控制与决策
Control and Decision

基于条件生成对抗网络的不平衡学习研究

赵海霞， 石洪波， 武建， 陈鑫

引用本文:
赵海霞, 石洪波, 武建, 等. 基于条件生成对抗网络的不平衡学习研究[J]. 控制与决策, 2021, 36(3): 619−628.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0522

基于条件生成对抗网络的不平衡学习研究

赵海霞[1], 石洪波[2†], 武 建[3,4], 陈 鑫[2]

(1. 山西财经大学 统计学院，太原 030006；2. 山西财经大学 信息学院，太原 030006;
3. 山西财经大学 应用数学学院，太原 030006；4. 太原理工大学 信息与计算机学院，太原 030600)

摘 要: 对于不平衡数据的分类,不平衡率并不是影响分类效果的唯一因素,类别间的重叠、正类样本的分离以及噪音样本的存在等均会对分类效果造成影响. 针对具有类别重叠的不平衡数据集,提出基于CGAN模型的重抽样方法(RECGAN). 该方法结合负类样本的欠抽样和正类样本的过抽样,既能够提高重叠区域正类样本的识别度,又可以克服以往均从样本点的局部邻域出发合成样本的缺陷. 实验结果表明,无论是从AUC和F_1的取值看,还是从数据集上的平均排序看,RECGAN方法均具有明显的优势.

关键词: 不平衡学习；类别重叠；重抽样方法；条件生成对抗网络

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0522 **开放科学(资源服务)标识码(OSID):**

引用格式: 赵海霞,石洪波,武建,等. 基于条件生成对抗网络的不平衡学习研究[J]. 控制与决策, 2021, 36(3): 619-628.

Research on imbalanced learning based on conditional generative adversarial networks

ZHAO Hai-xia[1], *SHI Hong-bo*[2†], *WU Jian*[3,4], *CHEN Xin*[2]

(1. School of Statistic，Shanxi University of Finance and Economics，Taiyuan 030006，China；2. College of Information，Shanxi University of Finance and Economics，Taiyuan 030006，China；3. Department of Applied Mathematics，Shanxi University of Finance and Economics，Taiyuan 030006，China；4. College of Information and Computer，Taiyuan University of Technology，Taiyuan 030600，China)

Abstract: For the classification of imbalanced data, the imbalance ratio is not the only factor affecting the classification effect. The class overlapping, the separation of positive samples and the noise samples will all have impact on the classification effect. For the imbalanced data with class overlapping, a re-sampling method with the conditional generotive adversarial networks(CGAN) model(RECGAN) is proposed. It not only improves the recognition of positive samples in overlapping regions, but also overcomes the defect of previous sample synthesis based on the local neighborhood of samples. The experimental results show that the RECGAN method has obvious advantages in terms of the values of AUC and F_1 and the average ordering on the dataset.

Keywords: imbalanced learning；class overlapping；re-sampling approach；conditional generative adversarial networks

0 引 言

在现实问题中经常会遇到很多类别不平衡的数据集,即数据集中某一类的样本量要远大于其他类别的样本量. 例如在医疗诊断中,需要特别关注的患者人数比普通患者要少之又少;在保险行业的欺诈检测中,相较于正常客户,具有欺诈行为的客户数量较少;类似的还有客户流失检测、银行风险分析、软件缺陷检测等领域. 在类别不平衡的数据集中(以二分类为例),通常将样本量多的类别称为负类,样本量少的类别称为正类,两类别间的样本数量之比称为不平衡率(imbalance ratio, IR),在很多实际问题的应用中,IR甚至可以达到1 000以上[1-2]. 传统的机器学习算法在应用于不平衡数据集时,其分类效果往往并不理想,通常表现为偏向负类样本,对正类样本难以识别[3-4]. 面对不平衡数据,关于如何克服分类器的偏误问题称为不平衡学习[5]. 多数情况下,正类样本虽然样例较少,但是能够提供更多重要的信息,对正类样本的误分代价会远高于负类样本. 例如在金融欺诈

收稿日期: 2019-04-24；**修回日期:** 2019-09-02.
基金项目: 国家社会科学基金项目(17BTJ010)；山西省自然科学基金项目(2014011022-2).
责任编委: 阳春华.
†通讯作者. E-mail: Shihb@sxufe.edu.cn.

检测、网络入侵检测、生物医学研究、电信管理与生物信息中稀有粒子的检测等领域,数量较少的正类样本都是数据挖掘应重点关注的研究对象.

目前,关于类别不平衡数据学习的研究策略大致可以分为3个方面:一是从数据层面进行研究,该类方法主要通过重抽样技术改变训练集的分布,降低类别间的IR,使得训练集趋于平衡,然后运用传统的分类算法进行研究;二是基于算法层面的研究,该类方法通过改进分类算法以降低偏向负类的误差,提高对正类的识别率,其中最为流行的便是代价敏感分类算法,通过给正类样本设置一个较高的错分代价因子以达到提高分类效果的目的,虽然在实际问题中合适的代价因子矩阵较难确定,但关于这方面的研究已有较多优秀成果[6-9];三是关于数据层面和算法层面的结合,该类方法主要将前面两种策略进行整合,同时可以减少各自的弱点,提取其优点,以提高算法的分类效果[10].

本文主要关注应用较为普遍的第1种策略,即从数据层面对不平衡学习问题进行研究. 从数据层面利用重抽样技术降低数据集的IR,最普遍的方法是随机过抽样和随机欠抽样. 随机过抽样方法通过对正类样本进行简单随机的重复抽取,而随机欠抽样方法则通过对负类样本进行随机地删减,使得数据集逐渐趋于平衡. 然而,简单的随机重抽样技术在不平衡分类中存在明显缺陷,简单随机过抽样方法容易产生过拟合,简单随机欠抽样方法又很容易删减掉负类中一些重要的样本.

