文章编号: 1001-0920(2017)10-1831-06 DOI: 10.13195/j.kzyjc.2016.1044

基于边界混合采样的非均衡数据处理算法

冯宏伟[1], 姚 博[1], 高 原[2], 王惠亚[3], 冯 筠[1†]

(1. 西北大学 信息科学与技术学院，西安 710127；2. 西北大学 经济管理学院，西安 710127； 3. 西北大学 数学学院，西安 710127)

摘 要: 针对非均衡数据分类效果差的问题,提出一种新的基于边界混合采样的非均衡数据处理方法(BMS). 首先通过引进“变异系数”找出样本的边界域和非边界域;然后对边界域中的少数类样本进行过采样,对非边界域中的多数类样本进行随机欠采样,以期达到训练数据基本平衡的目标. 实验结果表明,BMS方法比其他3种流行的非均衡数据处理方法在对7个公开数据集的分类性能上平均提高了5%左右,因此,该方法可以广泛应用于非均衡数据的处理和分类中.

关键词: 非均衡数据；欠采样；变异系数；分类

中图分类号: TP181 **文献标志码:** A

Imbalanced data processing algorithm based on boundary mixed sampling

FENG Hong-wei[1], *YAO Bo*[1], *GAO Yuan*[2], *WANG Hui-ya*[3], *FENG Jun*[1†]

(1. School of Information Science and Technology，Northwest University，Xi'an 710127，China；2. School of Economics and Management，Northwest University，Xi'an 710127，China；3. School of Mathematics，Northwest University，Xi'an 710127，China)

Abstract: Aiming to solve the poor performance of imbalanced data classification, an novel imbalanced data classification algorithm based boundary mixed sampling(BMS) is proposed. This method firstly introduces coefficient of variation is to find out the boundary and non-boundary samples and then deal with them in different ways. The minority samples in boundary are over sampled while the non-boundary majority ones are under sampled to achieve a basic balance of samples. Experimental results show that the proposed method achieves the better classification performance by 5% than other three popular methods in seven UCI datasets, thus this method can be widely used in imbalanced data processing and classification.

Keywords: imbalanced datasets；under-sampling；variable coefficient；clustering

0 引 言

近年来,非均衡数据挖掘作为数据分析领域10大挑战性难题之一,受到业界学者的广泛关注[1]. 现实应用中存在大量非均衡数据,例如信用卡欺诈数据[2]、网络入侵数据[3]、癌症病人的诊断数据[4]、客户流失数据等. 这类应用中的少数类样本通常蕴含重要的信息,是数据分析的重要目标,已成为数据挖掘研究的热点之一.

针对非均衡数据分类问题,人们主要从分类算法层面和数据处理层面展开研究. 基于算法层面的方法主要通过对传统分类算法加入惩罚机制或利用集成学习对数据分布的非平衡性进行补偿,例如代价敏感学习[5]、主动学习[6]、集成学习方法[7]和单类学习方法[8]等. 这类方法保留了最原始的数据分布,但使用范围相对较窄,同时对算法的改进难度较大. 基于数据处理层面的方法则通过抽样或内插的方法改变数据的样本分布,以改善样本数据的不平衡性. 其中常用的方法是随机欠采样和随机过采样.

随机欠采样[9]通过随机删除数据集中的多数类样本降低数据的不平衡程度,但是随机删除样本容易舍弃多数类中有用的样本信息从而造成信息丢失. 文献[10]针对随机欠采样在生成新样本过程中存在的盲目性问题,先对多数类样本进行谱聚类,然后在每个聚类中选择具有代表意义的信息点来实现数据样本间的数据平衡,但若与少数类样本距离较远的多数类样本包含大量有价值的信息,则删除他

收稿日期: 2016-08-15；**修回日期:** 2016-12-19.

基金项目: 陕西省教育厅科学研究计划自然科学专项项目(15JK1738)；陕西省自然科学基金项目(2014JQ8367).

