基于NKSMOTE算法的非平衡数据集分类方法

王 莉 陈红梅
(西南交通大学信息科学与技术学院 成都 611756)
(云计算与智能技术高校重点实验室(西南交通大学) 成都 611756)

摘 要 SMOTE(Synthetic Minority Over-sampling TEchnique)在进行样本合成时只在少数类中求其K近邻,这会导致过采样之后少数类样本的密集程度不变的问题。鉴于此,提出一种新的过采样算法NKSMOTE(New Kernel Synthetic Minority Over-Sampling Technique)。该算法首先利用一个非线性映射函数将样本映射到一个高维的核空间,然后在核空间上计算少数类样本在所有样本中的K个近邻,最后根据少数类样本的分布对算法分类性能的影响程度赋予少数类样本不同的向上采样倍率,从而改变数据集的非平衡度。实验采用决策树(Decision Tree,DT)、误差逆传播算法(error BackPropagation,BP)、随机森林(Random Forest,RF)作为分类算法,并将几类经典的过采样方法和文中提出的过采样方法进行多组对比实验。在UCI数据集上的实验结果表明,NKSMOTE算法具有更好的分类性能。

关键词 SMOTE算法,过采样,核空间,非平衡度,分类
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.09.043

NKSMOTE Algorithm Based Classification Method for Imbalanced Dataset

WANG Li CHEN Hong-mei
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)
(Key Laboratory of Cloud Computing and Intelligent Technology (Southwest Jiaotong University), Chengdu 611756, China)

Abstract In SMOTE(Synthetic Minority Over-sampling TEchnique), only minority class samples nearest to neighbors are computed when samples are synthesized, causing the problem that the density of the minority class samples remains unchanged after oversampling. This paper proposed an improved NKSMOTE(New Kernel Synthetic Minority Over-Sampling Technique) algorithm to overcome the shortage of SMOTE. Firstly, a nonlinear mapping function is used to map samples to a high-dimensional kernel space, and then the K nearest neighbors of samples of minority class from the whole samples are computed. In addition, different over-sampling rates are set on different minority samples to change the imbalanced multiplying power according to the influence caused by the distribution of minority class samples on the classification performance of algorithm. In the experiments, some classical oversampling methods were compared with the proposed oversampling method, and Decision Tree(DT), error BackPropagation(BP) and Random Forest(RF) were chosen as base classifier. Experimental results on UCI data sets show better classification performance of NKSMOTE algorithm.

Keywords SMOTE algorithm, Over-sampling, Kernel space, Imbalanced rate, Classification

1 引言

非平衡数据集分类是指各类样本数目不相等的情况下的分类问题[1]。以二分类问题为例,即在数据集中正类(负类)的样本数大大超过负类(正类)的样本数。在实际应用中,非平衡数据集分布广泛,比如医疗诊断、金融诈骗、网络入侵检测、电信用户检测、石油勘探等[2-6]。传统的分类算法大都以提高分类器的总体分类精度为目标,应用于非平衡数据集时很容易造成少数类的错分。

近年来,国内外学者通过不断的研究,从算法层面和数据层面提出了许多改进优化算法[7-9]。算法层面主要有集成学习、代价敏感和单类别学习等方法。集成学习法通过对多个基分类器进行集成来获得更强的泛化性能。胡小生等通过将随机欠采样和SMOTE算法相结合的方式来解决集成学习对非平衡数据效率不高的问题[10]。Galar等提出在集成学习中通过选择基分类器来提高其分类效率[11]。Kim等提出GM-

到稿日期:2017-08-12 返修日期:2017-12-19 本文受国家自然科学基金(61572406)资助。
王 莉(1992—),女,硕士,CCF会员,主要研究方向为数据挖掘,E-mail:13618022145@163.com;陈红梅(1971—),女,博士,教授,博士生导师,CCF会员,主要研究方向为智能信息处理、数据挖掘,E-mail:hmchen@swjtu.edu.cn(通信作者)。

