一种基于样本分层的双向过采样方法

周晓敏 曹付元 余丽琴
(山西大学计算机与信息技术学院 太原 030006)
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 重采样技术由于简单、直观,逐渐成为解决非平衡数据分类问题的一个重要方向。但是在数据集很小的情况下,重采样技术中的欠采样可能会丢失数据集的重要信息,因此过采样是非平衡数据分类问题的研究重点。现有的过采样方法虽然有效地解决了类间不平衡问题,但是有可能造成少数类的密集区域更加密集,甚至引起样本重叠。此外,由于少数类样本可能存在噪音,现有的过采样方法可能会在噪音周围生成新样本,从而造成少数类样本的分布更加混乱。针对这些问题,文中提出了一种基于样本分层的双向过采样方法,该方法首先基于最高密度点和类内平均距离将少数类样本划分成密集层和稀疏层,然后对密集层边界区样本和稀疏层的样本进行双向过采样。为了验证所提算法的有效性,在 9 个 UCI 数据集上将提出的算法和其他过采样算法进行了比较。实验结果和 Friedman 等检验结果显示,提出的算法在处理非平衡数据分类问题时具有一定优势。

关键词 非平衡数据,分类,双向过采样,密集层,稀疏层
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/jsjkx.190400053

Bi-directional Oversampling Method Based on Sample Stratification

ZHOU Xiao-min CAO Fu-yuan YU Li-qin
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)
(Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan 030006, China)

Abstract Resampling technology has gradually become an important direction to solve the problem of classification for imbalanced data because of its simplicity and intuition. However, in the case of small data sets, under-sampling in resampling technology may lose important information of data sets, so oversampling is the focus of classification for imbalanced data. Although the existing oversampling methods effectively overcome the imbalance between classes, they may cause dense areas of minority class to be denser, even lead to overlapping of samples. In addition, due to the noise of minority class, the existing oversampling methods may generate new samples around the noise, which makes the distribution of minority class more confusing. Aiming at these problems, this paper proposed a bi-directional oversampling method based on sample stratification. It firstly divides the minority samples into dense area and sparse area based on the highest density point and the intra-class average distance. And then the bi-directional oversampling is performed in the boundary region of dense area and the sparse area. In order to verify the effectiveness of the proposed algorithm, comprehensive experiments were conducted on 9 data sets of UCI database. The experimental results and Friedman test results show the superiority of the proposed algorithm for the task of imbalanced data classification.

Keywords Imbalanced data, Classification, Bi-directional oversampling, Dense area, Sparse area

1 引言

非平衡数据广泛存在于各种领域,如医疗诊断、雷达图像检测、诈骗检测、电信设备故障预测[1]、文本分类[2]、金融交易[3]等。在这些领域中,少数类往往是更具有价值的一类,将少数类样本判别成多数类样本的代价远高于将多数类样本判别成少数类样本的代价[4]。例如,在癌症诊断中,患者远远少于正常人,将癌症患者误诊为正常人的代价比将正常人误诊

到稿日期:2019-04-09 返修日期:2019-08-09 本文受国家自然科学基金项目(61573229),山西省重点研发计划项目(201803D31022),山西省留学基金项目(2016-003),山西省留学基金择优资助项目(2016-001)资助。
周晓敏(1995-),女,硕士生,主要研究方向为非平衡数据分类学习,E-mail:2641401859@qq.com;**曹付元**(1974-),男,博士,教授,博士生导师,CCF 会员,主要研究方向为数据挖掘与机器学习,E-mail:cfy@sxu.edu.cn(通信作者);**余丽琴**(1992-),女,博士生,主要研究方向为数据挖掘与机器学习。

为癌症患者的代价高得多。又如,在电信领域中,欺诈电话远少于正常电话,但是将一个欺诈电话预测为正常电话比将正常电话预测为欺诈电话造成的经济损失大得多[5]。但是,在分类过程中,由于少数类样本数量很少,其所能表达的信息受到了限制,因而在分类时很难正确分析出数据的分布及其内部规律,导致少数类的分类精度下降[6]。因此,针对非平衡数据的分类问题成为目前的研究热点。