作者简介: 冯宏伟(1964−), 男, 副教授, 从事数据挖掘、图形图像处理、模式识别与人工智能等研究；姚博(1990−), 男, 硕士生, 从事数据挖掘与金融大数据分析的研究.

†通讯作者. E-mail: fengjun@nwu.edu.cn

们容易造成多数类样本有价值的信息丢失. 随机过采样[11]通过精确复制少数类样本来增加少数类样本的数量,因而容易出现样本重叠和过拟合现象. 文献[12]针对随机过采样在非均衡数据分类中的不足,提出一种线性插值的SMOTE(synthetic minority oversampling technique)算法,根据过抽样率,从其k个最近邻中随机选出若干个近邻,在该样本和被选的近邻之间插入合成新的样本. 但由于SMOTE算法是对所有少数类样本进行过采样处理,也可能导致分类器出现过拟合现象.

针对欠采样方法可能导致样本信息丢失和过采样方法可能导致分类器出现过拟合现象,人们提出了混合采样的非均衡数据处理方法. 文献[13]首先删除噪声样本,然后删除边界样本中的少数类样本,对非边界样本采用改进的SMOTE合成新的样本,对比实验结果表明该算法能够有效地提高少数类的识别效果. 文献[14]提出了一种基于SMOTE-Clustering的混合采样算法,该算法首先对少数类样本采用改进的SMOTE算法进行合成,然后使用聚类的数据欠采样算法删除冗余和噪声数据,使得原始数据集达到平衡.但上述方法均直接删除噪声数据,在处理过程中会不可避免地误删部分少数类样本,造成少数类样本信息丢失.

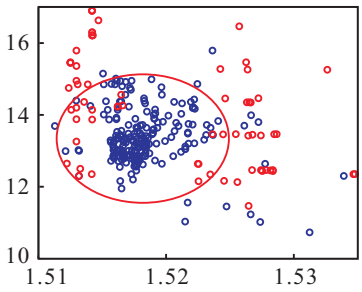
针对上述处理噪声数据时误删少数类样本造成的样本信息丢失问题,本文提出一种新的基于边界混合采样的非均衡数据处理算法(BMS). 该算法通过设置变异系数[15]阈值划分样本的边界域和非边界域,并使用SMOTE算法和用基于欧氏距离的随机欠采样方法(OSED)[16]对边界域中的少数类样本和非边界域中的多数类样本分别进行处理. 将本文提出的算法与基于样本特性欠采样的不均衡支持向量机算法、基于SMOTE过采样的SVM算法以及基于随机欠采样与SMOTE相结合的SVM算法[17]进行比较,实验结果表明,本文算法在不同数据集上的分类性能均优于上述3种算法,同时通过对比SVM和决策树分类器,表明SVM分类器更适合非均衡数据的分类.

1 基于边界混合采样的非均衡数据处理

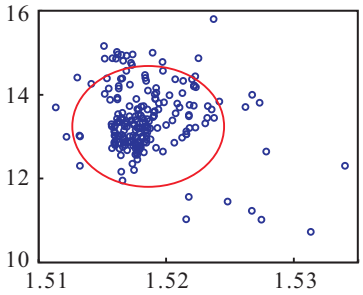
本文提出一种基于边界混合采样的非均衡数据处理方法,该算法分两阶段对样本的边界域和非边界域数据进行处理,得到的数据样本在平衡化的同时可最大程度地保留原始样本的分布特性,降低多数类样本的信息流失率,从而保障分类器的分类性能.

1.1 边界区域检测

一个非均衡数据样本集的分布如图1所示:大圈内代表样本密集的区域,称为非边界区域;圈外样本分布稀疏的区域称为边界区域.



(a) 原始样本分布



(b) 合成后的新样本

图1 SMOTE算法合成新的少数类后的样本分布

传统的方法在对数据集进行分类前,先将边界区域数据作为噪声数据直接剔除,使得数据边缘变得光滑,然后再对数据进行分类.然而,由于非均衡数据中正负样本数量差异较大,如果边界区域数据中的少数类样本数量过多,则处理噪声时直接删除边界区域数据会导致少数类样本与多数类样本的数量差比例更大,增加数据集的非均衡程度. 对此,本文提出基于边界混合采样(BMS)的非均衡数据处理算法. 首先采用基于变异系数[15]的边界点检测方法识别非均衡数据的边界区域和非边界区域,基本思想如下.