Boost 算法,该算法通过修改分类器的评价标准,然后与采样算法相结合,来解决不平衡分类问题[12]。数据层面主要通过采样技术对数据集进行重构,以降低非平衡度,进而提高分类准确率。常见的采样方法有:过采样和欠采样。过采样通过增加少数类样本使数据集分布相对平衡;欠采样通过减少多数类样本使数据集分布相对平衡。Chawla 等于 2002 年提出 SMOTE 算法,其基本思想是,分别求出与样本集中所有少数类样本距离较近的少数类,然后在它们之间通过线性插值的方式来生成少数类,以降低数据集的非平衡度[13]。Han 等基于此算法进行改进,提出了 Borderline-SMOTE 算法,该算法考虑到边界样本对分类器的重要性,对位于边界的少数类样本进行过采样,解决了 SMOTE 算法对所有少数类样本进行过采样而导致过拟合的问题[14]。Dong 等提出 Random-SMOTE 算法,该算法是对每个少数类样本在不同方向上应用 SMOTE 算法,以解决 SMOTE 算法新合成的样本相对集中的问题[15]。王超学等把支持度概念和轮盘赌选择技术引入到 SMOTE 算法,提高了新合成样本的质量[16]。

为解决数据非线性可分的问题,学者们提出了核学习的方法。核学习方法的核心思想是通过非线性映射将原空间的样本映射到一个高维的核空间,然后在核空间进行线性分类。该方法的主要代表有支持向量机[17]、核主分量分析[18]。陶新民等采用谱聚类的方法,在核空间中对多数类样本进行谱聚类,然后在每个聚类中根据聚类大小和该聚类与少数类样本间的距离,选择具有代表意义的信息点,使数据集分布相对平衡[19]。曾志强等在核空间中对少数类样本进行过采样,然后通过原空间与核空间的距离关系寻找所合成样本在原空间的原像[20]。

针对 SMOTE 算法对所有的少数类样本一视同仁,未考虑不同少数类样本对分类器的重要度不同,而且 SMOTE 算法只是在少数类样本中求其 K 近邻,没有充分利用多数类样本信息的问题,本文在已有文献的基础上提出基于核空间的改进 SMOTE 算法——NKSMOTE,并采用了核映射的方法在核空间中计算少数类样本的近邻。在 UCI 数据集上与 SMOTE 算法、Random-SMOTE 算法、Borderline-SMOTE 算法、SSMOTE 算法进行性能对比,结果验证了 NKSMOTE 算法的有效性。

2　SMOTE 算法简介

SMOTE 算法的基本思想是通过人工合成新的少数类样本来改变样本分布。其基本原理是在相距较近的少数类样本之间进行线性插值,从而生成新的少数类样本。下面对 SMOTE 算法的关键步骤进行简单的介绍。

SMOTE 算法对少数类样本进行过采样,对于某个少数类样本 x,首先找到距其最近的 K 个少数类样本,若向上采样倍率为 N,则从 K 个少数类样本中随机选择 N 个少数类样本,记为 $y_1,y_2,\cdots,y_N$;最后,x 分别和 N 个少数类样本进行随机线性插值,生成 N 个新的少数类样本,记为 $X_{new1},X_{new2},\cdots,X_{newN}$,如式(1)所示:

$$X_{newj}=x+rand(0,1)*(y_j-x),j=1,2,\cdots,N \tag{1}$$

其中,$rand(0,1)$是指区间$(0,1)$内的一个随机数。

3　NKSMOTE 的基本原理及算法

针对 SMOTE 算法对所有的少数类样本一视同仁,未考虑不同少数类样本对分类器的不同重要度以及未充分利用 K 近邻中多数类样本信息的问题,本文提出一种基于核空间的改进 SMOTE 算法——NKSMOTE。NKSMOTE 算法的基本思想是在核空间中把少数类分成不同的类别,并根据类别的不同赋予其不同的向上采样倍率,然后在整个数据集中求少数类的 K 个近邻,最后按合成规则合成新的少数类样本。

3.1　NKSMOTE 的基本原理

(1)对于所有的少数类样本 x,在核空间上寻找距其最近的 K 个样本。根据 K 个样本中少数类样本的个数与多数类样本的个数,将少数类样本 x 分为安全样本、危险样本、噪声样本。若少数类样本的个数多于多数类样本的个数,则少数类样本 x 为安全样本;若少数类样本的个数少于多数类样本的个数,且存在少数类样本,则少数类样本 x 为危险样本;若全是多数类样本,则该少数类样本 x 为噪声样本。