目前,对非平衡数据的处理方法大致分为两种,即数据层面的方法和算法层面的方法[7]。算法层面主要是代价敏感学习,由于分类不平衡问题中正确识别少数类比正确识别多数类更有价值,即错分少数类比错分多数类要付出更大的代价[8],因此代价敏感学习为少数类赋予更高的错分类代价[9]。算法层面还有集成学习,通过聚集多个模型的预测结果来提高分类性能[8]。数据层面主要是重采样技术,包括欠采样、过采样以及混合采样。欠采样是指在多数类样本中剔除一些样本,使之减少到少数类的样本数量;过采样是指在少数类样本中加入新的样本,使其和多数类的样本数量达到平衡[10];混合采样是过采样和欠采样的结合。

由于重采样技术简单、直观,而且重采样主要是改变数据集本身,而不是改变分类器[11],因此大多数研究者在解决非平衡数据的分类问题时都是使用过采样或者欠采样,但是欠采样可能会丢失数据集的重要信息,特别是当数据集很小时[12],因此过采样技术是非平衡问题的研究重点。2002 年,Chawla 提出的 SMOTE[13](Synthetic Minority Over-sampling Technique)算法是过采样的一种经典算法,现在很多算法都是在 SMOTE 算法的基础上改进的。SMOTE 的基本思想是先找到少数类中每个样本的 k 个近邻,然后随机选择其中的一个近邻样本,在给定样本和近邻样本之间的连线上合成新样本,使少数类样本数量增多,最后再利用传统分类器对其进行分类,有效地提高了少数类样本的分类精度。但是 SMOTE 算法对所有少数类样本不加区别地进行过采样,有可能造成少数类的密集区域内样本更加密集,甚至引起样本重叠,并且在有噪音的情况下有可能在多数类样本区域合成少数类样本,这些情况都会使少数类的分类精度下降。文献[14]提出的 Borderline-SMOTE 算法改变了传统的思路,认为边界样本具有更重要的信息,只对靠近边界的少数类样本进行过采样。Borderline-SMOTE 算法虽然加强了边界,在一定程度上有利于分类,但是风险太大,有可能大部分合成样本被判别为多数类。文献[15]提出的一种基于 k-means 和 SMOTE 的启发式过采样方法(Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE)通过对全体数据进行聚类,然后对选择出来的类进行 SMOTE 采样,从而有效解决了类间不平衡和类内不平衡问题。但是,基于 k-means 和 SMOTE 的启发式算法对聚类个数的选择、初始类中心比较敏感。

针对以上问题,本文提出了一种基于样本分层的双向过采样方法,该方法首先基于最高密度点和类内平均距离将少数类样本划分成密集层和稀疏层,然后找到密集层的边界样本,最后对这些边界样本和稀疏层的样本进行双向过采样。由于 LibSVM 简单、易于使用,本文使用 LibSVM 作为基准分类器,选择 F-measure 和 G-mean 作为评价指标,在 9 个 UCI 数据集上进行测试。实验结果表明,本文提出的算法可以对非平衡数据进行有效的分类,算法的整体性能较好。

2 一种基于样本分层的双向过采样方法

2.1 算法思想

传统的 SMOTE 算法对所有少数类样本不加区别地进行过采样,有可能造成少数类的密集区域内样本更加密集,甚至引起样本重叠,而稀疏区域依然比较稀疏;此外,在有噪音的情况下,SMOTE 算法会在多数类样本区域合成少数类样本,从而使少数类样本的分布更加混乱。这些情况都会造成分类器的整体性能不佳,不仅使少数类样本的分类精度不高,甚至使多数类样本的分类精度有所下降,如图 1 所示。

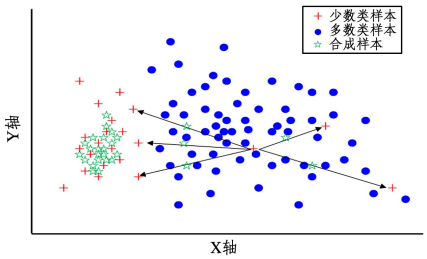


图 1 SMOTE 算法存在的问题

Fig. 1 Problems in SMOTE algorithm

针对图 1 的问题,本文提出了一种基于样本分层的双向过采样方法,简记为 SBO(Stratified Bi-directional Oversampling)。首先基于最高密度点和类内平均距离将少数类样本划分成密集层和稀疏层,然后找到密集层的边界样本,最后对这些边界样本和稀疏层的样本进行双向过采样,如图 2 所示。该算法有效地减小了样本重叠以及噪音的影响,使合成的新样本更加有利于分类。