1) 对于任意的正整数k和数据集D,p的k距离$k_\mathrm{dist}(p)$[15]定义为对象p到对象o的距离$\mathrm{dist}(p,o)$,且满足:

i) 至少有k个对象$q(q\in D$且$q\neq p)$使得

$$\mathrm{dist}(p,q)\leqslant \mathrm{dist}(p,o); \tag{1}$$

ii) 至多有$k-1$个对象$q(q\in D$且$q\neq p)$使得

$$\mathrm{dist}(p,q)< \mathrm{dist}(p,o). \tag{2}$$

2) 计算数据对象p到其k距离邻居距离之和的平均值

$$M_p=\sum \mathrm{dist}(p,m)\Big/N_{k_{\mathrm{dist}}(p)}. \tag{3}$$

其中:对象m为p的k距离邻居,$N_{k_{\mathrm{dist}}(p)}$为p的k距离邻居的个数.

3) 用平均值的倒数作为每个点的密度,有

$$\mathrm{Ld}(p)=\frac{1}{\sum \mathrm{dist}(p,m)\Big/N_{k_{\mathrm{dist}}(p)}}. \tag{4}$$

4) 将p的变异系数[15]定义为该对象与其k距离邻居密度的标准差与它们的平均值之比值,即

$$V_p=\frac{\sqrt{\dfrac{\sum\left(\mathrm{Ld}(p)-\dfrac{\sum\mathrm{Ld}(p)}{N_{k_{\mathrm{dist}}(p)}+1}\right)^2}{N_{k_{\mathrm{dist}}(p)}}}}{\dfrac{\sum\mathrm{Ld}(p)}{N_{k_{\mathrm{dist}}(p)}+1}}. \tag{5}$$

通过变异系数阈值寻找数据样本的边界域和非边界域.

1.2　边界区域样本处理

对于边界区域中的少数类样本,采用SMOTE算法进行合成,从而使得数据样本的边缘区域变得光滑,正负类样本均衡,在保留数据的原始信息的同时减少了噪声样本对分类器的分类性能造成的影响. 图1(b)展示了SMOTE算法合成新的少数类样本后的样本分布,其中大圆圈为采用SMOTE算法对边界区域的少数类样本进行合成后的新样本.

1.3　非边界区域数据的欠采样处理

样本中的非边界数据,即图1中圈内的点,是样本分布中的密集区域. 为了避免数据量大造成训练时间加长,同时为了尽可能地使数据样本分布均匀,本文采用基于欧氏距离的欠采样方法对非边界域中的多数类样本进行处理. 该方法首先找到所有非边界区域数据样本的中心点;然后计算每个样本点到中心点的距离并对其进行排序,根据欠采样倍率对样本进行删减. 经过处理后的数据样本集合尽可能地保留了样本原有的分布状态,同时保证非边界区域的正负类样本均衡.

1.4　基于边界混合采样的非均衡数据算法流程

基于边界的混合采样方法分别实现边界区域中的少数类样本和非边界区域中的多数类样本处理,在保持原有数据分布特征的基础上,实现数据集的均衡处理,提高非均衡数据分类的性能.

假设S为原始非平衡数据集,则S中的边界样本、非边界样本、边界样本中的少数类样本、边界样本中的多数类样本、非边界样本中的多数类样本、非边界类中的少数类分别定义为S_1、S_2、$S_1^{\min}$、S_1^{most}、S_2^{most}、$S_2^{\min}$,其相应的大小分别定义为N_{S_1}、N_{S_2}、$N_{S_1^{\min}}$、$N_{S_1^{\mathrm{most}}}$、$N_{S_2^{\mathrm{most}}}$、$N_{S_2^{\min}}$.