为避免简单随机重抽样方法的缺陷,SMOTE (synthetic minority over-sampling technique)方法被提出且被广泛应用,该方法将正类中的每个样本作为一个种子,寻找其同类的k-近邻样本,然后按照一定的比例在其与种子样本之间插入生成样本[11]. 显然,SMOTE方法对于正类中的每个样本都进行样本合成的做法,会使得那些不安全的样本更加难以被正确分类;另外其在合成样本的过程中并没有考虑到邻域中的负类样本. 为克服这样的问题,在SMOTE方法的基础上进行了一系列的改进:如运行SMOTE之后,结合ENNR方法对样本进行过滤,且通过实验验证其分类效果较好[12];Borderline-SMOTE方法只是将正类样本中边界上的样本作为种子进行样本的合成,可以有效地避免噪音样本的过抽样[13];ADASYN方法分析了每个种子样本k-邻域中负类样本的数量情况,结合数据不平衡率合成样本,在一定程度上避免了噪音样本的生成[14];Safe Level SMOTE、LN-SMOTE不仅分析种子样本局部子区域的分布,而且对其所选的邻域进行分析,通过选取合适的权重进行样本合成[15];Napierala等[16]对正类中每个样本的邻域进行了分析,并将其分为安全样本、边界样本、稀有样本和异常值4种情况,这不仅对正类样本的数据合成提供了指导,而且对于不平衡学习的研究也提供了一种新的思路,即从数据集的结构特征出发,分析影响分类效果的因素.

以上所述重抽样方法均是从样本点的局部邻域出发,并没有考虑到数据集的整体分布情况,如果能从数据的分布直接进行抽样,对训练集进行平衡处理,将是数据层面理想的不平衡学习方法. 生成对抗网络(generative adversarial networks, GAN)由Goodfellow等[17]于2014年提出,其采用内部对抗机制对网络进行训练,由于GAN模型可以学习到实际的数据分布,且效果较好,使得GAN模型不仅在学术界甚至工业界都受到了广泛关注[18-21]. 条件生成对抗网络[22](conditional generative adversa-rial nets, CGAN)是在GAN的基础上增加一个外部条件信息,去指导网络的训练及数据的生成. Douzas等[23]首次运用CGAN模型解决不平衡学习问题,并验证用模型中的生成网络对正类样本进行过抽样时,分类效果有一定的提升,但是仍存在明显的不足. 文中只考虑了IR对不平衡学习的影响,运用生成网络对正类样本进行过抽样,降低数据集的不平衡率. 然而研究表明,在不平衡数据的分类中,IR并不是影响分类效果的主要因素,数据集的结构特征例如:类别间的重叠、正类样本的分离以及噪音样本的存在等,均是影响分类效果的关键因素[16,24],且类别间重叠的程度越高,对分类效果的影响越大[25-26]. 所以对于类别间存在重叠的不平衡数据集,如果只是一味地对正类样本进行过抽样,则必然会导致边界样本的增加,对分类效果造成影响.

本文基于以上考虑,针对存在类别重叠的二分类不平衡数据集,提出一种基于CGAN模型的重抽样方法(re-sampling method based on CGAN, RECGAN). 首先,分析数据集的结构特征,移除训练集中负类的噪音样本,并对重叠区域的负类样本进行欠抽样,降低重叠区域负类样本的比重;然后,引入数据集的标签变量作为CGAN模型的条件变量,搭建并训练CGAN模型,并运用模型中的生成网络对训练集中的正类样本进行过抽样,使训练集趋于平衡;最后,为验证所提出RECGAN方法在不平衡学习中的优势,基于公开的不平衡数据集,利用4种分类器和两种评价

准则,与其他几种重抽样方法的性能进行了实验比较. 实验结果表明,无论是从AUC和F_1的取值看,还是从数据集上的平均排序看,RECGAN方法均具有明显的优势.

1　生成对抗网络

1.1　GAN模型

GAN模型是近年深度学习领域应用较广的一种生成式模型,其设计的思想源于博弈论中二人零和博弈的思想,模型中包含两个网络:生成网络G和判别网络D,这两个网络之间便是一种对抗博弈的关系. 生成网络基于一组随机噪声生成样本,其目的在于生成与真实样本相似的样本,混淆判别网络使其无法判别;判别网络的目的在于判别出输入是生成样本还是来自于真实数据. 因此,GAN模型其实就是生成网络和判别网络不断优化的一个训练过程,当生成网络成功学习到真实数据的样本分布时便达到了最终的平衡点. GAN模型的整体框架如图1所示.

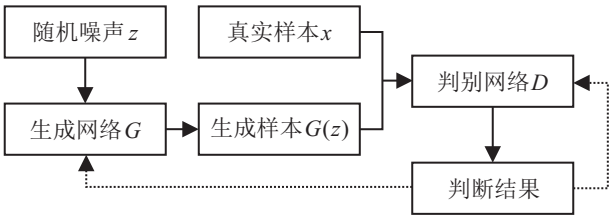


图1　GAN模型框架

记随机噪声z从先验分布$P_z(z)$中采样,x从真实数据分布$P_{\text{data}}(x)$中采样. 生成网络G和判别网络D均为多层感知机,G的输入为任意维度的随机噪声z,输出为尽量与真实数据相似的生成样本$G(z)$. 真实样本x和生成样本$G(z)$同时进入判别网络D中,D为二分类判别器,输出为样本来自于真实数据的概率. GAN模型的目标函数为

$$\min_G \max_D V(D,G) = \boldsymbol{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] + \boldsymbol{E}_{z\sim P_z(z)}[\log(1-D(G(z)))]. \tag{1}$$

训练过程采用生成网络G和判别网络D交替优化的方法进行,通常先固定G,优化D,最大化其判别正确率;然后固定D,优化G,尽量混淆判别网络,最小化其判别正确率. 在文献[17]中已证明,当$p_g = P_{\text{data}}$,即判别器的输出为1/2时,达到全局最优解.

