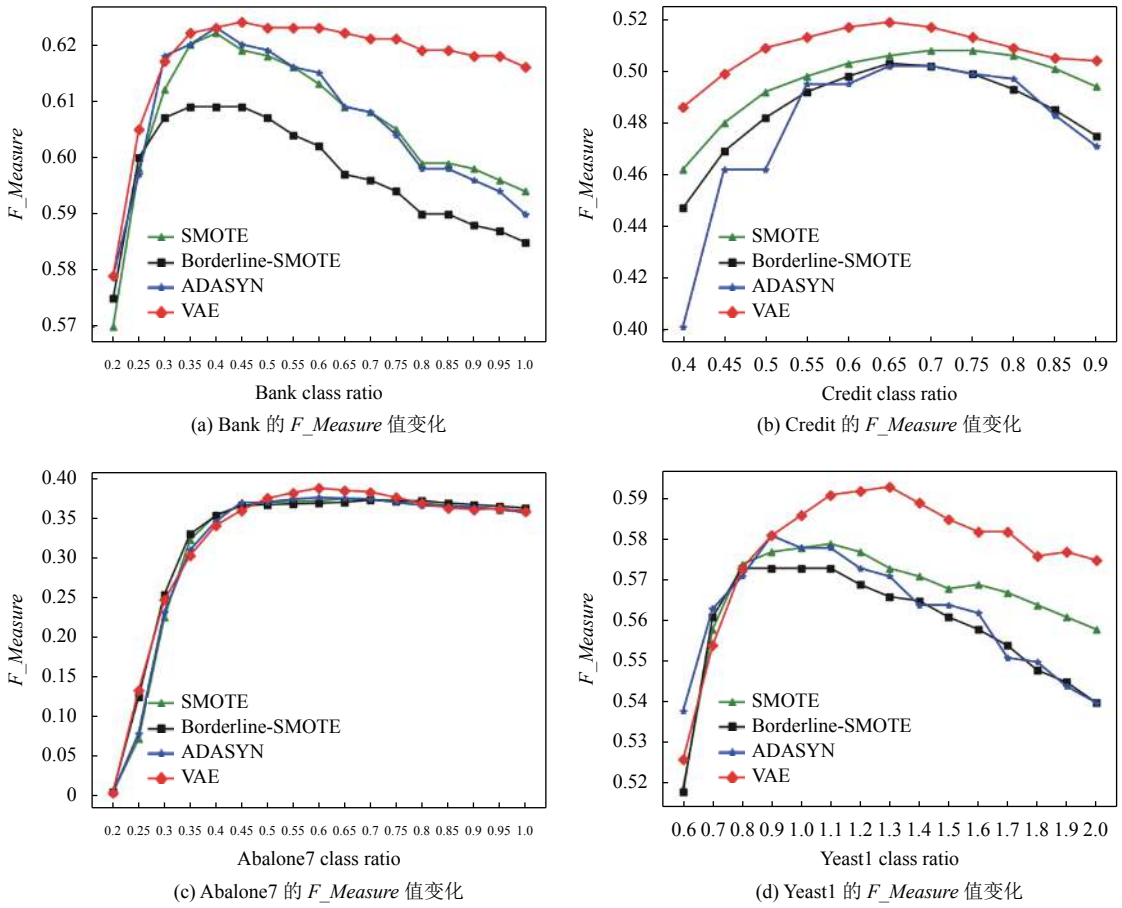
	预测正例	预测反例
真实正例	真正例 (TP)	伪反例 (FN)
真实反例	伪正例 (FP)	真反例 (TN)

3.3 数据采样量对分类的影响

为验证 VAE 数据预处理解决不平衡数据问题的有效性, 实验对比了不进行数据预处理以及使用 SMOTE、Borderline-SMOTE、ADASYN 过采样方法后的分类性能. 实验均采用十次五折交叉验证的平均值作为实验结果.

不平衡率对分类器的性能起至关重要的作用, 图 6 展示了不断增加少数类样本后不平衡率对应的 *F_Measure* 值, 图 7 展示了不断增加少数类样本后不平衡率对应的 *G_mean* 值.

通过图 6 和图 7 看出, 不同数据集对应的最优采样率不同. 经过对比发现 Bank 采样后不平衡率到达 0.4 最优, Credit 采样后不平衡率到达 0.65 最优, Abalone7 采样后不平衡率到达 0.6 最优, Yeast1 采样后不平衡率到达 1.1 最优. 虽然 Bank 数据集和 Abalone 数据集使用 VAE 预处理的 *G_mean* 值略低于其它预处理方法的 *G_mean* 值, 但是经过 VAE 预处理的 *F_Measure* 值几乎都高于其他预处理方法的 *F_Measure* 值, 可以看出 VAE 预处理具有更好的分类性能.



