一种基于 SMOTE 的不平衡数据集重采样方法

张天翼 丁立新
(武汉大学计算机学院 湖北 武汉 430072)

摘 要 不平衡数据集是指在数据集中,某一类样本的数量远大于其他类样本的数量,其会影响分类结果,使基本分类器偏向多数类。合成少数样本过采样技术(SMOTE)是处理数据不平衡问题的一种经典过采样方法,以两个少数样本对应的线段为端点生成一个合成样本。提出一种基于 SMOTE 的少数群体过采样方法,改进生成新样本的方式,在合成样本的过程中参考两个以上的少数类样本,增加合成样本的多样性。实验结果表明,在不同的基本分类器下该方法可以获得更好的接收者操作特征曲线面积(ROC-AUC)和稳定性。

关键词 不平衡数据集 过采样 样本合成 分类

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.09.043

A NEW RESAMPLING METHOD BASED ON SMOTE FOR IMBALANCED DATA SET

Zhang Tianyi Ding Lixin
(*School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China*)

Abstract The imbalanced data set refers to more instances in one class than that in other classes, which can influence classification results, and make basic classifiers have bias towards the majority class. Synthetic minority over-sampling technique (SMOTE) is one of over-sampling methods dealing with data imbalance problem, this method generates one synthetic sample according to a line segment of two minority samples as endpoint. This paper proposes a new over-sampling method of the minority class based on SMOTE. This method made improvement on how to generate new samples, it took more than two real samples into account to generate one synthetic sample, which increased diversity of synthetic samples. The experimental results show that this method achieves better area under curve and stability.

Keywords Imbalanced dataset Over-sampling Sample synthesis Classification

0 引 言

现实中的数据集通常是不平衡的,不平衡数据集中的实例分布十分不均衡。当基于不平衡数据集构造分类器时,分类器的预测结果可能会偏向多数类,这些分类器很容易将少数样本误分类为多数类。但是有时少数类样本才是问题的主要研究对象,在这种情况下,少数类样本的错误分类可能会带来严重的问题和风险。例如,在医学数据集中,健康人的样本通常远远多于患者样本,如果基于此数据集构建分类器,那么输入一个测试样本,分类器大概率会将输入样本预测为健康人,但是将患者误分类为健康人的风险远高于将健康人误分类为患者的风险。数据失衡不仅出现在医学检测中,而且也出现在许多其他实际应用中,例如海上雷达图像中油污泄露区域检测[1]、电信欺诈检测[2]等。

研究者们已经开发出了许多方法来消除数据不平衡所带来的影响,这些方法大都在算法层面或数据层面来解决不平衡问题。算法层面的方法主要包括集成学习法和成本敏感型学习法。传统分类算法的目标是平衡的数据集,因此数据集中的所有样本都具有相同的重要性,并且将 A 误分类为 B 和将 B 误分类为 A 的代价是相同的。但是在不平衡的数据集中,对于少数类而言,拥有与多数类样本相等的误分类成本并不公

收稿日期:2019-11-12。广东省珠海市产学研合作项目(2010A090200067,2016B090918097,2012D0501990016,2012D0501990026)。
张天翼,硕士生,主研领域:机器学习。**丁立新**,教授。

平。因为在一些问题中,少数类相比于其他类具有更大的研究价值。成本敏感型学习方法则修改了各类错误的惩罚因子,分类器将少数类样本误分类为多数类样本会受到更大的惩罚,在迭代过程中会逐渐减少这类错误,因此可以弱化或消除分类器的错误偏差。AdaCost[3]是一种典型的成本敏感型学习方法。AdaCost 在迭代学习过程中为少数样本的错误分类提供了更大的惩罚因素,这使得少数样本在总体成本函数中占主导地位。