1.2　CGAN模型

相较于其他生成式模型,GAN模型中生成网络和判别网络对抗训练的方式不再要求一个假设的数据分布,而是基于分布直接进行采样,在理论上可以完全逼近真实数据. 然而,简单GAN模型也存在一个很大的缺点:模型太过自由,导致模型的训练过程很难达到稳定. 为此,Mirza等[22]在简单GAN的基础上加上外部信息条件扩展为CGAN模型,在生成网络和判别网络中均引入一个条件变量y,达到指导数据生成的目的. 如果引入的条件变量y是数据的类标签,则CGAN模型可以看作是将无监督的GAN模型扩展到有监督的学习模型. CGAN模型的训练过程与GAN模型完全相似,其目标函数如下:

$$\min_G \max_D V(D|G) = \boldsymbol{E}_{x\sim P_{\text{data}}(x)}[\log D(x|y)] + \boldsymbol{E}_{z\sim P_z(z)}[\log(1-D(G(z|y)))]. \tag{2}$$

2　基于CGAN模型的重抽样方法

本文对于不平衡学习问题的研究,通过分析数据集的结构特征,针对具有类别重叠的不平衡数据集,基于CGAN模型提出一种欠抽样和过抽样相结合的重抽样方法. 该部分内容首先介绍不平衡数据集的结构特征分析,然后介绍基于CGAN模型的重抽样方法(RECGAN).

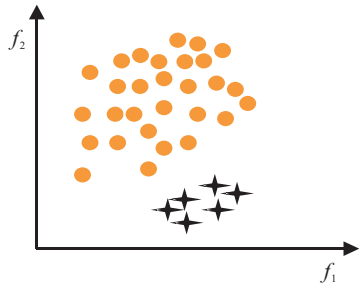
2.1　不平衡数据集的结构特征分析

设不平衡数据集为$D_0 = \{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$. 其中:$N$为数据集的样本容量,$x_i = (x_{i1}, x_{i2},\ldots,x_{id_x})$表示一个维数为$d_x$的样本点,$y_i \in \{-1,1\}$为数据集的标签变量.

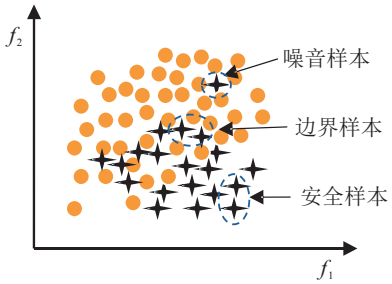
在不平衡学习问题的研究中,类别间的不平衡率并不是导致学习困难的唯一因素,只要各类分布可以被该类样本数据完全表达,且类别间不存在重叠,即使不平衡率非常高,使用传统的分类器依然可以得到很好的分类效果(见图2(a)). 那么在不平衡数据分类中导致分类器效果下降的关键因素,主要是在正类中存在一些较难分类的样本,包括位于类别边界附近和重叠区域的边界样本,以及噪音样本等[26-28](见图2(b)). 本文主要基于样本的邻域特征对样本进行划分,取每个样本的5个最近邻样本,在5个最近邻样本中与该样本同类的个数记为k,k的取值为0～5,根据k的取值对样本进行如下划分:安全样本(k为5或4)、边界样本(k为3、2或1)、噪音样本(k为0).

由于以往关于不平衡学习的研究文献中,对于不平衡数据集类别重叠度的度量并没有明确定义,本文在对样本进行划分的基础上,利用正类样本中边界样本所占的比例度量类别重叠度. 如图2(a)所示不平衡数据集中,正类样本中几乎无边界样本,均属于安全样本,因此其类别重叠度为0%,而图2(b)所示不平衡

数据集的类别重叠度将近70 %.



(a) 类别间无重叠的不平衡数据集



(b) 类别间存在重叠的不平衡数据集

图 2 类别不平衡及样本分类

2.2 RECGAN方法

在对数据集进行结构特征分析的基础上,考虑到正类样本在不平衡数据集中较少,但所含信息更为重要,因此对于两类样本应给予区别对待,尤其是噪音样本. 负类中的噪音样本在很大程度上是真实的噪音,它们可以导致正类样本的分离,增加分类的难度,因此可以考虑移除;而正类中的噪音样本却不能简单地移除,由于正类中样例较少,并不能对正类的分布进行完全表达,正类中的噪音样本极有可能是不能被其他样本代表的稀有的有效样本[29-30],即使有个别真正的噪音样本,也应将其保留在训练集中. 为克服以往主要从样本点局部邻域出发进行重抽样的缺陷,所提出RECGAN的重抽样方法大致分为两个步骤:负类样本的欠抽样、正类样本的过抽样,具体如下.

step 1:负类样本的欠抽样. 关于负类样本的欠抽样主要包括两部分,首先移除负类样本中的噪音样本,减少其对分类的干扰;其次,对重叠区域的负类样本进行适当的删减. 由于在重叠区域负类样本的数量远大于正类样本,使得重叠区域的正类边界样本往往较难识别,应降低负类在重叠区域的比重. 采用k-近邻算法对重叠区域的负类样本进行删减,结合第2.1节不平衡数据集结构特征分析中对样本的划分定义,将k-近邻算法中的k取为3,若k-近邻样本中有正类样本,则将其移除. 这样既可以达到对负类样本欠抽样的目的,又可以避免对负类样本的过度删减.

step 2:正类样本的过抽样. 首先,在对负类样本进行适当删减的基础上,将不平衡数据集的类标签变量作为CGAN模型的条件变量y,对CGAN模型进行搭建和训练;其次,当CGAN模型训练完成后,模型中的生成网络G便可以当作一种过抽样方法,有目的地生成样本,此时将随机噪声z和不平衡数据集中正类样本的类标签作为生成网络G的输入,则输出的生成样本便可以看作是基于正类样本的数据分布直接采样所得. 因此,运用训练后的CGAN模型中的生成网络对正类样本进行过抽样,以降低类别不平衡对分类的影响.