BMS算法的流程见图2. 该算法首先通过边界点检测算法计算样本集中每一个样本点的变异系数,设置变异系数阈值,将变异系数大于该阈值的样本划分为边界区域样本集,反之为非边界区域样本集;然后对于边界域中的少数类样本,采用SMOTE数据重构技术合成N倍采样率的新样本,加入训练集中;而对于边界域中的多数类样本不做处理,直接加入训练集中;最后,对于非边界中的多数类样本,采用OSED算法形成新的多数类样本子集并加入训练样本中.

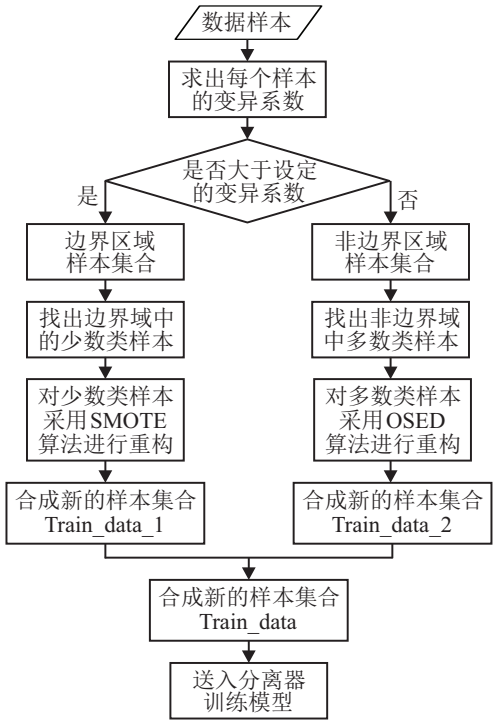


图 2　BMS算法流程图

首先,对算法中使用的两个常用倍率指标进行定义. 其中:过采样倍率N为SMOTE算法合成新的少数类样本的倍率,欠采样倍率M为基于欧氏距离的随机欠采样删除多数类样本的倍率. 具体表达式为

$$N=N_{S_1^{\mathrm{most}}}/N_{S_1^{\min}}, \tag{6}$$

$$M=N_{S_2^{\mathrm{most}}}/N_{S_2^{\min}}. \tag{7}$$

本文提出的边界混合采样的BMS算法的具体步骤描述如下.

输入:非平衡数据集$S(i=1,2,\cdots,T,\ T$为样本数量),过采样率N,欠采样率M,变异系数阈值V_a;

输出:分类模型H.

Step 1: for $i=1$ to $T, V_{x_i}=\mathrm{Variable\ Index}(x_i)$ //计算每个样本x_i的变异系数.

Step 1.1:算出样本x_i到每个样本的距离并加入Dist集合中,对Dist集合进行排序;

Step 1.2:找到距离样本x_i最近的k个样本,计算样本x_i分别到k个样本的距离之和并记为Distsum;

Step 1.3:求得每个样本的局部密度$\mathrm{Ld}(x_i)=\mathrm{Distsum}/k$;

Step 1.4:求得每个样本的变异系数,记为$V_{(x_i)}$.

Step 2:若$V_{(x_i)}>V_a$,则将x_i加入样本边界集合S_1;否则将x_i加入非边界样本集合S_2.

Step 3: for $i=1$ to N_{S_1}.

若x_i是多数类样本,则加入集合S_1^{most};否则加入集合$S_1^{\min}$.

Step 4:将S_1中的$S_1^{\min}$样本集合使用SMOTE方法进行复制$(N-1)$倍,得到复制后的样本$C_{S_1^{\min}}$,将此样本与$S_1^{\min}$、S_1^{most}合成新的样本集Train-data-1.

Step 5: 将S_2集合中的多数类样本采用OSED算法进行删减.

Step 5.1: 计算S_2集合中的样本均值M_{S_2};

Step 5.2: 计算S_2集合中每个样本点到M_{S_2}的距离并加入D集合中,对D集合中的距离进行排序;

Step 5.3: 设置计数器Counter $=$ 0,遍历排序后的D集合中每个样本点,每经过一个样本点,Counter加1,当且仅当Counter为M的整数倍时,将该样本点加入集合Train-data-2中.