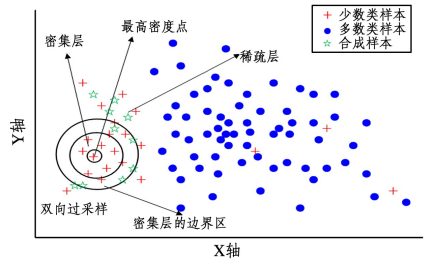


图 2 本文提出的 SBO 算法示意图

Fig. 2 Illustration of proposed SBO algorithm

2.2 样本分层

令训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,$x_i\in R^d$,$y_i\in\{-1,1\}(i=1,2,\cdots,n)$,$T$ 为包含 n 个样本点的二分类数据集,$y_i=1$ 表示对应的样本点属于少数类,记为 $P=\{(x_1,1),(x_2,1),\cdots,(x_p,1)\}$,$y_i=-1$ 表示对应的样本点属于多数类,记为 $M=\{(x_1,-1),(x_2,-1),\cdots,(x_m,-1)\}$,$T=P\cup M$。

为方便计算,首先给出两个样本的欧氏距离的定义。对于 $\forall x_i,x_j\in P$,属性个数为 d,x_i,x_j 之间的欧氏距离

$D(x_i,x_j)$的定义如下：

$$D(x_i,x_j)=\sqrt{\sum_{k=1}^{d}(x_{ik}-x_{jk})^2} \tag{1}$$

其中，$\forall x_i\in P(i\in\{1,2,\cdots,p\})$，$x_i$ 密度可以定义为：

$$Density(x_i)=-\sum_{j=1}^{p}D(x_i,x_j)/p \tag{2}$$

样本密度越大，说明样本周围的点越多，进一步可以获得少数类样本中的最高密度点，记为 $x_{max\_den}=\arg\max_{x_i\in P}(Density(x_i))$。

定义 1(类内平均距离)　给定 $P=\{(x_1,1),(x_2,1),\cdots,(x_p,1)\}$，则少数类内平均距离可以定义为 $dist\_aver$：

$$dist\_aver=\frac{\sum_{i=1}^{p}\sum_{j=1}^{p}D(x_i,x_j)}{p\times p} \tag{3}$$

给定 $P=\{(x_1,1),(x_2,1),\cdots,(x_p,1)\}$，以最高密度点 x_{max_den} 为中心、类内平均距离 $dist\_aver$ 为半径，包含的所有少数类样本点构成少数类密集层，记为 Den_area，密集层以外的少数类样本构成稀疏层，记为 $Spar\_area$，定义如下：

$$Den\_area=\{x_j\mid D(x_j,x_{max\_den})\leqslant dist\_aver,x_j\in P\} \tag{4}$$

$$Spar\_area=\{x_j\mid D(x_j,x_{max\_den})>dist\_aver,x_j\in P\} \tag{5}$$

定义 2(密集层的边界区)　$\forall x_i\in Den\_area$，$\forall x_j\in Spar\_area$，

$$border=\max_{1\leqslant i\leqslant|Den\_area|}(D(x_i,x_{max\_den}))-\min_{\substack{1\leqslant i\leqslant|Den\_area|\\1\leqslant j\leqslant|Spar\_area|}}(D(x_i,x_j)) \tag{6}$$

密集层的边界区可以定义为：

$$bor\_area=\{x_k\mid border\leqslant D(x_k,x_{max\_den})\leqslant\max_{1\leqslant i\leqslant|Den\_area|}(D(x_i,x_{max\_den})),x_k\in Den\_area\} \tag{7}$$

2.3　双向过采样

由于数据集中少数类样本可能存在噪音及离群点，传统的单向随机选择合成样本可能会在噪音周围生成新样本，造成少数类样本的分布更加混乱。本文采用双向过采样在互为近邻的少数类样本之间合成新样本，避免了噪音以及离群点的影响。

定义 3　$\forall x_i,x_j\in Spar\_area\cup bor\_area$，$N_k(x_i)$表示 x_i 的 k 个近邻，则进行双向过采样的样本对的集合 Bi_Sample 可以定义为：

$$Bi\_Sample=\{(x_i,x_j)\mid x_i\in N_k(x_j)\wedge x_j\in N_k(x_i)\} \tag{8}$$

例 1　假设数据集中有 10 个少数类样本($x_1\sim x_{10}$)，对这 10 个少数类样本进行双向过采样，利用欧氏距离找 k 近邻，这里取 $k=2$，如图 3 及表 1 所示。