(2)若少数类 x 不是噪声样本,则从其 K 近邻中随机选择 2 个样本,在 3 个样本之间按照一定的合成规则合成 N 个新样本,其中 N 值是向上采样倍率。若选中的两个样本 y_1 和 y_2 是多数类,则利用式(2)和式(3)得到新的少数类样本。

1)在 y_1 和 y_2 之间进行随机线性插值,生成 N 个临时样本 $t_{newj}(j=1,2,\cdots,N)$:

$$t_{newj}=y_1+rand(0,0.5)*(y_2-y_1) \tag{2}$$

其中,$rand(0,0.5)$表示区间$(0,0.5)$内的一个随机数。

2)在 t_{newj} 和 x 之间随机线性插值,构造新的少数类样本 $X_{newj}(j=1,2,\cdots,N)$:

$$X_{newj}=x+rand(0,1)*(t_{newj}-x) \tag{3}$$

其中,$rand(0,1)$表示区间$(0,1)$内的一个随机数。

若选中的两个样本 y_1 和 y_2 中有一个是少数类,则利用式(4)和式(3)得到新的少数类样本。

1)在 y_1 和 y_2 之间进行随机线性插值,生成 N 个临时样本 $t_{newj}(j=1,2,\cdots,N)$:

$$t_{newj}=y_1+rand(0,1)*(y_2-y_1) \tag{4}$$

2)在 t_{newj} 和 x 之间利用式(3)进行随机线性插值,构造新的少数类样本 $X_{newj}(j=1,2,\cdots,N)$。

(3)若少数类 x 是噪声样本,对其进行过采样时给数据集引入了噪声的风险,但是噪声样本又有其积极作用。为了使风险降到最低,设定其向上采样倍率 N 为 1。在少数类中随机选择一个少数类 y,在 x 与 y 之间进行随机线性插值,其中增量因子是一个在$(0.5,1)$区间上服从均匀分布的随机数,使新的少数类更靠近少数类 y。

$$X_{new}=x+rand(0.5,1)*(y-x) \tag{5}$$

3.2　核距离

将样本通过非线性映射函数映射到核空间,在核空间中样本之间的距离被称为核距离。NKSMOTE 算法中涉及核距离的计算,对于核空间中任意两个样本点 $\varphi(x)$ 和 $\varphi(y)$,其核距离的计算式如下:

$$\begin{aligned}d(\varphi(x),\varphi(y))&=\sqrt{\|\varphi(x)-\varphi(y)\|^2}\\&=\sqrt{K(x,x)+K(y,y)-2K(x,y)}\end{aligned} \tag{6}$$