Step 6: 将Train-data-1和Train-data-2加入Train-data中.

Step 7: 将Train-data进行训练,得出分类器模型H.

2 实验结果与分析

2.1 评价准则

在传统的分类方法中,一般采用分类精度作为评价指标,即正确分类的样本个数占总样本个数的百分比. 但是,分类精度指标未能考虑非均衡样本中少数类样本与多数类样本错分代价的差异,所以造成评价指标与实际情况不符. 因此,人们提出了采用混淆矩阵(见表1)结合F-value[18]及G-mean[19]实现非平衡数据分类性能的科学评价指标.

表 1 混淆矩阵

类别	被分为正类	被分为负类
实际为正类	T_P	F_N
实际为负类	F_P	T_N

表1中T_P是实际为正类且被分为正类的样本数量,F_N是实际为正类且被分为负类的样本数量,F_P是实际为负类且被分为正类的样本数量,T_N是实际为负类且被分为负类的样本数量.

1) F-value是一种不平衡数据分类问题的评价准则,主要针对正类的分类精度进行评价,其定义为

$$\text{F-value} = \frac{(1+\beta^2)\times R_C \times P_R}{\beta^2 \times R_C + P_R}, \tag{8}$$

$$R_C = T_P/(T_P + T_N), \tag{9}$$

$$P_R = T_P/(T_P + F_N). \tag{10}$$

其中: R_C为查全率; P_R为查准率; β表示R_C和P_R的相对重要性,在二分类问题中,β设置为1. 只有当R_C和P_R相对较大时,F-value才会相应较大,因此,F-value能够合理地评价分类器对于非平衡数据集的分类性能.

2) G-mean反映了对正类和负类样本分类能力的均衡程度,是一种衡量数据集整体分类性能的评价指标,其定义为

$$\text{G-means} = \sqrt{P_A \times N_A}. \tag{11}$$

其中:真正率P_A和真负率N_A分别定义为

$$P_A = R_C,\ N_A = T_N/(T_N + F_P). \tag{12}$$

本文采用以上两种指标来评价基于边界混合采样算法处理非均衡数据的分类准确率.

2.2 数据集描述

为了评价本文提出的BMS算法的有效性,选用国际机器学习标准数据库UCI中的7组具有非均衡性特征的数据集进行实验,数据特征信息见表2. 其中: haberman数据集的第2类作为少数类,第1类作为多数类; german数据集的B类作为少数类,A类作为多数类; pima数据集的第1类作为少数类,第0类作为多数类; abalone数据集中的第10类作为少数类,第9类作为多数类; ecoli数据集的第im类为少数类,其他合起来为多数类; glass数据集的第3、第5和第6类合起来为少数类,其他的为多数类; pageblocks数据集的第1类为少数类,其他的合起来为多数类.

表 2 实验数据集描述

数据集	维数	少数类/多数类	不平衡比例
haberman	4	126/225	1:1.78
german	25	300/700	1:2.33
pima	9	268/500	1:1.86
abalone	8	634/689	1:1.10
ecoli	8	77/259	1:3.36
glass	10	51/163	1:5.48
pageblocks	11	560/4 913	1:8.77

2.3 实验设计与结果分析

为了评估本文提出的基于边界混合采样算法在对非均衡数据进行分类时的性能,设计实验如下:

1) 为了观察不同分类器对不同非均衡数据集分类性能的影响,将表2中的7个非均衡数据集直接送入SVM和C4.5分类器进行分类,实验结果见表3.