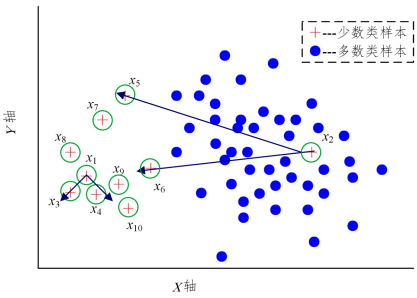


图 3　双向过采样

Fig. 3　Bi-directional oversampling

表 1　$k=2$ 时少数类样本的 k 近邻

Table 1　k-nearest neighbors of minoritg samples when $k=2$

少数类样本	近邻($k=2$)	
x_1	x_3	x_4
x_2	x_5	x_6
x_3	x_1	x_4
x_4	x_1	x_3
x_5	x_7	x_8
x_6	x_9	x_{10}
x_7	x_5	x_8
x_8	x_1	x_7
x_9	x_4	x_{10}
x_{10}	x_6	x_9

由图 3 可知，对于少数类样本 x_1，它的 2 近邻是 x_3 和 x_4，同时 x_3 和 x_4 的 2 近邻也有 x_1，因此 $Bi\_Sample=\{(x_1,x_3),(x_1,x_4)\}$。

对于噪音样本 x_2，它的 2 近邻是 x_5 和 x_6，但是 x_5 和 x_6 的 2 近邻没有 x_2，因此在 x_2 和 x_5、x_2 和 x_6 之间不生成样本，这样就避免了生成不必要的噪音。

2.4　算法流程

首先计算多数类样本和少数类样本数量的差值 N，N 就是要合成的少数类样本数量，合成的新样本占原始少数类样本的比例记为采样比例 $radio=N/P$。SBO 算法如算法 1 所示。

算法 1　SBO(Stratified Bi-directional Oversampling)

输入：训练集 T 中的少数类样本集合 P，采样比例 radio，近邻个数 Q

输出：合成的少数类样本 Sample

Step 1　根据式(2)计算 P 中每个样本的密度，得到最高密度点 x_{max_den}；

Step 2　根据式(3)得到类内平均距离 dist_aver；

Step 3　根据式(4)和式(5)得到 P 的密集层 Den_area 和稀疏层 Spar_area；

Step 4　根据式(7)得到密集层 Den_area 的边界区 bor_area；

Step 5　对 bor_area 和 Spar_area 中的样本点进行双向过采样；

Step 6　输出合成的少数类样本 Sample。

2.5　时间复杂度分析

本文提出的 SBO 算法主要包括样本分层和双向过采样两部分。计算少数类样本的最高密度点和类内平均距离的时间复杂度为 $O(d*p^2)$；计算少数类的密集层和稀疏层的时间复杂度为 $O(p)$；计算少数类密集层边界区的时间复杂度为 $O(d*(|Den\_area|+|Den\_area|*|Spar\_area|)+|Den\_area|)$；进行双向过采样的时间复杂度为 $O(d*|Den\_area|*|Spar\_area|+d*|Bi\_sample|)$。

综上，SBO 算法的时间复杂度为 $O(d*p^2)$。由于非平衡数据中少数类样本数量 p 相对于多数类样本较少，因此提出的算法可以应用到大数据集中。

3　实验与结果分析

3.1　评价指标

非平衡数据的二分类问题往往使用混淆矩阵来评估，如表 2 所列，其中列表示类的预测结果，行表示类的实际类别[16]。表 2 中，TN 表示多数类样本中被分类正确的样本数；TP 表示少数类样本中被分类正确的样本数；FN 表示多数类

样本中被预测为少数类的样本数;FP 表示少数类样本中被预测为多数类的样本数。

表 2 二分类问题中的混淆矩阵

Table 2 Confusion matrix of two class problem

分类	预测为多数类	预测为少数类
实际为多数类	TN	FN
实际为少数类	FP	TP

准确率 Precision 和召回率 Recall 是分类中最基本的两个指标[17]。准确率 Precision 也称为查准率,召回率 Recall 也称为查全率[18]。其定义如下:

$$Precision=\frac{TP}{TP+FP} \tag{9}$$

$$Recall=\frac{TP}{TP+FN} \tag{10}$$

F-measure 是准确率 Precision 和召回率 Recall 的调和平均,F-measure 的值越高,分类器性能就越好,其定义为:

$$F\text{-}measure=\frac{(1+\alpha^2)\times Recall\times Precision}{\alpha^2\times Recall+Precision} \tag{11}$$