其中，$\varphi()$是非线性映射函数，$K(,)$是核函数。

目前常用的核函数有高斯核函数、多项式核函数、S 型核函数等。

(1)高斯核函数

$$K(x,z)=\exp(-\frac{\|x-z\|^2}{2\sigma^2}) \tag{7}$$

(2)多项式核函数

$$K(x,z)=(x\cdot z+1)^p \tag{8}$$

(3)S 型核函数

$$K(x,z)=\tanh[\alpha(x,z)+c] \tag{9}$$

本文采用的是高斯核函数。

3.3 NKSMOTE 算法

根据 NKSMOTE 的基本原理，给出具体算法，如算法 1 所示。

算法 1 NKSMOTE 算法

输入：少数类的数量 T；向上采样倍率 N；K 近邻个数(K>3)；少数类数据集 D_{min}；少数类样本 x_i，i=1,2,…,T

输出：合成的少数类样本

```
1. if N<100%
2. T←N * T;
3. N←100%;
4. endif
5. for i=1 to T
6. Di←NearestExample(xi,K);//在核空间从整个数据集中求 xi 的 K 个近邻
7. m←NumMajorityNeighbor(xi,K);//xi 的 K 近邻中多数类的个数
8. if m==K
9. y←RandomExample(Dmin);//在少数类数据集中随机选择一个样本
10. Xnew←xi+rand(0.5,1) * (y-xi);//合成新的少数类样本
11. else
12. for j=0 to N
13. y1←RandomExample(Di);//在 K 近邻中随机选择一个样本
14. y2←RandomExample(Di);
15. tnewj←CreatExample(y1,y2);//调用合成样本函数
16. Xnewj←xi+rand(0,1) * (tnewj-xi);
17. j++;
18. endfor
19. endif
20. i++;
21. endfor
22. return Xnewj,Xnew
    CreatExample(y1,y2)//合成样本函数
23. if y1.label==0 && y2.label==0
24. tnewj←y1+rand(0,0.5) * (y2-y1);
25. else
26. tnewj←y1+rand(0,1) * (y2-y1);
27. endif
```

4 实验设计与结果分析

为了验证本文所提算法的有效性，在实际的非平衡数据集上分别用 SMOTE[13]，Random-SMOTE[15]，Borderline-SMOTE[14]，SSMOTE[16]和 NKSMOTE 对其进行预处理，然后在新的数据集上用决策树(J48)、误差逆传播算法(BP)、随机森林(RF)进行分类。通过对实验结果进行分析，得出 NKSMOTE 算法比上述几种过采样算法更优的结论。

4.1 数据集

实验所用数据集来自 UCI 数据库[1)]，表 1 列出了这些非平衡数据集的基本信息，包括数据集名称、样本数、属性数、少数类样本数以及非平衡度(Imbalanced Rate，IR)。

表 1 数据集

Table 1 Date set

名称	样本数	属性数	少数类样本数	非平衡度
ecoli3	336	7	35	8.60
ecoli1	336	7	77	3.36
wisconsin	683	9	239	1.86
yeast0	1004	8	99	9.14
yeast3	1484	8	163	8.10

4.2 评价标准

在传统的分类学习方法中，一般采用分类精度作为评价指标，然而对于非平衡数据集而言，用分类精度来评价分类器的性能是不合理的。在机器学习领域中，非平衡数据分类的常用评价标准包括受试者工作特征曲线 ROC(Receiver Operating Characteristic)、AUC(Area Under ROC Curve)以及基于混淆矩阵的 F-value 和 G-mean。在非平衡数据学习中，少数类对应为正类，多数类对应为负类。表 2 列出了二分类问题的混淆矩阵。

表 2 混淆矩阵

Table 2 Confusion matrix

	预测正类	预测负类
实际正类	TP(True Positives)	FN(False Negatives)
实际负类	FP(False Positives)	TN(True Positives)

根据混淆矩阵可以得到以下评价指标。

(1)真正率

$$TP_{rate}=\frac{TP}{TP+FN} \tag{10}$$

(2)真负率

$$TN_{rate}=\frac{TN}{TN+FP} \tag{11}$$

(3)假正率

$$FP_{rate}=\frac{FP}{TN+FP} \tag{12}$$

(4)正类预测值

$$PP_{value}=\frac{TP}{TP+FP} \tag{13}$$

其中，真实正类率 TP_{rate} 又称为查全率 *recall*，正类预测值 PP_{value} 又称为查准率 *precision*。

$$F\text{-}value=\frac{(1+\beta^2)precision\cdot recall}{(\beta^2\cdot recall+precision)} \tag{14}$$

其中，β 用于调节 *precision* 和 *recall* 的相对重要度，通常取值为 1。

1) http://archive.ics.uci.edu/ml

如果同时关注两个类的性能,可以使用 *G-mean* 评价算法在两个类上的性能。

$$G\text{-}mean=\sqrt{TP_{\text{rate}}\times TN_{\text{rate}}} \tag{15}$$

ROC 曲线是 TP_{rate} 与 FP_{rate} 关系的可视化表示,其中 TP_{rate} 是纵坐标,FP_{rate} 是横坐标,所以曲线越靠近左上角表示分类器的性能越好。当两条 ROC 曲线相交时,很难判断出哪条曲线更好,因此计算 ROC 曲线下方的面积 *AUC* 作为评价指标。