集成学习方法从数据集中生成多个独立的预测模型作为弱分类器,然后将这些模型组合为强分类器。当每个弱分类器具有相对较低的错误率时,组合的强分类器将具有比任何弱分类器低得多的错误率。研究人员已经开发了基于提升算法的改进方法来解决数据不平衡问题,例如文献[4]提出少数类合成提升算法(SMOTEBoost)、文献[5]提出随机欠采样提升算法(RUSBoost)、文献[6]提出干扰修正提升算法(PCBoost)、文献[7]提出基于模型的样本合成提升算法(MBSBoost)、文献[8]提出基于过采样的不平衡数据集成分类算法(SDPDBoost)。SMOTEBoost 使用 SMOTE 进行样本合成,并且把新样本加入到数据集中。这些新样本可以给弱分类器带来更多有关少数群体分类的信息,经过多次迭代,最终的强分类器可以得到针对少数类样本分类的提升。RUSBoost 则采用欠采样方法,随机删除一些多数类样本,然后使用处理后的数据构造弱分类器。PCBoost 算法首先对少数类进行随机过采样,然后使用信息增益率构造弱分类器。错误分类的过采样样本在最后阶段会被删除。除了基于提升算法的方法,还有其他的方法,如文献[9]提出的概率阈值袋装法,利用袋装法首先获得校准良好的后验估计,然后根据性能指标选取适当的阈值,以使其最大化。

数据层面的方法采用的主要策略是合成新样本和重采样。这些方法会重塑数据集,因此可以通过重塑每个类别中的样本数来消除数据不平衡。主要有三种重采样方式:多数类样本欠采样、少数类样本过采样和混合方法。欠采样方法会丢弃多数类中的某些内部样本,或将某些样本替换为合成样本,然后通过某种标准选择丢弃的样本或替换后的样本,以便剩余的多数样本可以保留尽可能多的原始数据信息。欠采样后,两种类型的采样数近似相等,数据集达到平衡。过采样方法通过生成新的少数类样本来消除偏斜分布的危害,生成的新样本加入数据集后,应使数据集达到平衡,并且基于这些数据集训练的分类器可以是无偏的。混合方法是上述方法的混合,它同时使用欠采样和过采样来使数据集平衡,经由数据层面的方法处理后的数据集是平衡的,因此基本分类器可以发挥其原始作用。

在以上不同类型的方法中,过采样是研究人员在解决数据不平衡问题中的一种流行策略[10],而使用较多的方法之一是少数类样本合成过采样技术(SMOTE)算法[11]。该方法根据少数样本的 k 个最近邻样本生成新的合成样本,合成样本是端点为两个最近邻少数类样本对应的线段上的随机点。由于缺乏多样性,已经有许多其他改进的算法被提出,例如文献[12]提出的边界线少数类样本合成技术(Borderline SMOTE)、文献[13]提出的自适应综合过采样(ADASYN)、文献[14]提出的基于类聚集程度的少数类样本合成(DB-SMOTE)、文献[15]提出的基于周围邻域的 SMOTE 和文献[16]提出的随机游走过采样(RWO)。针对多分类不平衡问题,文献[17]提出了基于马氏距离的适应性过采样方法(AMDO)。为了使合成的样本更具多样性,本文提出了一种改进的合成技术。与其选择两个点来构建一条线,不如在合成过程中涉及更多样本来构建平面或空间。除了过采样策略,还有许多欠采样的方法被用来解决不平衡问题,如文献[18]提出的去噪欠采样(Noise-filtered Under-sampling Scheme)。

1 背景知识

1.1 少数类样本合成过采样技术

解决数据不平衡问题的一种典型的过采样方法是 SMOTE 算法,该方法旨在弥补少数类随机过采样的缺陷。对少数类样本进行随机过采样不会使得少数类样本更具识别性,因为过采样过程其实是对样本进行复制,这种复制会使样本的决策判定越来越严格,越来越具体,导致分类过拟合。例如,如果原始决策为[0,10],则在随机过采样后,由于复制了多个少数样本,这使分类器确信少数类在较窄的范围内,分类器将给出更具体的决策区域,例如[3,6]。SMOTE 算法则采用了合成新样本的方法来增加少数类样本的数量,其基本步骤如下:

Step1 从少数类样本 $\boldsymbol{A}$ 的 K 最近邻少数类中随机选取一个 $\boldsymbol{B}$,$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的样本特征的差向量为($\boldsymbol{B}-\boldsymbol{A}$)。

Step2 从区间(0,1)中随机选取一个实数 i 作为权值。将权值 i 与差向量相乘得到 $i(\boldsymbol{B}-\boldsymbol{A})$。