记输入的不平衡数据集为D_0,正、负类样本数分别为n_0、n_1;记正类样本的类标签为y_0;负类样本中的边界样本集合为S_1,噪音样本集合为S_2;将移除负类噪音样本的操作记作D_0-S_2;记S_1中的样本数为$\mathrm{num}(S_1)$;训练迭代次数为NI,小批次的样本量为m. 下文进一步给出RECGAN重抽样算法的伪代码描述.

算法1 RECGAN重抽样.

输入:不平衡的数据集D_0;

输出:经平衡化处理的数据集.

step 1:根据D_0,计算n_0、n_1、S_1、S_2,并对D_0进行结构特征分析.

/* 负类样本的欠抽样 */

step 2:令$D_1=D_0-S_2$.

step 3:利用最近邻算法移除负类重叠区域的部分边界样本.

for $i=1\rightarrow \mathrm{num}(S_1)$ do

计算并获得$x_i\in S_1$的3个最近邻样本,若其中有正类样本,则将其从数据集D_1中移除,得到数据集D_2.

end for

/*CGAN模型的训练 */

step 4:基于数据集D_2训练CGAN模型.

for $j=1\rightarrow$ NI do

从$U[-1,+1]$中抽取容量为m的随机噪声$\{(z_1,y_1),(z_2,y_2),\dots,(z_m,y_m)\}$;

从数据集D_2中抽取容量为m的样本$\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$;

根据如下目标函数更新判别网络:

$$\boldsymbol{E}_D=\frac{1}{m}\sum_{i=1}^{m}\{\log D(x_i|y_i)+\log(1-D(z_i|y_i))\};$$

从$U[-1,+1]$中抽取容量为m的随机噪声$\{(z_1,y_1),(z_2,y_2),\dots,(z_m,y_m)\}$;

根据如下目标函数更新生成网络:

$$\boldsymbol{E}_G=\frac{1}{m}\sum_{i=1}^{m}\log(1-D(z_i|y_i)).$$

end for

/*基于生成网络对正类样本进行过抽样*/

step 5: 从$U[-1,+1]$中抽取容量为$n=n_1-n_0$的随机噪声, 并将$\{(z_1,y_0),(z_2,y_0),\ldots,(z_n,y_0)\}$作为生成网络$G$的输入, 即可生成$n$个正类样本.

step 6: 将生成的正类样本与数据集D_2进行合并, 得到经过平衡化处理的数据集D_3.

step 7: 输出D_3.

在算法1中, CGAN模型的训练采用Adam优化算法进行, 训练中通常将条件变量转换成独热编码的形式, 记其维数为d_y, 文中重点考察不平衡的二分类数据集, 因此d_y的取值为2. CGAN模型中生成网络G和判别网络D均采用单隐层的神经网络, 且隐层均采用Relu(rectified linear unit)激活函数. 模型中生成网络G的输入为随机噪声z(记其维数为d_z)及其标签变量y, 输出为与真实数据维数相同的生成样本, 因此, 生成网络G的输入和输出节点数分别为d_z+d_y和d_x. 由于本文将生成网络G作为一种过抽样方法应用于不平衡学习, 为使生成样本能够更接近真实数据, 在生成网络G的输出层没有采用激活函数. 判别网络D作为一个二分类的判别器, 其目的在于判别出其输入是生成样本还是真实数据, 因此判别网络D的输入和输出节点数分别为d_x+d_y和1, 其输出层采用sigmoid激活函数, 输出为样本来自真实数据的概率. 将随机噪声z的先验分布设为$[-1,1]$的均匀分布, 即假设随机噪声z在均匀分布$U[-1,1]$中进行采样.

3　数值实验

为了验证RECGAN方法在具有类别重叠的不平衡数据集中的优势, 选取KEEL数据库(http://www.keel.es)中8个不平衡数据集进行实验, 基本信息如表1所示. 为突出数据集的结构特征, 将数据集的类别重叠度也列于表中, 采用正类样本中边界样本所占的比例度量.

表1　实验数据集的基本信息

数据集	正类样本数	负类样本数	IR	属性数	重叠度(%)
ecoli1	77	259	3.36	7	35.06
yeast1	429	1055	2.46	8	65.97
yeast-0-3-5-9_ vs _7-8	50	456	9.12	8	50.00
page-blocks0	559	4913	8.79	10	27.19
haberman	81	225	2.78	3	71.6
glass	70	144	2.06	9	48.57
newthyroid	35	180	5.14	5	40.00
cleveland	54	243	4.5	13	51.85

实验对于RECGAN方法与ROS[28]、SMOTE[11]、SMOTEENN[12]、ADASYN[14]以及FAST-CBUS[31]等几种重抽样方法的数据平衡化性能, 在决策树、k-近邻、支持向量机、逻辑回归4种分类器上进行实验比较. 实验硬件环境为CPU Corei7-5257U 2.70 GHz, 16 GB内存, 软件环境为Windows10操作系统, 实验工具采用python3.5.2, 其中CGAN模型的训练利用Tensorflow实现.