(a) Bank 的 *F_Measure* 值变化　(b) Credit 的 *F_Measure* 值变化

(c) Abalone7 的 *F_Measure* 值变化　(d) Yeast1 的 *F_Measure* 值变化

图 6　4 个数据集采样后不平衡率对应的 *F_Measure* 值

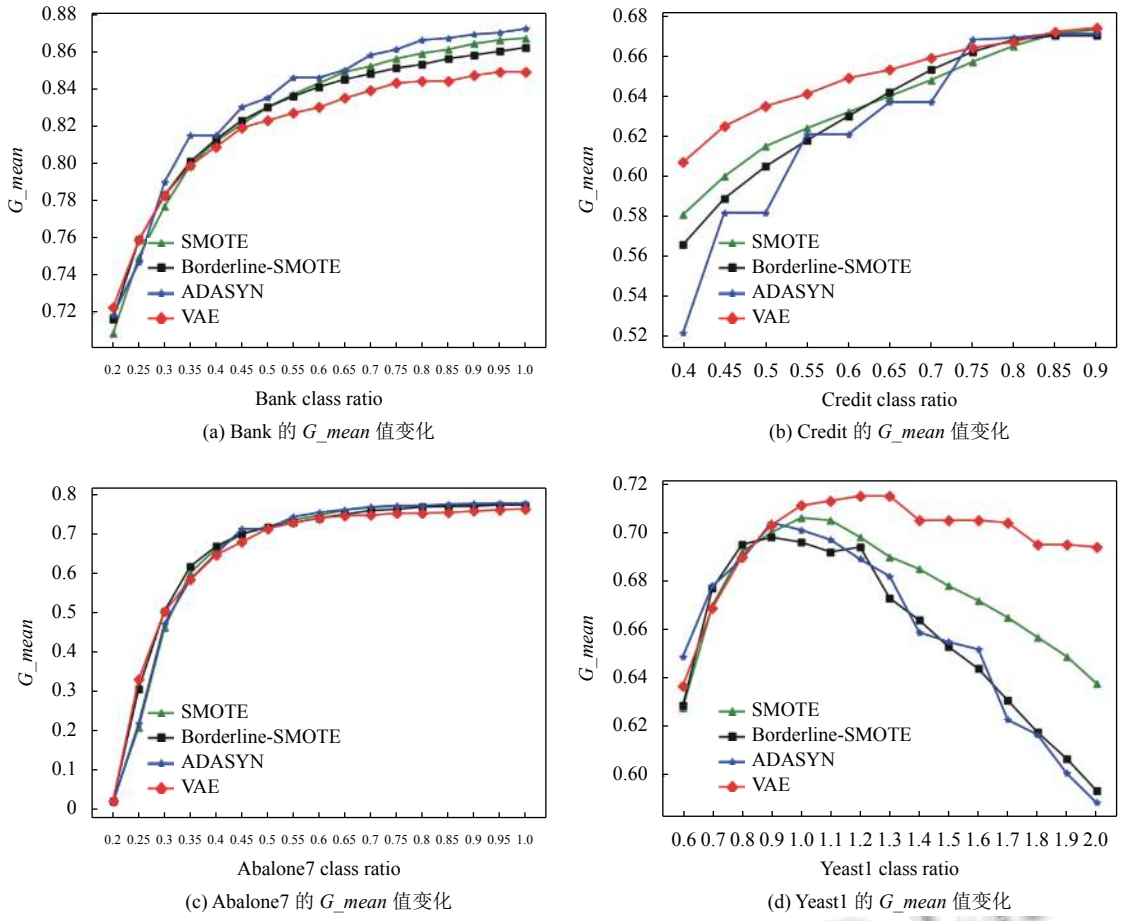
3.4　实验结果

表 3 分别列出了 4 个数据集使用不同预处理方法在最优采样率下的 *F_measure*, *G_mean* 和准确率. LR 表示直接对不平衡数据使用逻辑回归分类器分类.

由表 3 可知, 数据集使用变分自编码器数据预处理对比其它预处理方法有明显提高. Bank 数据集使用 VAE 预处理对比直接进行分类提高了 25.6%, 对比使用 SMOTE 方法提高了 0.16%, 对比使用 Borderline-SMOTE 提高了 2.30%, 对比使用 ADASYN 提高了 0.00%.credit 数据集使用 VAE 预处理对比直接进行分类提高了 49.1%, 对比使用 SMOTE 方法提高了 2.57%, 对比使用 Borderline-SMOTE 提高了 3.18%, 对比使用 ADASYN 提高了 3.39%. Abalone7 数据集使用 VAE 预处理对比直接进行分类提高了 387 倍, 对比使用 SMOTE 方法提高了 4.30%, 对比使用 Borderline-SMOTE 提高了 4.86%, 对比使用 ADASYN 提高了 3.19%. Yeast1 数据集使用 VAE 预处理对比直接进行分类提高了 76.9%, 对比使用 SMOTE 方法提高了 2.07%, 对比使用 Borderline-SMOTE 提高了 3.14%, 对比使用 ADASYN 提高了 2.25%. 对于 4 个数据集 *F_Measure* 平均值, 变分自编码器使用 VAE 预处理对比不进行数据预处理提高 80.3%, 对比 SMOTE 方法提高 2.12%, 对比 Borderline-SMOTE 方法提高 3.31%, 对比 ADASYN 方法提高 2.12%.