Step3　把 Step2 的结果与样本 A 的特征向量相加得到合成样本 $\boldsymbol{A}+i(\boldsymbol{B}-\boldsymbol{A})$。

该技术通过生成人工样本来拓宽决策区域,因为添加到数据集中的样本位于原始样本的附近的合成样本,而不是样本本身。与带有替换的随机过采样相比,决策区域更为通用。实验表明,SMOTE 算法可以提高少数类的分类器准确性,并且 SMOTE 算法和欠采样的组合比单纯使用欠采样效果更好。SMOTE 算法在低维不平衡数据集中运行良好,但在一些实验中能观察到,SMOTE 在高维上的性能不如在低维上的性能[19]。SMOTE 包含一个参数 k,代表了取最近邻的个数,文献[20]介绍了如何选取合适的 k 值。

1.2　已有的 SMOTE 算法改进

尽管 SMOTE 算法是解决数据不平衡问题的有效工具,但它仍有一些局限性。其没有考虑多数类别即可生成合成样本,由于新样本的生成过程是随机的,因此新生成的样本可能会出现在多数类的决策区域中。随机生成的结果是两种类别的决策区域的重叠的概率会增加,这使得两个类别更难以区分[11]。前人已经提出了 SMOTE 算法的一些改进版本,大多数的改进算法都在寻找一个合适的生成区域生成新样本并尽量避免重叠的增大。文献[12]提出的 Borderline SMOTE 将少数样本划分为噪声点、危险点和安全点,首先删除噪声点,仅使用危险点进行样本合成。Borderline SMOTE 在生成过程中不仅使用少数样本,还使用多数样本,通过此方法可以加强类之间的边界。自适应 SMOTE 考虑了最近邻居和被选取的少数样本的距离[13],设置了最近邻距离的阈值,避免了样本到合成样本之间的距离过长,并根据不同样本集的内部分布特征调整阈值。基于周围邻域的 SMOTE 算法使用了最近邻的不同定义[14],该方法使用了最近的质心邻域和 Graph 邻域,以确保最近的邻域距离不太远。基于局部线性嵌入的 SMOTE 算法将局部线性嵌入算法部署到少数样本[15]。随机游走过采样(RWO)引入了基于中心极限定理的过采样方法[16],它以新生成的少数样本均值遵循原始分布的方式创建样本。当使用带有 SVM 的 SMOTE 算法作为分类器时,合成采样方法会影响 SVM 的内核归纳特征空间的性能,基于内核的 SMOTE 算法直接在 SVM 的特征空间中生成合成样本[23]。文献[24]结合了 K-means 聚类和 SMOTE 算法来创建新样本,避免了噪声的产生,有效地克服了类之间和类内部的不平衡。

2　改进算法

2.1　SMOTE 算法的局限性

SMOTE 算法首先找出每个少数类的 k 个最近邻样本,然后随机选择一个最近邻样本和一个实数来合成新样本。根据算法的描述,对于单个合成样本,只有两个真实的少数样本参与合成,并且合成样本选自两个真实样本所对应的线段上。换言之,合成样本的特征向量是两个真实样本特征向量的线性组合。整个少数类中新样本的潜在出现范围是每个少数类样本对之间的一组线段上。在低维特征空间中,这种方法足以描述潜在的少数类样本分布特点。但当特征空间维度较高时,线性关系太单调以致不足以描述潜在的少数样本的分布。因为在低维度空间中可能的真实样本落在一条线段上的概率较高,但是随着维度的增大,潜在的真实样本落入在两个样本之间线段上的可能性则会越来越小。

另外,原有的合成策略不足以改变某些分类器的偏差。例如,支持向量机分类器使用支持向量来找出分隔不同类的边界,支持向量是靠近边界的样本向量,是分类算法的核心,如果将 SVM 应用于通过 SMOTE 算法进行过采样的数据集,参与单个样本合成的真实样本存在三种可能性,即两个都是支持向量、两个都不是支持向量、一个是支持向量且一个是非支持向量,后两种可能性的合成样本几乎不能成为支持向量,因此新样本对边界的计算没有帮助。对于第一种情况,新样本不会显著改变原始边界,因为它们位于支持向量的直线上,并且这些直线与边界线段趋近平行。总体而言,SMOTE 算法在高维度上缺乏多样性,并且可能不会大大改变某些分类器的偏差。