3.1　评价准则

作为机器学习中常用的评价分类效果的指标——整体的分类准确率, 当其应用于不平衡数据集的分类评价时, 往往会偏向负类样本, 不能正确识别正类样本, 而正类样本在数据集中又具有较为重要的信息, 因此整体的分类准确率并不能真实有效地反映分类效果. 本文基于不平衡数据的分类效果混淆矩阵, 采用更适合于不平衡数据分类评价的标准——F_1值、AUC对几种重抽样方法的性能进行评价.

文中将数据集中的少数类样本定义为正类, 多数类样本定义为负类, 二分类问题的混淆矩阵见表2.

表2　关于二分类问题的混淆矩阵

	预测为正类	预测为负类
实际为正类	TP	FN
实际为负类	FP	TN

由混淆矩阵可以得到如下度量指标:

真正率

$$\mathrm{TPrate}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

正类预测值

$$\mathrm{PPvalue}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

假正率

$$\mathrm{FPrate}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}},$$

真负率

$$\mathrm{TNrate}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

1) F_1值.

真正率又称为查全率(recall), 正类预测值又称为查准率(precision), 有

$$F_1=\frac{2\cdot\mathrm{recall}\cdot\mathrm{precision}}{\mathrm{recall}+\mathrm{precision}}.$$

查全率是分类器对正类样本正确分类是否全面的一个度量; 查准率用来表示正类样本被正确分类的比重, 通常会受到数据集不平衡率的影响, 不平衡率越高对其影响越大. 通常在不平衡分类问题中, 查全率和查准率之间是相互矛盾的, 不会同时取得大值. F_1值对查全率和查准率进行了综合, 是两者的调

和平均值,其值越高表明数据的分类效果越理想.

2) AUC.

ROC曲线用图示法反映真正率和假正率的相互关系,是对不平衡数据分类性能评价的综合指标,其X轴为假正率(FPrate),Y轴为真正率(TPrate),每一组衡量指标(FPrate, TPrate)对应不同的概率阈值,根据此概率阈值对每个样本进行类别划分. ROC曲线详情见文献[32-33].

ROC曲线下方的面积即为AUC(area under the ROC),AUC是从总体上评价分类器性能更便利的一种方法. 一般而言AUC值越大,分类器的分类效果越好,分类的理想状态下AUC取值为1,当AUC取值为0.5时,表示分类器的分类效果相当于随机猜测.

3.2 超参数的说明与优化

在CGAN模型的训练中,将生成网络和判别网络交替迭代的次数取为1,此外涉及到的一些超参数分别为:随机噪声z的维数d_z、生成网络G的隐层节点数d_G、判别网络D的隐层节点数d_D以及算法迭代过程中小批次的样本量. 在将RECGAN方法应用于不平衡学习时,生成网络G作为一种过抽样方法必须经过模型的训练过程,因此与传统的重抽样方法相比,其耗时会相对较长. 在多次实验过程中发现,实验结果对d_z的变化不太敏感,因此考虑到实验的时间问题以及生成网络G和判别网络D的输入层构造,将随机噪声的维数d_z取作相应数据集的维数d_x;小批次的样本量取为每次迭代中训练样本量的1/30(根据所选数据集通过实验选取);d_G和d_D在数据集维数d_x的2∼10倍的整数倍中选取[23].

实验采用五折交叉验证的方法对各种重抽样方法进行比较,记每一折训练集和测试集分别为T_i和$V_i, i = 1,2,\ldots,5$,对每一折的T_i采用重抽样算法进行平衡处理,之后在该训练集上训练分类器,在测试集V_i上评估分类器的分类效果,采用5次AUC和F_1的平均值作为分类效果的评价指标.

在每一折训练集T_i上进行模型中超参数的优化,需将T_i再次划分为训练集T_{it}和验证集T_{iv},实验中将T_i的80%作为训练集T_{it},20%作为验证集T_{iv}. 在T_{it}上进行算法超参数的选取并训练分类器,算法超参数的选取标准为:使得分类器在验证集T_{iv}上的AUC达到最大. 对于随机过抽样、SMOTE以及ADASYN几种重抽样算法超参数的选取过程,与RECGAN方法超参数的优化过程一致.

在CGAN模型的训练中,将模型的训练次数取为10 000时,以支持向量机分类器为例,在各个数据集上G和D的隐层节点数的最优选择如表3所示.

表3 数据集对应CGAN模型中G和D的隐层节点数

数据集	d_D	d_G
ecoli1	14	70
yeast1	40	40
yeast-0-3-5-9_ vs _7-8	80	24
page-blocks0	20	70
haberman	6	9
glass	45	72
newthyroid	50	25
cleveland	52	26

3.3 实验结果及分析

实验对于具有类别重叠的不平衡数据集,利用支持向量机、逻辑回归、决策树及k-近邻分类器评估RECGAN重抽样方法的数据平衡化性能,并与随机过抽样(ROS)、SMOTE、SMOTEENN、ADASYN、FAST-CBUS等重抽样方法的性能进行实验比较. 实验中把没有进行任何重抽样处理的分类器的分类效果也作为比较对象之一,以便对各种重抽样方法的性能有更清晰的了解. 实验结果主要分两个方面进行分析:一方面是基于不同分类器,结合各种重抽样方法对数据进行分类,比较分类器在所有数据集上的AUC(表4)和F_1值(表5);另一方面是基于不同分类器,在AUC和F_1值两种评价准则下,分别比较各种重抽样方法在数据集上的平均排序(表6).

1) 基于AUC和F_1值的分类性能比较.