$$AUC=\frac{1+TP_{\text{rate}}-FP_{\text{rate}}}{2} \tag{16}$$

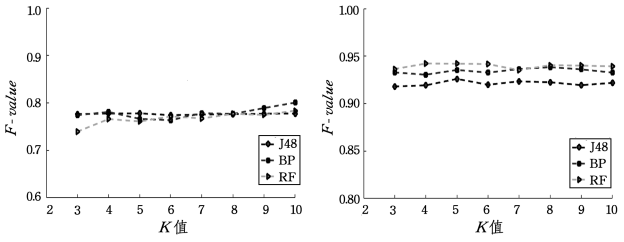
本文采用 *F-value*,*G-mean*,*AUC* 作为评价度量。

4.3　近邻 K 和高斯核参数 σ 的选择

近邻的数量和高斯核函数的参数对 NKSMOTE 算法会产生一定的影响,下文将分别进行讨论。

4.3.1　近邻参数 K 的选取

在核空间中求少数类的 K 个近邻,然后根据 K 个近邻中少数类样本的个数将该少数类分为不同的类别,并赋予其不同的采样倍率。因此,K 的大小对少数类的划分有一定影响。本文假设 NKSMOTE 算法中 K 的取值范围为 3～10,比较不同分类算法在不同 K 值下的 *F-value* 值,进而确定 K 值。图 1 给出了数据集 ecoli1,wisconsin 在 J48,BP,RF 算法上近邻参数 K 与 *F-value* 值的关系。



(a)ecoli1　　(b)wisconsin

图 1　近邻参数 K 与 *F-value* 的关系

Fig. 1　Relationship between K and *F-value*

由图 1 可以看出,在数据集 ecoli1 上,J48 在 $K=5$ 时取得最优 *F-value* 值,BP 和 RF 在 $K=10$ 时取得最优 *F-value* 值;在数据集 wisconsin 上,3 种分类算法在 $K=4,5,6$ 时都可以取得较高的 *F-value* 值。为了保证在其他数据集上的效果,本文在实验中设 NKSMOTE 的参数 K 为 5。

4.3.2　高斯核参数 σ 的选取

本文是在核空间中求少数类的近邻,为了研究高斯核参数 σ 对分类效果的影响,将 σ 分别设置为 0.05,0.1,0.3,0.5,0.7,1.0,在数据集 ecoli1 和 wisconsin 上采用五折交叉验证法对算法 J48,BP,RF 进行分类性能测试。图 2 给出了数据集 wisconsin 在 J48,BP,RF 算法上 σ 与 *F-value* 值的关系。图 3 给出了数据集 ecoli1 和 wisconsin 在 J48 算法上 σ 与 *F-value* 值的关系。

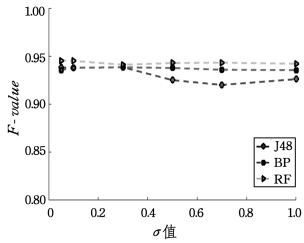


图 2　高斯核参数 σ 与 *F-value* 的关系

Fig. 2　Relationship between σ and *F-value*

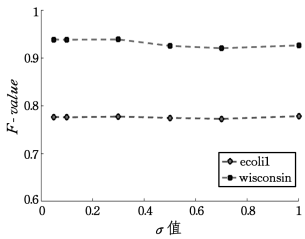


图 3　J48 算法上 σ 与 *F-value* 的关系

Fig. 3　Relationship between σ and *F-value* on J48 algorithm

由图 2 可以看出,3 种分类算法取得较优 *F-value* 值时的 σ 值不同,随着 σ 值的变化,不同算法的最优 *F-value* 值的变化较小。由图 3 可以看出,J48 在数据集 ecoli1 上取得较优 *F-value* 值时 σ 为 1,在数据集 wisconsin 上取得较优的 *F-value* 值时 σ 为 0.3。虽然在不同数据集上取得最优 *F-value* 值时的 σ 值不同,但是同一数据集在不同 σ 值时 *F-value* 值的变化不大。

综上所述,在不同算法、不同数据集上的 σ 值不同时,最优 *F-value* 的取值可能不同,但是对 *F-value* 值的影响范围不大。为方便讨论,本文实验中取 σ 值为 1。

4.4　实验结果及分析

实验采用五折交叉验证法,用 Java 实现 SMOTE,Random-SMOTE,Borderline-SMOTE,SSMOTE,NKSMOTE 算法,本文使用 J48,BP,RF 分类算法。