2.2　改进 SMOTE 算法设计

基于以上分析,SMOTE 算法的缺点实际上有着相同的原因,即合成方法太单调,并且线段关系太简单以致无法适应潜在的少数类特征。为合成样本添加一些垂直偏移可以增加多样性,一种有效的方法是在生成过程中涉及更多的少数类样本。

因此,本文提出了一种改进的 SMOTE 算法,与原始的 SMOTE 算法相比,本文使用 D 个少数类样本创建了人工样本,这里 D 是特征空间的维数。首先,对于所有少数样本,计算它们的 k 个相同类别的最近邻样本集,然后对于每个少数样本,选择 D 个邻居和0 到 $1/D$ 的实数以创建新样本。该方法将合成样本空间从一维空间扩展到 D 维空间,从而使新样本更加多样

化。改进的 SMOTE 算法描述如算法 1 所示。

算法 1 改进的 SMOTE 算法

输入:训练集中的正类样本集合(少数类集合)$P=\{P_1,P_2,\cdots,P_{min}\}$;正类样本的个数 min;每一个正类需合成样本的数量 N;近邻个数 k;参与合成的近邻个数 $D(D<k)$。

输出:一个合成样本集合 *Synthetic samples*。

使用少数类集合 P 构建 Kd 树;

for $i=1$ to min **do**

找出 P_i 的 k 近邻集合:

$knn_i=\{knn_{i1},knn_{i2},\cdots,knn_{ik}\}$;

for $a=1$ to N **do**

从 knn_i 中随机选择 D 个紧邻样本:

$knn'_a=\{knn'_{a1},knn'_{a2},\cdots,knn'_{aD}\}$;

从 $\left[0,\frac{1}{D}\right]$ 中随机选择 D 个实数(可重复):

$d_a=\{d_{a1},d_{a2},\cdots,d_{aD}\}$;

计算被选取的近邻与样本 P_a 的向量差:

$diff_{an}=knn'_{an}-P_a$;

计算合成样本的向量:

$$newSample_a=P_a+\sum_{n=1}^{D}d_{an}*diff_{an};$$

将生成的新样本计入集合;

End for

End for

这里,失衡率是多数类样本个数与少数类样本个数的比值,$N=INT(maj/min-1)$,maj 是多数样本的数量。

上述的改进算法中有两个参数 k 和 D,其中 k 表示最近邻样本的数量,D 表示生成新样本过程中涉及的样本数量。原始 SMOTE 算法始终将参数 D 设置为 1,这使得样本出现的范围在真实的少数类样本的线段上。如果将 D 设置为 2,则合成样本将在平面上而不是在线段上。如果将 D 设置为特征向量的大小,则合成样本的可能范围将扩展到整个特征空间。在一些特殊情况下,合成样本的可能范围会小于预期,如选取的最近邻中存在某个样本是其他样本的线性组合,此时依然能够生成足够多样的合成样本。

新算法会输出 $min\times D$ 个合成样本,这些样本分布在整个特征空间而不是线段中。因此,本文改进的 SMOTE 算法的结果会更加多样化,并且能够表示潜在分布特征。与 RWO 算法相比,本文改进的 SMOTE 算法生成的人工样本具有更多的局部分布特征。

3 实 验

准确率是衡量分类器性能的通用指标,但是当数据集不平衡时,准确率并不能很好地体现分类器对于少数类样本的分类性能。由于数据集中包含大量多数类样本,因此多数类的准确率主导了整体准确率。为了评估分类器的整体性能,研究者们使用了许多其他指标,例如 AUC 和 F-measure。AUC 是 ROC(接收器工作特性曲线)曲线下的面积,ROC 曲线是表示在不同分类标准下真阳率和假阳率变化的曲线。根据不同的分类器,标准也有所不同。由于 ROC 曲线下的面积是不同标准的积分结果,针对分类器的整体度量,因此该度量仅与分类器和数据集有关。ROC 曲线如图 1 所示,其中:曲线最左边点的坐标为(0,0),最右边点的坐标为(1,1)。AUC 则是 ROC 曲线下的的面积,即 ROC 在[0,1]区间的定积分。