一般令 α=1,因此 F-measure 定义为:

$$F\text{-}measure=\frac{2\times Recall\times Precision}{Recall+Precision} \tag{12}$$

本文中的另一个指标是 G-mean 值,用来衡量分类器对于两类样本分类的平均性能[19],是对算法性能的总体评价,G-mean 的值越高越好,其定义为:

$$G\text{-}mean=\sqrt{Recall\times\frac{TN}{TN+FP}} \tag{13}$$

3.2 数据集

为了测试本文提出的 SBO 算法的有效性,从 UCI 数据库中选择 9 个非平衡度不同的数据集,其中非平衡度小于 9∶1 的有 5 个数据集,大于 9∶1 的有 4 个数据集,详细信息如表 3 所列。

表 3 数据集的基本信息

Table 3 Basic information of data sets

数据集	属性	样本数量	多数类样本数量	少数类样本数量	非平衡度
yeast3	9	1484	1321	163	8.1∶1
ecoli1	8	336	259	77	3.4∶1
glass0	10	214	144	70	2.1∶1
new_thyroid2	6	215	180	35	5.1∶1
page_blocks0	11	5472	4913	559	8.8∶1
yeast-2_vs_4	9	514	463	51	9.1∶1
yeast05679vs4	9	528	477	51	9.4∶1
ecoli4	8	336	316	20	15.8∶1
glass4	10	214	201	13	15.5∶1

3.3 实验结果与分析

本文各种算法的编写与编译是在 Matlab R2016a 中实现的,实验环境为 Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz 3.60GHz,8GB 内存,64 位操作系统,Windows 7 专业版。

实验首先对每个数据集进行归一化处理,然后对其进行 20 次 5 折交叉验证(共 100 次实验),最后取所有结果的平均值和标准差。本文将所提算法与经典的 SMOTE[13] 算法、Borderline-SMOTE[14]算法以及一种基于 k-means 和 SMOTE 的启发式过采样方法[15]进行了比较,使用 LibSVM 作为基准分类器来进行测试,设近邻个数 $Q=5$,对于 kmeans-SMOTE 算法取聚类个数 $K=3$。

实验结果如表 4 和表 5 所列,"±"左边的数值代表平均值,"±"右边的数值代表标准差,标准差后面括号中的数字表示这 4 种算法的性能高低排序,"1"表示算法的性能最好,"2"表示算法的性能次好,以此类推,最后一行表示每种算法在 9 个数据集上的平均序值。

表 4 4 种算法在 9 个数据集上的 F-measure 比较结果

Table 4 Comparison results of F-measure of 4 algorithms on 9 data sets

数据集	SMOTE	Borderline-SMOTE	*k-m*eans-SMOTE	SBO
yeast3	0.7272±0.0186(3)	0.6616±0.0118(4)	0.7443±0.0199(2)	**0.7580±0.0026(1)**
ecoli1	0.8088±0.0302(2)	0.7687±0.0142(4)	0.8083±0.0269(3)	**0.8249±0.0066(1)**
glass0	0.7154±0.0208(3)	0.7129±0.0171(4)	0.7172±0.0246(2)	**0.7248±0.0079(1)**
new_thyroid2	0.9624±0.0224(2)	0.9476±0.0041(4)	0.9552±0.0285(3)	**0.9659±0.0067(1)**
page_blocks0	0.7812±0.0075(3)	0.7036±0.0058(4)	0.8020±0.0060(2)	**0.8025±0.0014(1)**
yeast-2_vs_4	0.7326±0.0395(2)	0.6895±0.0175(4)	0.7137±0.0316(3)	**0.7330±0.0074(1)**
yeast05679vs4	0.5126±0.0433(2)	0.4512±0.0232(4)	0.4641±0.0386(3)	**0.5450±0.0024(1)**
ecoli4	0.7829±0.0386(3)	0.7844±0.0033(2)	0.7772±0.0328(4)	**0.8399±0.0139(1)**
glass4	0.7795±0.0525(4)	0.7844±0.0216(3)	0.7937±0.0626(2)	**0.7961±0.0201(1)**
平均序值	2.6667	3.6667	2.6667	**1**