数据集使用变分自编码器数据预处理整体提高了 *G_mean* 值. 对于 Bank 数据集和 Abalone7 数据集, 本文方法略低于其它数据预处理方法, 但是在 Credit 数据集上实验, 本文方法对比不进行预处理提高了 38.1%, 对比 SMOTE 方法提高了 2.03%, 对比 Borderline-SMOTE 提高了 1.71%, 对比 ADASYN 方法提高了 2.5%. 在 Yeast1 数据集上实验, 本文方法对比不进行数据预处理提高了 51.7%, 对比 SMOTE 方法提高了 1.13%, 对比 Borderline-SMOTE 提高了 3.03%, 对比 ADASYN 方法提高了 2.30%. 对于 4 个数据集 *F_Measure* 平均

值, 变分自编码器使用 VAE 预处理对比不进行数据预处理提高了 86.9%, 对比 SMOTE 方法提高了 0.41%, 对比 Borderline-SMOTE 方法提高了 0.97%, 对比 ADASYN 方法提高了 0.41%.



(a) Bank 的 *G_mean* 值变化 (b) Credit 的 *G_mean* 值变化

(c) Abalone7 的 *G_mean* 值变化 (d) Yeast1 的 *G_mean* 值变化

图 7 4 个数据集采样后不平衡率对应的 *G_mean* 值

表 3 数据集使用不同方法的 *F_Measure*, *G_mean*, *accuracy*

	方法	Bank	Credit	Abalone7	Yeast1	average
F_Measure	LR	0.496	0.348	0.001	0.334	0.294
	SMOTE	0.622	0.506	0.372	0.579	0.519
	Borderline-SMOTE	0.609	0.503	0.370	0.573	0.513
	ADASYN	0.623	0.502	0.376	0.578	0.519
	VAE	0.623	0.519	0.388	0.591	0.530
G_mean	LR	0.619	0.473	0.001	0.470	0.390
	SMOTE	0.812	0.640	0.748	0.705	0.726
	Borderline-SMOTE	0.813	0.642	0.741	0.692	0.722
	ADASYN	0.815	0.637	0.755	0.697	0.726
	VAE	0.809	0.653	0.741	0.713	0.729
accuracy	LR	0.910	0.809	0.906	0.742	0.841
	SMOTE	0.902	0.805	0.372	0.691	0.692
	Borderline-SMOTE	0.895	0.798	0.369	0.663	0.681
	ADASYN	0.903	0.803	0.771	0.667	0.786
	VAE	0.905	0.800	0.792	0.723	0.805

数据集使用变分自编码器数据预处理对比其它预处理方法整体上提高了分类准确率. Bank 数据集使用 VAE 预处理对比直接进行分类降低了 0.55%, 对比使用 SMOTE 方法提高了 0.33%, 对比使用 Borderline-

SMOTE 提高了 1.12%, 对比使用 ADASYN 提高了 0.22%. Credit 数据集使用 VAE 预处理对比使用 SMOTE 方法降低了 1.11%, 对比使用 Borderline-SMOTE 降低了 0.62%, 对比使用 ADASYN 降低了 0.37%. Abalone7 数据集使用 VAE 预处理对比使用 SMOTE 方法提高了 1.13 倍, 对比使用 Borderline-SMOTE 提高了 1.15 倍, 对比使用 ADASYN 提高了 17.9%. Yeast1 数据集使用 VAE 预处理对比使用 SMOTE 方法提高了 4.63%, 对比使用 Borderline-SMOTE 提高了 9.04%, 对比使用 ADASYN 提高了 8.40%. 对于 4 个数据的 *accuracy* 平均值, 变分自编码器预处理对比 SMOTE 方法提高 11.3%, 对比 Borderline-SMOTE 方法提高 18.2%, 对比 ADASYN 方法提高 2.42%.

3.5 实验结果分析

由表 3 可以看出, 直接使用逻辑回归分类器对不平衡数据分类的 *F_Measure* 和 *G_mean* 值很低, 分析其原因, 是由于逻辑回归算法平等的看待每一类样本, 而少数类样本提供给分类器的有效信息极少, 分类器将大部分样本预测为多数类样本以保证较高的准确率, 导致少数类样本的准确率严重降低. 为解决此问题, 本文在样本输入到分类器之前进行过采样处理均衡正负类样本提高少数类样本精度.