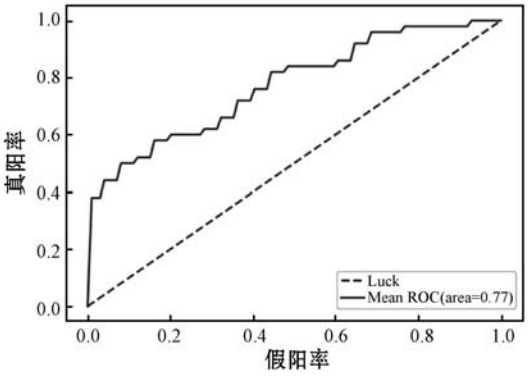

图 1 ROC 曲线示例

F-measure 是精度和召回率的加权谐波平均值。由于多数类的权重更多地取决于准确性,因此手动为少数类设置适当的权重可以对分类器进行公平的评估。F 测度的公式为:

$$F\text{-}measure=\frac{(1+\beta^2)\times recall\times precision}{\beta^2\times precision+recall} \tag{1}$$

式中:$recall$ 为召回率,$recall=TP/(TP+FN)$;$precision$ 为准确率,$precision=TP/(TP+FP)$;β 为谐波系数,设置 $\beta=1$;*F-measure* 为 F1-measure,本文实验中也使用了 F1-measure 作为衡量指标之一。实验中选择 AUC、少数类召回率、少数类准确率及 F1 量度作为度量标准,因为这些衡量指标更多地集中于分类器的整体表现。

本文使用的数据集来自 UCI 机器学习数据库。数据集包括 Adults、Forest、Phoneme 和 Pima。在这些数据集中 Adults、Phoneme 和 Pima 是二分类集,Forest 是多分类集。由于 Forest 数据集具有两个以上的类别,所以手动选择一个类作为少数类,并将其余的类合并为一个类作为多数类。某些数据集包含名义属性,SMOTE 算法是为数字属性设计的,不能用于名词性属性。为了方便起见,将这些名词性属性删除。改进算法中的参数 D 则会根据数据集的属性数有所改变。Adult、Forest、Phoneme 和 Pima 的参数 D 分别为 14、12、5

和 8。表 1 展示了每个数据集的详细信息。

表 1　数据集信息

数据集名称	标签数	属性数	少数类样本数量	多数类样本数量	不平衡率
Adult	2	14	11 687	37 155	3.18
Forest	7	12	2 747	35 754	13.02
Phoneme	2	5	1 586	3 818	2.41
Pima	2	8	268	500	1.86

实验中应用了不同的机器学习算法作为过采样数据集的分类器,包括 KNN、CART、朴素贝叶斯分类器(Bayes)和支持向量机(SVM)。这些分类器是基于 scikit-learn(https://scikit-learn.org/)构建。

实验测试了三种过采样方法,分别为 SMOTE 算法、本文改进的 SMOTE 算法和 RWO 算法。SMOTE 算法是本文中改进算法的原算法。RWO 算法是一种基于中心极限定理的过采样算法,在合成新样本的过程中首先会计算出所有少数类样本的正态分布,再根据这个分布产生新样本。所以新样本是根据所有少数类样本产生的,并且在所有属性上都具有多样性。本文算法是针对原算法在合成样本多样性上的改进,因此选用 SMOTE 算法和 RWO 算法作为对照比较。由于所有这些方法均包含随机因素,因此单次实验无法有效反映算法的性能。针对每种过采样方法和分类器进行了 30 次重复实验,最终结果是所有结果的平均值。每种过采样算法和分类算法的实验结果如表 3 – 表 5 所示。4 个指标通过十折交叉验证进行评估,每个指标的评估将产生 10 个实验结果,并且表中显示的结果是所有验证结果的均值。

表 2　Adult 数据集实验结果

过采样方法	分类器	准确率	召回率	F1-measure	ROC-AUC
本文方法	KNN	0.725	0.707	0.716	0.746
	CART	0.867	0.864	0.853	0.852
	Bayes	0.880	0.310	0.458	0.865
	SVM	0.578	0.987	0.729	**0.773**
SMOTE	KNN	0.728	0.701	0.714	0.738
	CART	0.865	0.892	0.875	0.833
	Bayes	0.875	0.304	0.451	0.839
	SVM	**0.581**	0.988	**0.732**	0.742
RWO	KNN	**0.953**	**0.770**	**0.815**	**0.901**
	CART	**0.873**	**0.872**	**0.860**	**0.859**
	Bayes	**0.951**	**0.823**	**0.854**	**0.866**
	SVM	0.576	**0.989**	0.728	0.642