表 3 7种数据集在SVM和C4.5上的对比结果

dataset	method	F-value	G-mean
haberman	SVM	**0.605**	**0.618**
	C4.5	0.375	0.533
german	SVM	**0.584**	**0.641**
	C4.5	0.451	0.487
pima	SVM	**0.626**	**0.682**
	C4.5	0.514	0.498
abalone	SVM	**0.545**	**0.513**
	C4.5	0.361	0.425
ecoli	SVM	**0.642**	**0.708**
	C4.5	0.541	0.647
glass	SVM	**0.764**	**0.828**
	C4.5	0.654	0.719
pageblocks	SVM	**0.778**	**0.641**
	C4.5	0.541	0.452

2) 选取上述较优的分类器SVM进行实验,将本文算法与基于SMOTE过采样的SVM算法(SMOTE-SVM)、基于随机欠采样与SMOTE相结合的SVM算法(RU-SMOTE-SVM)和基于样本特性欠采样的不均衡支持向量机算法(SPU-SVM)进行对比实验. 对

于每一个数据集,采用SVM分类器对均衡化后的数据进行训练,为防止随机影响,采用十折交叉验证法对每一组数据集进行实验,最后计算这些实验的F-value和G-mean性能评测指标的统计平均值,对比结果如表4所示.

表 4　非均衡数据集F-value和G-mean性能比较

dataset	method	F-value	G-mean
haberman	Smote-SVM	0.597	0.588
	RU-Smote-SVM	0.661	0.599
	SPU-SVM	0.599	0.612
	BMS-SVM	**0.684**	**0.626**
german	Smote-SVM	0.626	0.629
	RU-Smote-SVM	0.531	0.555
	SPU-SVM	**0.719**	0.679
	BMS-SVM	0.685	**0.731**
pima	Smote-SVM	0.789	0.725
	RU-Smote-SVM	0.738	0.687
	SPU-SVM	0.833	0.733
	BMS-SVM	**0.889**	**0.795**
abalone	Smote-SVM	0.753	0.623
	RU-Smote-SVM	0.716	0.595
	SPU-SVM	0.740	0.630
	BMS-SVM	**0.782**	**0.681**
ecoli	Smote-SVM	0.654	0.615
	RU-Smote-SVM	0.712	0.687
	SPU-SVM	0.730	0.788
	BMS-SVM	**0.871**	**0.891**
glass	Smote-SVM	0.758	0.894
	RU-Smote-SVM	0.764	0.743
	SPU-SVM	0.814	0.867
	BMS-SVM	**0.831**	**0.922**
pageblocks	Smote-SVM	0.657	0.725
	RU-Smote-SVM	**0.784**	0.758
	SPU-SVM	0.705	0.799
	BMS-SVM	0.735	**0.824**

由表3可以看出,对于每一个数据集,SVM分类器对于少数类样本的识别率F-value和数据整体的识别率G-mean均高于C4.5. 因此相对于C4.5而言,SVM分类器更适合非均衡数据的分类.

7个数据集的F-value和G-mean性能指标的对比结果如表4所示. 对于每一个数据集,分别比较了4种不同的非均衡数据处理后的分类性能,并将最大F-value和G-mean值用黑体表示. 同时为了更加直观地观察不同方法在不同数据集上的F-value和G-mean值的对比情况,将表4中的对比数据画成柱状图如图3、图4所示.