相较于 SMOTE、Borderline-SMOTE、ADASYN 算法, 本文提出的算法其准确率、*F_Measure* 和 *G_mean* 更高. 可见, 相比仅通过欧几里得距离及其改进算法衡量数据间的分布情况, 本文通过含有多个神经元的隐含层线性学习并使用激活函数非线性变换, 学习样本的不同特征分布, 以此学习到的分布更接近真实数据分布. 另外, 借助变分自编码器的思想, 使采样的数据通过解码器生成的少数类样本更符合原始数据的特征.

变分自编码器充分考虑了少数类样本不同层次的特征, 可以生成更广泛的少数类样本, 从而有效提高了分类器的泛化能力. 因此, 本文提出方法训练的分类器预测测试样本时, 准确率、*F_Measure* 和 *G_mean* 都较高.

综上所述, 使用变分自编码器均衡不平衡数据集改善了原始数据集中多数类样本占主导作用使少数类样本准确率降低的问题, 其生成的样本增加了分类时少数类样本的有效信息并提高了少数类样本的分类识别率, 具有更高的分类精确度.同时, 变分自编码器通过神经网络多次非线性特征转换学习到的数据分布特征更接近真实数据, 改善了传统过采样技术产生无效的“人造样本”影响少数类样本分布导致模糊正负类边界的问题. 融入变分自编码器的过采样技术在提高少数类样本精度的同时兼顾了多数类样本准确率.

4 结论

本文结合变分自编码器和过采样技术解决数据不平衡导致传统分类器分类性能较差的问题, 该方法通过变分自编码器学习少数类样本的分布, 使用其生成器生成相似的数据以均衡数据集. 实验结果表明变分自编码器生成的样本更接近真实数据, 融合变分自编码器的数据预处理技术保证了较高准确率的同时提高了少数类样本的精确度, 改善了不平衡数据的分类偏斜问题和传统过采样的过拟合问题.

参考文献

1 Zeng M, Zou BJ, Wei FR, *et al*. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. Proceedings of 2016 IEEE International Conference of Online Analysis and Computing Science. Chongqing, China. 2016. 225–228.

2 Sarakit P, Theeramunkong T, Haruechaiyasak C. Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. Proceedings of the 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications. Chonburi, Thailand. 2015. 1–2.

3 Wang AN, Liu LM, Jin X, *et al*. Adapting TSVM for fault diagnosis with imbalanced class data. Proceedings of 2016 Chinese Control and Decision Conference. Yinchuan, China. 2016. 2919–2923.

4 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.

5 Han H, Wang WY, Mao BH. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Proceedings of 2005 International Conference on Intelligent Computing. Hefei, China. 2005. 878–887.

6 Barua S, Islam M, Yao X, *et al*. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405–425. [doi: 10.1109/TKDE.2012.232]

7 Kingma DP, Welling M. Auto-encoding variational Bayes.

Proceedings of 2014 International Conference on Learning Representations. Banff, Canada. 2014. arXiv:1312.6114.
8 An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf, 2015.
9 Tan S, Sim KC. Learning utterance-level normalisation using Variational autoencoders for robust automatic speech recognition. Proceedings of 2016 IEEE Spoken Language Technology Workshop. San Diego, CA, USA. 2016. 43–49.
10 宋辉, 代杰杰, 张卫东, 等. 基于变分贝叶斯自编码器的局部放电数据匹配方法. 中国电机工程学报, 2018, 38(19): 5869–5877.
11 叶志飞, 文益民, 吕宝粮. 不平衡分类问题研究综述. 智能系统学报, 2009, 4(2): 148–156.
12 Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter-Special Issue on Learning from Imbalanced Datasets, 2004, 6(1): 20–29.
13 Castro CL, Braga AP. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(6): 888–899. [doi: 10.1109/TNNLS.2013.2246188]
14 Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models. Proceedings of 2009 IEEE Symposium on Computational Intelligence and Data Mining. Nashville, TN, USA. 2009. 324–331.
15 贾文娟, 张煜东. 自编码器理论与方法综述. 计算机系统应用, 2018, 27(5): 1–9.
16 Doersch C. Tutorial on variational autoencoders. Stat, 2016, 1050: 13.
17 Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 2014, 62: 22–31. [doi: 10.1016/j.dss.2014.03.001]
18 Yeh IC, Lien CH. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 2009, 36(2): 2473–2480. [doi: 10.1016/j.eswa.2007.12.020]
19 Prusty MR, Jayanthi T, Velusamy K. Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors. Progress in Nuclear Energy, 2017, 100: 355–364. [doi: 10.1016/j.pnucene.2017.07.015]