表 3　Forest 数据集实验结果

过采样方法	分类器	准确率	召回率	F1-measure	ROC-AUC
本文方法	KNN	0.936	0.988	**0.960**	**0.978**
	CART	0.962	0.949	0.952	0.963
	Bayes	0.801	0.845	0.821	0.889
	SVM	**1.000**	0.079	0.144	**0.871**
SMOTE	KNN	0.928	**0.994**	0.959	0.970
	CART	0.959	**0.980**	**0.969**	**0.956**
	Bayes	0.799	0.845	0.820	0.888
	SVM	**1.000**	0.229	0.357	0.796
RWO	KNN	**0.955**	0.928	0.920	0.957
	CART	**0.966**	0.928	0.926	0.944
	Bayes	**0.999**	**0.928**	**0.943**	**0.964**
	SVM	0.518	**1.000**	**0.682**	0.502 5

表 4　Phoneme 数据集实验结果

过采样方法	分类器	准确率	召回率	F1-measure	ROC-AUC
本文方法	KNN	0.892	**0.942**	**0.916**	**0.966**
	CART	0.905	0.899	0.902	0.890
	Bayes	0.736	0.843	0.786	0.820
	SVM	0.823	**0.933**	0.874	0.927
SMOTE	KNN	0.893	0.934	0.913	0.962
	CART	0.902	0.904	0.903	0.85
	Bayes	0.736	0.842	0.785	0.819
	SVM	0.821	0.932	0.873	0.915
RWO	KNN	**0.948**	0.872	0.903	0.965
	CART	**0.927**	**0.906**	**0.915**	**0.895**
	Bayes	**0.849**	**0.876**	**0.860**	**0.890**
	SVM	**0.882**	0.925	**0.901**	**0.938**

表 5　Pima 数据集实验结果

过采样方法	分类器	准确率	召回率	F1-measure	ROC-AUC
本文方法	KNN	0.725	**0.864**	**0.787**	0.838
	CART	0.755	0.748	0.748	0.745
	Bayes	0.776	**0.719**	**0.745**	**0.837**
	SVM	**0.528**	0.987	**0.688**	**0.880**
SMOTE	KNN	0.727	0.851	0.783	0.831
	CART	0.747	0.746	0.745	0.737
	Bayes	0.774	0.708	0.738	0.826
	SVM	0.524	0.993	0.686	0.806

续表 5

过采样方法	分类器	准确率	召回率	F1-measure	ROC-AUC
RWO	KNN	**0.806**	0.732	0.750	**0.856**
	CART	**0.764**	**0.766**	**0.755**	**0.764**
	Bayes	**0.814**	0.595	0.628	0.824
	SVM	0.524	**0.996**	0.686	0.691