在表4和表5中,每行加粗字体表示该行的最高值,从中可以看出,在利用决策树、支持向量机和k-近邻分类器进行分类时,RECGAN方法明显优于其他几种重抽样方法,其在绝大部分数据集上的AUC和F_1值是最高的,且在所有数据集上的平均AUC和F_1值也是最高的. 在利用逻辑回归分类器进行分类时,虽然在AUC的评价准则下RECGAN方法没有表现出太明显的优势,一方面可能与部分数据集的结构特征有关系,另一方面,由于CGAN模型的训练时间问题,实验中对于模型中生成网络和判别网络隐层节点数的选择并不是很充分,模型的训练次数等部分超参数的选择均与数据集有很大的关系;但在F_1值的评价准则下,RECGAN方法仍明显优于其他几种重抽样方法,其在数据集上的平均F_1值仍是最高的.

2) 平均排序比较.

在AUC和F_1值两种评价准则下,几种重抽样方法在数据集上的平均排序如表6所示. 包括不作重抽样处理的分类效果在内,共有7种进行比较的重抽样

方法,因此表中的排名值应分布在1～7之间,且其值越小表明该种重抽样方法的排名越靠前,其数据平衡化性能越好越有助于数据的分类. 表6中加粗字体表示对应的重抽样方法在数据集上的排名是最优的.

由表6可见,当运用支持向量机、决策树及k-近邻分类器进行分类时,在两种评价准则下RECGAN方法均体现出了明显的优势,其在几种重抽样方法中的排名均是最优的;当运用逻辑回归分类器进行分类时,在F_1值的评价准则下RECGAN方法的排名是最优的,表现出了很大的优势.

表 4　重抽样方法在4种分类器上的AUC值

	DateSet	None	SMOTE	ROS	ADASYN	SMOTEENN	FASTCBUS	RECGAN
支持向量机分类器	ecoli1	0.801	0.879	0.877	**0.886**	0.884	0.801	0.880
	yeast1	0.524	0.696	0.704	0.696	0.685	0.575	**0.709**
	yeast-3-5-9_vs_7-8	0.528	0.665	0.652	**0.686**	0.638	0.607	0.609
	page-blocks0	0.582	0.641	0.610	0.634	0.644	0.604	**0.648**
	haberman	0.514	0.511	0.511	0.503	0.604	0.514	**0.639**
	glass	0.618	**0.770**	0.753	0.750	0.736	0.720	0.739
	newthyroid	0.629	0.800	0.643	0.743	0.786	**0.864**	0.766
	cleveland	0.500	0.597	0.612	0.606	0.592	0.504	**0.619**
	平均值	0.587	0.695	0.670	0.688	0.696	0.649	**0.701**
决策树分类器	ecoli1	0.856	0.875	0.844	0.842	0.881	0.791	**0.890**
	yeast1	0.649	0.671	0.651	0.650	0.713	0.636	**0.718**
	yeast-3-5-9_vs_7-8	0.638	0.588	0.551	0.645	0.670	0.586	**0.675**
	page-blocks0	0.913	0.922	0.896	0.912	**0.954**	0.907	0.939
	haberman	0.571	0.559	0.572	0.566	0.624	0.557	**0.647**
	glass	0.766	**0.834**	0.794	0.766	0.818	0.741	0.832
	newthyroid	0.932	0.935	0.932	**0.980**	0.960	0.923	0.938
	cleveland	0.478	0.508	0.491	0.505	0.473	**0.527**	0.492
	平均值	0.725	0.736	0.716	0.733	0.762	0.709	**0.766**
逻辑回归分类器	ecoli1	0.774	0.884	0.899	**0.906**	0.881	0.839	0.886
	yeast1	0.591	0.713	0.713	0.702	0.707	**0.790**	0.701
	yeast-3-5-9_vs_7-8	0.549	0.750	0.719	0.732	0.701	**0.792**	0.660
	page-blocks0	0.772	0.906	0.907	0.906	**0.927**	0.916	0.888
	haberman	0.552	0.632	0.646	0.659	0.654	0.680	**0.686**
	glass	0.649	0.747	0.727	0.737	0.743	0.728	**0.763**
	newthyroid	**0.997**	0.994	0.992	0.992	0.980	0.990	0.990
	cleveland	0.494	**0.600**	0.598	0.595	0.573	0.533	0.589
	平均值	0.672	0.778	0.775	0.779	0.771	**0.784**	0.770
k-近邻分类器	ecoli1	0.859	0.885	0.865	0.869	**0.904**	0.835	0.892
	yeast1	0.639	0.683	0.671	0.682	0.691	0.631	**0.701**
	yeast-3-5-9_vs_7-8	0.640	0.687	0.692	0.698	**0.712**	0.638	0.692
	page-blocks0	0.859	0.927	0.929	0.932	0.933	0.875	**0.933**
	haberman	0.580	0.615	0.569	0.588	0.608	0.541	**0.666**
	glass	0.777	0.775	0.782	0.811	0.766	0.802	**0.855**
	newthyroid	0.871	0.926	0.935	0.946	0.910	**0.952**	0.930
	cleveland	0.480	0.384	0.474	0.448	0.399	**0.569**	0.508
	平均值	0.713	0.735	0.740	0.747	0.740	0.730	**0.772**