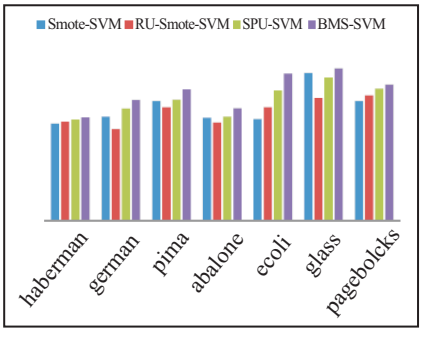


图 3　4种方法在7个UCI数据集的G-mean对比

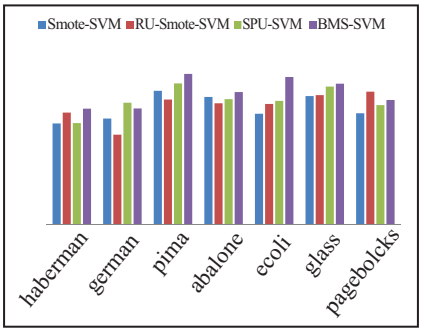


图 4　4种方法在7个UCI数据集的F-value对比

由图3可以看出,本文算法在7个数据集分类上的G-mean值均优于其他非均衡数据SVM分类算法. 这是由于本文算法对边界域和非边界域中的多数类和少数类样本分别进行处理,最大程度地降低了边界样本点对于分类性能的影响,同时对于非边界域中的多数类样本,并没有采用传统的随机欠采样技术随机删除某些多数类样本,而是采用基于欧氏距离的随机欠采样方法进行删除,避免了有价值的多数类样本的信息丢失,保留了原始数据集的信息,因此,本文算法处理后的非均衡数据集上的分类,在整体样本的识别上优于其他算法.

观察F-value指标,如图4所示,对于数据集german,SPU-SVM的F-value值高于本文算法,这是因为german数据集的多数类样本分布相对比较集中,有很大一部分靠近分类界面,而SPU-SVM算法通过样本密度信息选择最具代表性的多数类样本点,因此german数据集在SPU-SVM算法上更有效. 同时,pageblocks数据集具有较高的不平衡比例和较大的样本个数,本文算法中的变异系数阈值这一关键参数对数据集的分类性能有所影响,导致RU-Smote-SVM算法的F-value高于本文算法. 但从整体上看,SPU-SVM方法是根据样本密度对多数类样本进行的删减,虽然在减少多数类的同时使得分类界面向多数类方向偏移,但会造成部分多数类样本的信息丢失;而BMS-SVM将过采样和欠采样方法进行融合,保留边界域的样本信息,同时选择性地删减分布密集的非边界样本,因此,BMS-SVM分类性能从整体上优于SPU-SVM算法,同时也优于其他算法,是较为有效和稳定的.

3　结　论

本文针对非均衡数据分类问题,提出了一种基于边界混合采样方法BMS来均衡化数据集. 该算法首先计算样本集中每一个样本点的变异系数,通过变异系数界定边界域和非边界域;然后,对于边界域中的少数类样本,采用SMOTE技术合成新的样本,对于非边界域中的多数类样本,采用基于欧氏距离的随机欠采样形成新的多数类样本子集;最后将均衡化后的

样本集送入SVM分类器中进行训练,并与其他非均衡数据分类算法进行比较. 实验结果表明,本文算法在不同数据集上的分类性能均优于其他算法. 通过对比SVM和决策树分类器,表明SVM分类器更适合非均衡数据的研究. 未来的研究拟在本文工作的基础上对分类算法进行改进,以达到更好的少数类识别效果.

参考文献(References)

[1] Gu Q, Cai Z, Zhu L, et al. Data mining on imbalanced data sets[C]. Int Conf on Advanced Computer Theory and Engineering. Phuket: IEEE, 2008: 1020-1024.

[2] Xiao Y, Wang H, Zhang L, et al. Two methods of selecting Gaussian kernel parameters for one-class SVM and their application to fault detection[J]. Knowledge-Based Systems, 2014, 59(2): 75-84.

[3] Miao Z, Zhao L, Yuan W, et al. Multi-class imbalanced learning implemented in network intrusion detection[C]. Int Conf on Computer Science and Service System. Nanjing: IEEE, 2011: 1395-1398.

[4] Liu Y Q, Wang C, Zhang L. Decision tree based predictive models for breast cancer survivability on imbalanced data[C]. The 3rd Int Conf on Bioinformatics and Biomedical Engineering. Beijing, 2009: 1-4.

[5] 蔡艳艳, 宋晓东. 针对非平衡数据分类的新型模糊SVM模型[J]. 西安电子科技大学学报:自然科学版, 2015, 42(5): 120-124.
(Cai Y Y, Song X D. A new fuzzy SVM model for imbalanced data classification[J]. J of Xi'an Electronic and Science University: Natural Science Edition, 2015, 42 (5): 120-124.)

[6] Chen X, Song E, Ma G. An adaptive cost-sensitive classifier[C]. The Int Conf on Computer and Automation Engineering. Singapore: IEEE, 2010: 699-701.