为了更加直观地比较三种方法的综合性能，本文特别比较了三种方法在不同数据集和分类算法下的 ROC-AUC 指数，如图 2 – 图 5 所示。

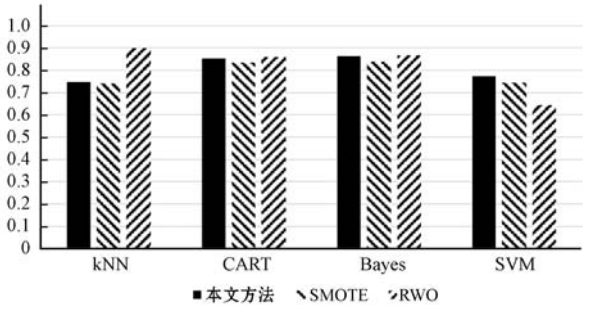

图 2 Adult 数据集 ROC-AUC 比较

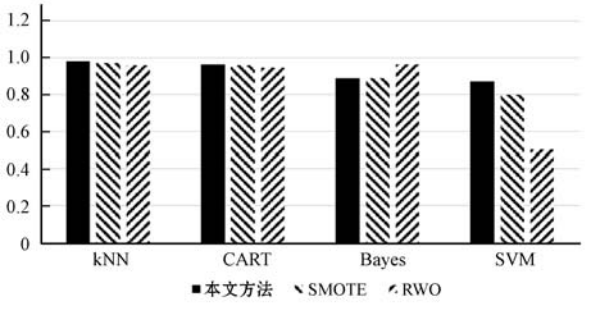

图 3 Forest 数据集 ROC-AUC 比较

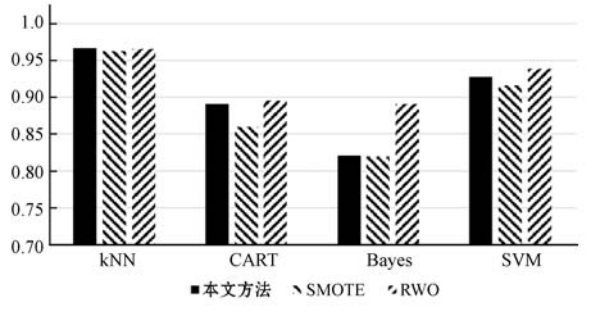

图 4 Phoneme 数据集 ROC-AUC 比较

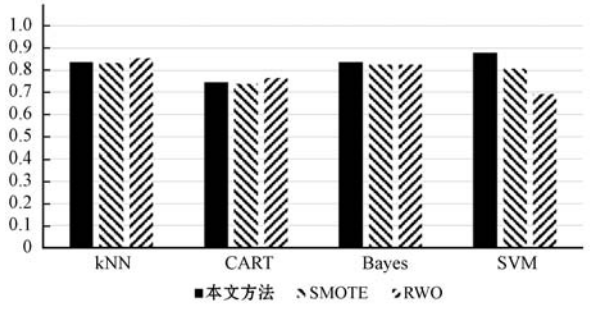

图 5 Pima 数据集 ROC-AUC 比较

根据 Adult 数据集的实验结果，本文方法具有比原始 SMOTE 算法更好的总体性能，尤其是在使用 SVM 分类器的 Forest 和 Pima 数据集的结果上，本文方法在这些数据集上实现了更高的 ROC-AUC。至于其他指标和测试，结果提升了 1% ~2%。当使用 CART 分类器对过采样的数据集进行分类时，SMOTE 算法的性能要优于本文方法，SMOTE 算法在召回率、F1 和 ROC-AUC 方面表现更好。在高失衡率数据集中，本文方法的性能不如 RWO 算法。但是在低失衡率数据集（如 Pima）中，本文方法具有与 RWO 算法类似的结果。综合结果表明本文方法优于其他两种方法，特别是在使用 SVM 时，而 RWO 算法在使用朴素贝叶斯分类器时具有更好表现。

4 结 语

数据不平衡会影响基本分类器的分类结果，使它们很难对少数类进行公平的分类。为了解决这个问题，SMOTE 算法被提出以通过生成少数样本的合成来达到平衡。本文提出了一种 SMOTE 方法的改进，使算法产生的合成样本更具多样性。实验表明，该方法在召回率、F1 和 ROC-AUC 方面比原始 SMOTE 算法具有更好的性能，并且在使用 SVM 分类器的低失衡率数据集上特别有效。本文算法比原始 SMOTE 算法在综合性能上也有一定的提升，在使用不同的分类算法时，本文方法和 RWO 算法也会有不同的表现。在使用朴素贝叶斯分类器时，RWO 算法优于本文方法；使用支持向量机时，本文方法则会有更好的综合性能。尽管在整体实验结果上，本文方法优于 SMOTE 算法，但是当数据集的不平衡率较高时，RWO 算法会比本文方法更好。因此，当数据集高度不平衡时，还需要探索更有效的改进策略。

本文方法比原始 SMOTE 算法多一个设定参数。对于不同的数据集，最佳参数是不同的，如何设置适当的参数是有待解决的问题。未来可尝试将其他的一些改进版本的 SMOTE 上使用的策略移植到本文方法上，多种策略融合或许是处理非平衡数据集分类问题的可选途径。

参 考 文 献

[1] Kubat M, Holte R C, Matwin S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning, 1998, 30: 195 – 215.