表 5 重抽样方法在4种分类器上的F_1值

	DateSet	None	SMOTE	ROS	ADASYN	SMOTEENN	FASTCBUS	RECGAN
支持向量机分类器	ecoli1	0.711	0.755	0.751	**0.757**	**0.757**	0.703	0.755
	yeast1	0.107	0.570	0.579	0.571	0.561	0.286	**0.585**
	yeast-3-5-9_vs_7-8	0.103	0.344	0.319	0.285	0.245	0.334	**0.339**
	page-blocks0	0.280	0.441	0.352	0.405	0.427	**0.569**	0.417
	haberman	0.068	0.212	0.172	0.233	0.434	0.070	**0.491**
	glass	0.410	0.681	0.665	0.662	0.649	0.586	**0.683**
	newthyroid	0.379	**0.731**	0.429	0.641	0.719	0.679	0.633
	cleveland	0.000	0.345	**0.367**	0.359	0.346	0.186	0.354
	平均值	0.257	0.510	0.454	0.489	0.517	0.427	**0.532**
决策树分类器	ecoli1	0.768	**0.791**	0.765	0.737	0.781	0.700	0.769
	yeast1	0.502	0.533	0.503	0.505	**0.590**	0.481	0.551
	yeast-3-5-9_vs_7-8	0.332	0.248	0.175	0.322	0.323	0.252	**0.354**
	page-blocks0	0.839	0.824	0.824	0.798	0.834	0.828	**0.841**
	haberman	0.370	0.356	0.372	0.373	0.454	0.352	**0.496**
	glass	0.685	**0.768**	0.720	0.687	0.737	0.667	0.754
	newthyroid	0.883	0.895	0.885	**0.960**	0.917	0.896	0.885
	cleveland	0.119	0.213	0.167	0.222	0.213	0.224	**0.228**
	平均值	0.562	0.579	0.551	0.575	0.606	0.550	**0.610**
逻辑回归分类器	ecoli1	0.685	0.762	0.774	**0.775**	0.744	0.767	0.763
	yeast1	0.340	**0.589**	0.589	0.576	0.582	0.451	0.576
	yeast-3-5-9_vs_7-8	0.173	**0.388**	0.360	0.335	0.294	0.313	0.368
	page-blocks0	0.675	0.708	0.714	0.650	0.719	0.720	**0.748**
	haberman	0.223	0.460	0.478	0.503	0.495	0.233	**0.535**
	glass	0.504	0.659	0.635	0.648	0.656	0.581	**0.674**
	newthyroid	**0.987**	0.973	0.960	0.960	0.940	0.930	0.958
	cleveland	0.000	0.350	**0.351**	0.346	0.330	0.222	0.341
	平均值	0.448	0.611	0.608	0.599	0.595	0.527	**0.620**
k-近邻分类器	ecoli1	**0.793**	0.778	0.751	0.743	0.791	0.745	0.761
	yeast1	0.470	0.553	0.539	0.554	0.564	0.480	**0.574**
	yeast-3-5-9_vs_7-8	**0.400**	0.336	0.367	0.341	0.330	0.353	0.380
	page-blocks0	0.794	0.786	0.804	0.766	0.765	0.801	**0.805**
	haberman	0.337	0.448	0.394	0.423	0.447	0.293	**0.515**
	glass	0.702	0.694	0.702	**0.730**	0.680	0.629	0.680
	newthyroid	0.845	0.857	0.852	0.855	0.794	**0.896**	0.864
	cleveland	0.024	0.142	**0.239**	0.215	0.220	0.216	0.184
	平均值	0.545	0.574	0.581	0.578	0.574	0.552	**0.595**

表 6　几种重抽样方法在数据集上的平均排序结果

	评价准则	None	SMOTE	ROS	ADASYN	SMOTEENN	FASTCBUS	RECGAN
支持向量机分类器	AUC	6.25	2.88	3.63	3.13	3.38	4.88	**2.50**
	F_1	6.63	2.75	3.75	3.38	3.50	4.50	**2.25**
逻辑回归分类器	AUC	6.00	**2.88**	3.38	3.13	3.88	3.38	3.63
	F_1	5.88	2.88	2.88	3.38	4.13	4.75	**2.50**
决策树分类器	AUC	4.75	3.38	5.00	4.13	2.63	5.75	**1.88**
	F_1	4.63	3.75	5.00	4.00	2.50	5.25	**2.00**
k-近邻分类器	AUC	5.38	4.38	4.25	3.13	3.38	4.88	**1.88**
	F_1	4.25	4.00	3.50	4.25	4.25	4.38	**2.50**

4　结　论

本文对于具有类别重叠的不平衡数据集,在分析数据集结构特征的基础上,提出了RECGAN的重抽样方法. 该方法包括对负类样本的欠抽样和对正类样本的过抽样,欠抽样主要移除负类的噪音样本和部分边界样本,这样既可以减少噪音对分类的干扰,也降低了重叠区域负类样本的比重,提高正类样本的识别度; 正类样本的过抽样是将数据的类别标签变量作为CGAN模型的条件变量并进行训练,去学习正类样本的数据分布,然后利用模型中的生成网络基于正类样本的分布对其进行过抽样.

利用支持向量机、逻辑回归、决策树及k-近邻4种分类器和F_1值、AUC两种评价标准,在具有类别重叠的不平衡数据集上,对所提RECGAN方法的数据平衡化性能与其他几种重抽样方法进行了比较,结果表明: 从F_1值和AUC的取值看,在决策树、支持向量机和k-近邻分类器中,RECGAN方法的性能均是最优的; 从平均排名看,RECGAN方法在利用支持向量机、决策树及k-近邻分类器分类时,在两种评价准则下均是最优的; 在运用逻辑回归分类器时,RECGAN方法在F_1值准则下同样具有明显的优势.

将CGAN模型应用于不平衡学习问题的研究在数据层面可以很好地解决不平衡问题,同时还存在很大的提升空间,下一步的研究重点是: 对于不平衡数据集研究更有效的模型训练方法和超参数的优化方法; 利用CGAN模型解决多类别的不平衡学习问题; 从数据集的结构特征出发,将更有效的欠抽样方法与CGAN模型相结合,对数据集进行更合理的不平衡学习.

参考文献(References)

[1] Czarnecki W M, Rataj K. Compounds activity prediction in large imbalanced datasets with substructural relations fingerprint and EEM[C]. IEEE Trustcom/BigDataSE/ISPA. Piscataway: IEEE, 2015: 192.