表 3－表 5 分别列出了以 J48,BP,RF 为分类算法时 5 种过采样算法在 5 个数据集上的 *F-value* 值、*G-mean* 值和 *AUC* 值的比较。其中,S 代表 SMOTE,RS 代表 Random-SMOTE,BS 代表 Borderline-SMOTE,SS 代表 SSMOTE,NKS 代表 NKSMOTE,w 代表数据集 wisconsin。

表 3　*F-value* 值的比较

Table 3　Comparison of *F-value* values

数据集	J48					BP					RF				
	S	RS	BS	SS	NKS	S	RS	BS	SS	NKS	S	RS	BS	SS	NKS
ecoli3	0.6292	0.5844	0.5920	0.5800	**0.6334**	0.6092	0.6324	0.6468	0.6372	**0.7248**	0.5566	0.5552	0.5568	0.5596	**0.5662**
ecoli1	0.7680	0.7740	0.7722	**0.7792**	0.7780	0.7862	0.7752	**0.8034**	0.7890	0.7664	0.7776	**0.7936**	0.7816	0.7782	0.7610
w	0.9170	0.9178	0.9248	0.9210	**0.9260**	0.9314	0.9338	0.9384	**0.9402**	0.9354	0.9396	0.9418	0.9330	0.9410	**0.9420**
yeast0	0.7436	0.7272	0.7446	0.7424	**0.7452**	**0.7722**	0.7636	0.7678	0.7388	0.7516	0.7960	0.7974	0.7774	0.7864	**0.8008**
yeast3	0.7240	0.7268	0.7340	0.7144	**0.7360**	0.7454	0.7376	**0.7474**	0.7304	0.7356	0.7336	0.7304	0.7232	0.7348	**0.7400**

表 4 *G-mean* 值的比较

Table 4 Comparison of *G-mean* values

数据集	J48					BP					RF				
	S	RS	BS	SS	NKS	S	RS	BS	SS	NKS	S	RS	BS	SS	NKS
ecoli3	0.7400	0.7092	0.6920	0.7108	**0.7488**	0.7590	0.7906	0.7822	0.7918	**0.8492**	0.6918	0.6972	0.6786	0.6994	**0.7124**
ecoli1	0.8348	0.8434	0.8336	**0.8476**	0.8472	0.8592	0.8536	**0.8682**	0.8562	0.8356	0.8376	**0.8502**	0.8420	0.8350	0.8232
w	0.9362	0.9376	0.9428	0.9396	**0.9436**	0.9484	0.9484	0.9556	**0.9566**	0.9522	0.9550	0.9534	0.9464	**0.9566**	0.9544
yeast0	0.8194	0.8100	0.8166	**0.8216**	0.8204	0.8568	**0.8752**	0.8576	0.8596	0.8736	0.8476	0.8606	0.8390	0.8472	**0.8626**
yeast3	0.8240	0.8194	0.8360	0.8264	**0.8380**	0.8608	0.8480	**0.8684**	0.8616	0.8516	0.8208	0.8122	0.8224	0.8250	**0.8288**

表 5 *AUC* 值的比较

Table 5 Comparison of *AUC* values

数据集	J48					BP					RF				
	S	RS	BS	SS	NKS	S	RS	BS	SS	NKS	S	RS	BS	SS	NKS
ecoli3	0.7816	0.7552	0.7428	0.7588	**0.7870**	0.7868	0.8146	0.8122	0.8156	**0.8634**	0.7358	0.7396	0.7282	0.7400	**0.7540**
ecoli1	0.8496	0.8564	0.8484	**0.8596**	0.8592	0.8708	0.8658	**0.8802**	0.8688	0.8510	0.8510	**0.8616**	0.8548	0.8484	0.8382
w	0.9456	0.9454	0.9510	0.9482	**0.9524**	0.9582	0.9574	0.9574	**0.9656**	0.9622	0.9616	0.9614	0.9552	**0.9626**	0.9622
yeast0	0.8400	0.8318	0.8384	0.8406	**0.8412**	0.8742	**0.8886**	0.8752	0.8750	**0.8864**	0.8638	0.8748	0.8576	0.8632	**0.8784**
yeast3	0.8416	0.8372	**0.8496**	0.8398	0.8492	0.8730	0.8624	**0.8784**	0.8740	0.8646	0.8402	0.8324	0.8400	0.8436	**0.8472**