[2] Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters,2006,27(8):861 – 874.

[3] Fan W, Stolfo S J, Zhang J X, et al. AdaCost: Misclassification cost-sensitive boosting[C]//Proceedings of the 16th International Conference on Machine Learning,1999:97 – 105.

[4] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]//7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Springer,2003:107 – 119.

[5] Seiffert C, Khoshgoftaar T M, Hulse J V, et al. RUSBoost: Improving classification performance when training data is skewed[C]//2008 19th International Conference on Pattern Recognition. IEEE,2008:1 – 4.

[6] Li X F, Li J, Dong Y F, et al. A new learning algorithm for imbalanced data-PCBoost[J]. Chinese Journal of Computers, 2012,35(2):202 – 209.

[7] Liu C L, Hsieh P Y. Model-based synthetic sampling for imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering,2020,32(8):1543 – 1556.

[8] 张菲菲,王黎明,柴玉梅. 一种改进过采样的不平衡数据集成分类算法[J]. 小型微型计算机系统,2018,39(10):2162 – 2168.

[9] Collell G, Prelec D, Patil K R. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data[J]. Neurocomputing,2018, 275: 330 – 340.

[10] Guo H X, Li Y J, Shang J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systems with Applications,2017,73:220 – 239.

[11] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16:321 – 357.

[12] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]//2005 International Conference on Intelligent Computing. Springer,2005:878 – 887.

[13] He H B, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE International Joint Conference on Neural Networks(IEEE World Congress on Computational Intelligence). IEEE,2008:1322 – 1328.

[14] Liu Y X, Liu S M, Liu T, et al. New oversampling algorithm DB_SMOTE[J]. Computer Engineering and Applications, 2014,50(6):95 – 95.

[15] García V, Sánchez J S, R. Martín-Félez R. Surrounding neighborhood-based SMOTE for learning from imbalanced data sets[J]. Progress in Artificial Intelligence,2012,1:347 – 362.

[16] Zhang H X, Li M F. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification[J]. Information Fusion,2014,20:99 – 116.

[17] Yang X B, Kuang Q M, Zhang W S, et al. AMDO: An over-sampling technique for multi-class imbalanced problems[J]. IEEE Transactions on Knowledge and Data Engineering,2018,30(9):1672 – 1685.

[18] Kang Q, Chen X S, Li S S, et al. A noise-filtered under-sampling scheme for imbalanced classification[J]. IEEE Transactions on Cybernetics,2017,47(12):4263 – 4274.

[19] Blagus R, Lusa L. Class prediction for High-Dimensional Class-Imbalanced Data[J]. BMC Bioinformatics, 2010, 11: 523.

[20] Yun J, Ha J, Lee J S. Automatic determination of neighborhood size in SMOTE[C]//Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. ACM,2016:1 – 8.

[21] 杨智明,乔立岩,彭喜元. 基于改进 SMOTE 的不平衡数据挖掘方法研究[J]. 电子学报,2007,35(S2):22 – 26.

[22] Wang J J, Xu M T, Wang H, et al. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding[C]//2006 8th international Conference on Signal Processing. IEEE,2006.

[23] Mathew J, Luo M, Pang C K, et al. Kernel-based SMOTE for SVM classification of imbalanced datasets[C]//IECON 2015- 41st Annual Conference of the IEEE Industrial Electronics Society. IEEE,2015:1127 – 1132.

[24] Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information Sciences,2018,465:1 – 20.

(上接第 261 页)

[16] Park S W, Kang M G. NLM algorithm with weight update[J]. Electronics Letters, 2010, 46(15):1061 – 1063.

[17] 刘晓明,田雨,何徽,等. 一种改进的非局部均值图像去噪算法[J]. 计算机工程,2012,38(4):199 – 201, 207.

[18] 张玉征. 基于改进的非局部均值图像去噪算法研究[D]. 南昌:南昌航空大学,2019.

[19] 曹璟,周宁宁,洪龙. 基于边缘检测的自适应非局部均值去噪算法[J]. 济南大学学报(自然科学版),2016,30(3):209 – 214.

[20] Haralick R M. Digital step edges from zero crossing of second directional derivatives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1984,6(1):58 – 68.