[2] Wei W, Li J J, Cao L B, et al. Effective detection of sophisticated online banking fraud on extremely imbalanced data[J]. World Wide Web-internet & Web Information Systems, 2013, 16(4): 449-475.

[3] Chawla N V, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets[J]. Sigkdd Explorations Newsletter, 2004, 6(1): 1-6.

[4] 陶新民, 郝思媛, 张冬雪, 等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报: 自然科学版, 2013, 25(1): 101-110.
(Tao X M, Hao S Y, Zhang D X, et al. Overview of classification algorithms for unbalanced data[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2013, 25(1): 101-110.)

[5] Japkowicz N, Stephen S. The class imbalance problem: A systematic study[J]. Intelligent Data Analysis, 2002, 6(5): 429-449.

[6] Zhang L, Zhang D. Evolutionary cost-sensitive extreme learning machine[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(12): 3045-3060.

[7] Khan S H, Hayat M, Bennamoun M, et al. Cost sensitive learning of deep feature representations from imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(8): 3573-3587.

[8] 尹华, 胡玉平. 一种代价敏感随机森林算法[J]. 武汉大学学报: 工学版, 2014, 47(5): 707-711.
(Yi H, Hu Y P. A cost-sensitive algorithm based on random forest[J]. Engineering Journal of Wuhan University, 2014, 47(5): 707-711.)

[9] 权鑫, 顾韵华, 郑关胜, 等.一种增量式的代价敏感支持向量机[J]. 中国科学技术大学学报, 2016, 46(9): 727-735.
(Quan X, Gu Y H, Zheng G S, et al. An incremental cost-sensitive support vector machine[J]. Journal of University of Science and Technology of China, 2016, 46(9): 727-735.)

[10] Wozniak M. Hybrid classifiers: Methods of data, knowledge, and classifier combination[M]. New York: Springer Heidelberg, 2013: 95-133.

[11] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 341-378.

[12] Batista G E A P A, Prati R C, Monard M C. A study of the behavior of several methods for balancing machine learning training data[J]. Sigkdd Explorations Newsletter, 2004, 6(1): 20-29.

[13] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[J]. Lecture Notes in Computer Science, 2005, 3644(5): 878-887.

[14] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. IEEE International Joint Conference on Neural Networks. Piscataway: IEEE, 2008: 1322-1328.

[15] Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data[C]. Computational Intelligence and Data Mining. Piscataway: IEEE, 2011: 104-111.

[16] Napierala K, Stefanowski J. Types of minority class examples and their influence on learning classifiers from imbalanced data[J]. Journal of Intelligent Information Systems, 2016, 46(3): 563-597.

[17] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[18] Bousmalis K, Silberman N, Dohan D, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks[J]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 95-104.

[19] Antoniou A, Storkey A, Edwards H. Data augmentation generative adversarial networks[J]. 2017, arXiv: 1711.04340.

[20] 孙亮, 韩毓璇, 康文婧, 等.基于生成对抗网络的多视图学习与重构算法[J]. 自动化学报, 2018, 44(5): 819-828.
(Sun L, Han Y X, Kang W J, et al. Multi-view learning and reconstruction algorithms via generative adversarial networks[J]. Acta Automatic Sinica, 2018, 44(5): 819-828.)

[21] 唐贤伦, 杜一铭, 刘雨微, 等. 基于条件深度卷积生成对抗网络的图像识别方法[J]. 自动化学报, 2018, 44(5): 855-864.
(Tang X L, Du Y M, Liu Y W, et al. Image recognition with conditional deep convolutional generative adversarial networks[J]. Acta Automatic Sinica, 2018, 44(5): 855-864.)

[22] Mirza M, Osindero S. Conditional generative adversarial nets[J]. 2014, arXiv: 1411.1784.

[23] Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks[J]. Expert Systems with Application, 2018, 91: 464-471.

[24] Napierala K, Stefanowski J, Wilk S. Learning from imbalanced data in presence of noisy and borderline examples[C]. International Conference on Rough Sets & Current Trends in Computing. Berlin: Springer-Verlag, 2010, 6086: 158-167.

[25] Denil M, Trappenberg T. A characterization of the combined effects of overlap and imbalance on the SVM classifier[J]. Computer Science, 2011, https://www.researchgate.net/publication/51937982.

[26] Stefanowski J. Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data[J]. Emergin Paradigms in Machine Learning, 2013, 13: 277-306.

[27] Saez J, Luengo J, Stefanowski J, et al. SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291: 184-203.

[28] Haixiang G, Yijing L, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications, 2017, 73: 220-239.

[29] Gamberger D. Experiments with noise filtering in a medical domain[C]. The 16th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 1999: 143-151.

[30] Zhu X, Wu X, Ying Y. Error detection and impact-sensitive instance ranking in noisy datasets[C]. National Conference on Artifical Intelligence. AAAI Press, 2014: 378-383.

[31] Ofek N, Rokach L, Stern R, et al. Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem[J]. Neurocomputing, 2017, 243: 88-102.

[32] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 33-35.
(Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 33-35.)

[33] Fawcett T. An introduction to ROC analysis[J]. Pattern Rcognition Letters, 2006, 27(8): 861-874.

作者简介

赵海霞(1979−), 女, 讲师, 博士生, 从事机器学习的研究, E-mail: Zhaohx@sxufe.edu.cn;

石洪波(1965−), 女, 教授, 博士, 从事数据挖掘等研究, E-mail: Shihb@sxufe.edu.cn;

武建(1978−), 男, 讲师, 博士生, 从事图论、图神经网络的研究, E-mail: wujian@sxufe.edu.cn;

陈鑫(1995−), 男, 硕士生, 从事数据挖掘的研究, E-mail: 462240557@qq.com.

(责任编辑：郑晓蕾)