表 5 4 种算法在 9 个数据集上的 G-mean 比较结果

Table 5 Comparison results of G-mean of 4 algorithms on 9 data sets

数据集	SMOTE	Borderline-SMOTE	*k*means-SMOTE	SBO
yeast3	0.7882±0.0139(3)	0.7263±0.0089(4)	0.8212±0.0272(2)	**0.8674±0.0031(1)**
ecoli1	0.8541±0.0227(3)	0.7983±0.0105(4)	0.8680±0.0236(2)	**0.8975±0.0024(1)**
glass0	0.7566±0.0193(3)	0.7465±0.0156(4)	0.7579±0.0213(2)	**0.7766±0.0065(1)**
new_thyroid2	0.9661±0.0167(2)	0.9620±0.0039(3)	0.9582±0.0219(4)	**0.9792±0.0065(1)**
page_blocks0	0.8183±0.0065(3)	0.7458±0.0043(4)	0.9269±0.0077(2)	**0.9363±0.0008(1)**
yeast-2_vs_4	0.8076±0.0305(2)	0.7751±0.0152(4)	0.8013±0.0287(3)	**0.8403±0.0046(1)**
yeast05679vs4	0.6343±0.0334(2)	0.5767±0.0188(4)	0.5825±0.0307(3)	**0.8292±0.0056(1)**
ecoli4	0.8829±0.0376(3)	0.8895±0.0032(2)	0.8772±0.0376(4)	**0.9568±0.0005(1)**
glass4	0.8310±0.0415(4)	0.8620±0.0261(2)	0.8415±0.0557(3)	**0.8625±0.0197(1)**
平均序值	2.7778	3.4444	2.7778	**1**

从表 4 可以看出，无论数据集的非平衡度是否大于 9∶1，SBO 算法的 *F-measure* 值都要高于其他算法。表 5 列出了 4 种算法在 9 个数据集上的 *G-mean* 值，*G-mean* 同时考虑两个类别的性能，少数类分类正确率和多数类分类正确率需要同时高。从表 5 可以看出，SBO 算法的 *G-mean* 值明显高于其他算法，尤其对于非平衡度大于 9∶1 的数据集。*F-measure* 和 *G-mean* 值的提升证明了 SBO 算法的有效性。因为 SBO 算法不是对所有的少数类样本进行过采样，而是对少数类先进行样本分层，再对密集层边界区样本和稀疏层样本进行双向过采样，这样可以减少样本的重叠以及噪声和离群点的影响，所以 F-measure 和 G-mean 值得到了提升。

为了对算法进行更全面的比较，本文使用 Friedman 检验和 Bonferroni-Dunn 检验来比较这 4 种算法的性能差异。假设 A 代表算法的个数，B 代表数据集的个数，C 代表评价指标的个数，则第 j 种算法的平均序值定义为 $R_j=(\sum_{i=1}^{C} r_i^j)/C$，其中 r_i^j 代表第 i 种评价指标下第 j 种算法的序值。显然，本文中 $A=4$，$B=9$，$C=2$，4 种算法的 R 值如表 6 所列。

表 6　4 种算法的 R 值

Table 6　R values of four algorithms

	SMOTE	Borderline-SMOTE	k-means-SMOTE	SBO
R	2.7223	3.5556	2.7223	1

假设所有算法没有差异，则它们的排序应该是相等的，即本文 4 种算法的排序均为 2.5。Friedman 检验用来测试比较的算法是否满足该假设，并且在该假设下服从自由度为 3 的 χ_F^2 分布，Friedman 检验如下：

$$\chi_F^2=\frac{12B}{A(A+1)}\left[\sum_{j=1}^{A}R_j^2-\frac{A(A+1)^2}{4}\right]$$

$$=\frac{12\times 9}{4\times(4+1)}[2.722\,3^2+3.555\,6^2+2.722\,3^2+1^2-\frac{4\times(4+1)^2}{4}]$$

$$\approx 18.7$$

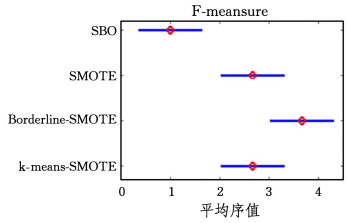
显然，$18.7>\chi_{0.1}^2(4-1)=6.251$，因此不满足该假设，则说明比较的算法具有显著性差异。