由表 3—表 5 可知，当 J48 作为分类算法且用 *F-value* 值作为评价度量时，NKSMOTE 比 SMOTE 最多可提高 1.2%；比 Random-SMOTE 提高 0.4%～4.9%；比 BSMOTE 最多可提高 4.14%；NKSMOTE 除了在数据集 ecoli1 上略逊于 SSMOTE，在其他数据集上比 SSMOTE 最多可提高 5.34%。用 *G-mean* 作为评价度量时，NKSMOTE 比 SMOTE 最多可提高 1.4%；比 Random-SMOTE 最多可提高 3.96%；比 BSMOTE 最多可提高 5.68%；NKSMOTE 除了在数据集 ecoli1 和 yeast0 上略逊于 SSMOTE，在其他数据集上比 SSMOTE 算法最多可提高 3.8%。用 *AUC* 作为评价度量时，NKSMOTE 比 SMOTE 最多可提高 0.96%；比 Random-SMOTE 最多可提高 3.18%；NKSMOTE 除了在数据集 yeast3 上略逊于 BSMOTE，在其他数据集上比 BSMOTE 最多可提高 4.42%；另外，NKSMOTE 除了在数据集 ecoli1 上略逊于 SSMOTE，在其他数据集上比 SSMOTE 最多可提高 2.82%。

当 BP 作为分类算法且用 *F-value* 值作为评价度量时，NKSMOTE 略逊于其他几种过采样算法。用 *G-mean* 作为评价度量时，NKSMOTE 在大多数数据集上优于 SMOTE 和 Random-SMOTE 算法，略逊于 BSMOTE 和 SSMOTE 算法。用 *AUC* 作为评价度量时，NKSMOTE 在大多数数据集上优于 SMOTE，Random-SMOTE，BSMOTE，略逊于 SSMOTE 算法。

当 RF 作为分类算法且用 *F-value* 值作为评价度量时，相较于其他几种过采样算法，NKSMOTE 除了在数据集 ecoli1 上表现略差之外，在其他数据集上都有不同程度的提高。用 *G-mean* 作为评价度量时，NKSMOTE 除了在数据集 ecoli1 和 wisconsin 上略逊于 SMOTE，SSMOTE，在其他数据集上比 SMOTE 最多可提高 2.06%，比 SSMOTE 最多可提高 1.54%；NSMOTE 算法除了在数据集 ecoli1 上略逊于 Random-SMOTE，在其他数据集上比 Random-SMOTE 最多可提高 1.66%；比 BSMOTE 最多可提高 3.38%。用 *AUC* 作为评价度量时，NKSMOTE 除了在数据集 wisconsin 上略差于 SSMOTE，在数据集 ecoli1 上略逊于其他几种过采样算法，在其他数据集上比 SMOTE，Random-SMOTE，BSMOTE，SSMOTE 都有不同程度的提高。

总的来说，NKSMOTE 算法比 SMOTE 算法、Random-SMOTE 算法、BSMOTE 算法、SSMOTE 算法更优，可以提高分类器的分类性能。

结束语 针对非平衡数据集的分类问题，本文基于数据层面的过采样算法。提出了 NKSMOTE 算法，该算法在核空间中求少数类的 K 近邻，并将 K 近邻的范围扩大到整个数据集，使得生成的少数类更具有真实性，并根据少数类的分布特点，赋予其不同的向上采样倍率。所提算法的时间复杂度与 SMOTE 算法相同。在数据集上将 NKSMOTE 算法和几种过采样算法的性能进行对比，实验结果表明 NKSMOTE 算法具有更好的分类性能。进一步的研究工作是将欠采样和集成学习的思想与本文算法融合起来，以解决非平衡数据分类困难的问题。

参 考 文 献

[1] WEISS G M, ZADROZNY B, SAAR M. Guest editorial: special issue on utility-based data mining[J]. Data Mining and Knowledge Discovery, 2008, 17(2): 129-135.