[7] 张银峰, 郭华平, 职为梅,等. 一种面向不平衡数据分类的组合剪枝方法[J]. 计算机工程, 2014, 40(6): 157-161.
(ZhangY F, Guo H P, Zhi W M, et al. A classification method of combination pruning for unbalanced data[J]. Computer Engineering, 2014, 40(6): 157-161.)

[8] García V, Sánchez J S, Mollineda R A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance[J]. Knowledge-Based Systems, 2012, 25(1): 13-21.

[9] 程险峰, 李军, 李雄飞. 一种基于欠采样的不平衡数据分类算法[J]. 计算机工程, 2011, 37(13): 147-149.
(Cheng X F, Li J, Li X F. An imbalanced data classification algorithm based on under sampling[J]. Computer Engineering, 2011, 37(13): 147-149.)

[10] 陶新民, 张冬雪, 郝思媛,等. 基于谱聚类欠取样的不均衡数据SVM分类算法[J]. 控制与决策, 2012, 27(12): 1761-1768.
(Tao X M, Zhang D X, Hao S Y, et al. SVM classification algorithm for imbalanced data based on spectral clustering under-sampling[J]. Control and Decision, 2012, 27(12): 1761-1768.)

[11] García S, Herrera F. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy[J]. Evolutionary Computation, 2009, 17(3): 275-306.

[12] 刘余霞, 刘三民, 刘涛,等. 一种新的过采样算法DB_SMOTE[J]. 计算机工程与应用, 2014, 50(6): 92-95.
(Liu Y X, Liu S M, Liu T, et al. A new algorithm of over- sampling DB_SMOTE[J]. Computer Engineering and Application, 2014, 50 (6): 92-95.)

[13] 古平, 欧阳源遊. 基于混合采样的非平衡数据集分类研究[D]. 重庆: 重庆大学计算机学院, 2014.
(Gu P, Ouyang Y Y. Imbalanced data classification researchbased on hybrid sampling[D]. Chongqing: College of Computer Science, Chongqing University, 2014.)

[14] 谷琼, 袁磊, 宁彬,等. 一种基于混合重取样策略的非均衡数据集分类算法[J], 计算机工程与科学, 2012, 34(10): 128-134.
(Gu Q, Yuan L, Ning B, et al. An imbalanced data set classification algorithm based on mixed resampling strategy[J]. Computer Engineering and Science, 2012, 34(10): 128-134.)

[15] 薛丽香, 邱保志. 基于变异系数的边界点检测算法[J]. 模式识别与人工智能, 2009, 22(5): 799-802.
(Xue L X, Qiu B Z. A method of boundary point detection based on coefficient of variation[J]. Pattern Recognition and Artificial Intelligence, 2009, 22 (5): 799-802.)

[16] 赵自翔, 王广亮, 李晓东. 基于支持向量机的不平衡数据分类的改进欠采样方法[J]. 中山大学学报:自然科学版, 2012, 51(6): 10-16.
(Zhao Z X, Wang G L, Li X D. An improved under-sampling method based on support vector machine for imbalanced data classification[J]. J of Sun Yat-sen University: Natural Science Edition, 2012, 51(6): 10-16.)

[17] 陶新民, 郝思媛, 张冬雪,等. 基于样本特性欠取样的不均衡支持向量机[J]. 控制与决策, 2013, 28(7): 978-984.
(Tao X M, Hao S Y, Zhang D X, et al. Imbalanced support vector machines based on sample characteristics under-sampling[J]. Control and Decision, 2013, 28(7): 978-984.)

[18] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]. Int Conf on Advances in Intelligent Computing. Hefei: Springer-Verlag, 2005: 878-887.

[19] Sophia Daskalaki, Ioannis Kopanas, Nikolaos Avouris. Evaluation of classifiers for an uneven class distribution problem[J]. Applied Artificial Intelligence, 2006, 20(5): 381-417.

(责任编辑: 李君玲)