进一步，使用 Bonferroni-Dunn 检验来区分各算法性能的差异。根据文献[20]，当 $\alpha=0.1$ 时 q_α 值为 2.128，计算出平均序值差别的临界值域如下：

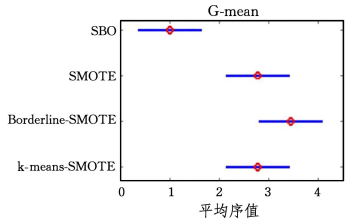
$$CD=q_\alpha\sqrt{\frac{A(A+1)}{6B}}=2.128\cdot\sqrt{\frac{4\times(4+1)}{6\times 9}}\approx 1.295$$

其中，B 为数据集个数，A 为算法个数。如果两种算法平均序值的差值大于临界值域 CD，就说明这两种算法的性能显著不同。图 4 显示了 4 种算法在 9 个数据集上的 Bonferroni-Dunn 检验，其中纵轴表示各种算法，横轴是平均序值，圆圈表示每种算法的平均序值，以圆圈为中心的横线段表示临界值域的大小。若两种算法的横线段没有重叠，则说明两种算法有显著差别。

从图 4 可以看到，SBO 算法和其他 3 种算法没有重叠区域，说明 SBO 算法和其他 3 种算法有显著差别。但是，SMOTE 算法和 k-means-SMOTE 几乎重叠，SMOTE、k-means-SMOTE和 Borderline-SMOTE 算法存在很大的重叠区域，说明这 3 种算法没有显著差别。由于 SBO 算法的平均序值高，并且与其他算法有显著差别，因此 SBO 算法的性能优于其他算法。



(a)F-measure



(b)G-mean

图 4　Bonferroni-Dunn 检验

Fig. 4　Bonferroni-Dunn test

表 7 为 4 种算法在 9 个数据集上的运行时间比较。

表 7　4 种算法在 9 个数据集上的运行时间比较

Table 7　Comparison results of running time of four algorithms on 9 data sets

（单位：s）

数据集	SMOTE	Borderline-SMOTE	kmeans-SMOTE	SBO
yeast3	0.0891	1.3735	0.0924	**0.0785**
ecoli1	0.0105	0.1260	**0.0086**	0.0233
glass0	**0.0060**	0.0511	0.0089	0.0195
new_thyroid2	**0.0033**	0.1022	0.0046	0.0051
page_blocks0	0.9769	5.5794	0.6409	**0.6318**
yeast-2_vs_4	0.0141	0.6570	0.0157	**0.0140**
yeast05679vs4	0.0303	0.2962	0.0332	**0.0173**
ecoli4	0.0053	0.2003	0.0072	**0.0052**
glass4	0.0059	0.1329	0.0056	**0.0032**

从表 7 可以看出，在大多数数据集上 SBO 算法的运行时间短于其他算法，这是由于 SBO 算法仅对密集层的边界样本和稀疏层样本进行双向过采样，减少了过采样的样本数量，因此运行时间能够与之前的一些算法保持相近并优于某些算法。

综上所述，本文提出的 SBO 算法分类性能良好，无论数据集的非平衡度是否大于 9∶1，其都能达到较高的分类正确率，而且在时间效率上能够与之前的一些算法保持相近并优于某些算法。

结束语　本文提出了一种基于样本分层的双向过采样方法(SBO)，该方法首先基于少数类中的最高密度点和类内平均距离将少数类样本划分成密集层和稀疏层，然后对密集层边界区样本和稀疏层的样本进行双向过采样。SBO 算法既有效地解决了 SMOTE 算法引起样本重叠的问题，不会使密集的区域内的样本更加密集，稀疏的区域的样本依然很稀疏，又避免了噪音以及离群点的影响。实验结果表明，该算法的分类性能良好，特别是对于非平衡度大于 9∶1 的数据集，有

着明显的优势。但是该算法解决的是比较简单的二分类问题，现实生活中存在的问题大多是多分类问题，因此下一步将针对多分类问题进行研究。

参 考 文 献

[1] HE H，GARCIA E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering，2009，21(9)：1263-1284.

[2] ZHENG Z，WU X，SRIHARI R，et al. Feature selection for text categorization on imbalanced data [J]. SIGKDD Explorations，2004，6(1)：80-89.

[3] HUANG H，HE Q M，CHEN Q，et al. Rare category detection algorithm based on weighted boundary degree [J]. Journal of Software，2012，23(5)：1195-1208. (in Chinese)
黄浩，何钦铭，陈奇，等. 基于加权边界度的稀有类检测算法[J]. 软件学报，2012，23(5)：1195-1208.

[4] LOU X J，SUN Y X，LIU H T. Clustering boundary over-sampling classification method for imbalanced data sets [J]. Journal of Zhejiang University (Engineering Science)，2013，47(6)：944-950. (in Chinese)
楼晓俊，孙雨轩，刘海涛. 聚类边界过采样不平衡数据分类方法[J]. 浙江大学学报(工学版)，2013，47(6)：944-950.

[5] WANG H，ZHOU Z M. An over sampling algorithm based on clustering [J]. Journal of Shandong University (Engineering Science)，2018，48(3)：134-139. (in Chinese)
王换，周忠眉. 一种基于聚类的过抽样算法[J]. 山东大学学报(工学版)，2018，48(3)：134-139.

[6] WANG J H，DUAN B Q. Research on the SMOTE method based on density [J]. CAAI Transactions on Intelligent Systems，2017(6)：865-872. (in Chinese)
王俊红，段冰倩. 一种基于密度的 SMOTE 方法研究[J]. 智能系统学报，2017(6)：865-872.

[7] ZHU Y Q，DENG W B. A method using clustering and sampling approach for imbalance data [J]. Journal of Nanjing University (Natural Sciences)，2015，51(2)：421-429. (in Chinese)
朱亚奇，邓维斌. 一种基于不平衡数据的聚类抽样方法[J]. 南京大学学报(自然科学版)，2015，51(2)：421-429.

[8] YU Q，JIANG S J，ZHANG Y M，et al. The impact study of class imbalance on the performance of software defect prediction models [J]. Chinese Journal of Computers，2018，41(4)：809-824. (in Chinese)
于巧，姜淑娟，张艳梅，等. 分类不平衡对软件缺陷预测模型性能的影响研究 [J]. 计算机学报，2018，41(4)：809-824.

[9] 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. (in Chinese)
李雄飞，李军，董元方，等. 一种新的不平衡数据学习算法 PCBoost [J]. 计算机学报，2012，35(2)：202-209.

[10] JIN X，WANG L，SUN G Z，et al. Under-sampling method for unbalanced data based on centroid space [J]. Computer Science，2019，46(2)：50-55. (in Chinese)
金旭，王磊，孙国梓，等. 一种基于质心空间的不均衡数据欠采样方法 [J]. 计算机科学，2019，46(2)：50-55.

[11] BARUA S，ISLAM M M，YAO X，et al. MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Transactions on Knowledge and Data Engineering，2014，26(2)：405-425.

[12] 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. IEEE Xplore，2008：1322-1328.

[13] CHAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research，2011，16(1)：321-357.

[14] HAN H，WANG W Y，MAO B H. Borderline-SMOTE：A new over-sampling method in imbalanced data sets learning [C]// International Conference on Intelligent Computing，Springer-Verlag Berlin Heidelberg，2005，3644(5)：878-887.

[15] GEORGIOS D，FERNANDO B，FELIX L. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote [J]. Information Sciences，2018，465：1-20.

[16] ZHANG X，SONG Q，WANG G，et al. A dissimilarity-based imbalance data classification algorithm [J]. Applied Intelligence，2015，42(3)：544-565.

[17] XU Y，YANG Z，ZHANG Y，et al. A maximum margin and minimum volume hyper-spheres machine with pinball loss for imbalanced data classification [J]. Knowledge-Based Systems，2016，95：75-85.

[18] NEKOOEIMEHR I，LAI-YUEN S K. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets [J]. Expert Systems with Applications，2016，46：405-416.

[19] ANWAR N，JONES G，GANESH S. Measurement of data complexity for classification problems with unbalanced data [J]. Statistical Analysis and Data Mining，2014，7(3)：194-211.

[20] DEMSAR J. Statistical comparisons of classifiers over multiple data sets [J]. Journal of Machine Learning Research，2006，7(1)：1-30.