[2] DEL C, SERRANO J. A multistrategy approach for digital text categorization from imbalanced documents[J]. Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining Explorations, 2004, 6(1): 70-79.

[3] WEI W, LI J, CAO L. Effective detection of sophisticated online banking fraud on extremely imbalanced data[J]. World Wide Web, 2013, 16(4): 449-475.

[4] HANG Z. Imbalanced data classification method and its application research for intrusion detection[J]. Computer Science, 2013, 40(4): 131-135.

[5] KUBAT M, HOLTE R C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning, 1998, 30(2): 195-215.

[6] ZHANG J W. Imbalanced data classification and its application in cancer recognition[D]. Hangzhou: China Jiliang Unversity, 2012. (in Chinese)
张金伟. 不平衡数据分类研究及在肿瘤识别中的应用[D]. 杭州：中国计量学院，2012.

[7] JASON V H, TAGHI K. Knowledge discovery from imbalanced and noisy data[J]. Data Knowledge Engineering, 2009, 68(12): 1513-1542.

[8] YANG Z M,QIAO L Y,PENG X Y. Research on datamining method for imbalanced dataset based on improved SMOTE[J]. Acta Electronica Sinica,2007,35(12):22-26. (in Chinese)
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(12):22-26.

[9] WANG L Y. Research of boosting classification algorithm for imbalance data[D]. Harbin: Harbin Institute of Technology, 2013. (in Chinese)
王璐林. 面向不平衡样本的 Boosting 分类算法研究[D]. 哈尔滨:哈尔滨工业大学,2013.

[10] HU X S,WEN J P,ZHONG Y. Imbalanced data ensemble classification using dynamic balance sampling[J]. CAAI Transactions on Intelligent Systems,2016,11(2):257-263. (in Chinese)
胡小生,温菊屏,钟勇. 动态平衡采样的不平衡数据集成分类方法[J]. 智能系统学报,2016,11(2):257-263.

[11] GALAR M,FERNANDEZ A,BARRENECHEA E. Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced data sets[J]. Information Sciences,2016,354(C):178-196.

[12] KIM M J,KANG D K,HONG B K. Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction[J]. Expert Systems with Applications,2015,42(3):1074-1082.

[13] CHAWLA N V,BOWYER K W,HALLO L O. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16(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]// Proc. of International Conference on Intelligent Computing. 2005:878-887.

[15] DONG Y,WANG X. A new over-sampling approach: Random-SMOTE for learning from imbalanced data Sets[C]// Internation al Conference on Knowledge Science, Engineering and Management. 2011:343-352.

[16] WANG C X,PAN Z M,DONG L L,et al. Research on classification for imbalanced dataset based on improved SMOTE[J]. Computer Engineering and Applications,2013,49(2):184-187. (in Chinese)
王超学,潘正茂,董丽丽,等. 基于改进 SMOTE 的非平衡数据集分类研究[J]. 计算机工程与应用,2013,49(2):184-187.

[17] CRISTIANINI N,SHAWE T J. An introduction to support vector machines: and other kernel-based learning methods[M]. Cambridge University Press,2000.

[18] SCHOIKOPF B,MIKA S,BURGES C J C. Input space versus featurespace in kernei-based methods[J]. IEEE Transactions on Neural Networks,1999,10(5):1000-1017.

[19] TAO X M,ZHANG D M,HAO S Y. SVM classifier for unbalanced data based on spectrum cluster-based under-sampling approaches[J]. Control and Decision,2012,27(12):1761-1768. (in Chinese)
陶新民,张冬梅,郝思媛. 基于谱聚类欠取样的不均衡数据 SVM 算法[J]. 控制与决策,2012,27(12):1761-1768.

[20] ZENG Z Q,WU Q,LIAO B S. A classfication method for imbalance data set based on kernel SMOTE[J]. Acta Electronica Sinica,2009,37(11):2489-2495. (in Chinese)
曾志强,吴群,廖备水. 一种基于核 SMOTE 的非平衡数据集分类方法[J]. 电子学报,2009,37(11):2489-2495.

(上接第 259 页)

[12] OTSUKA E,CHIU D. Design and evaluation of a Twitter hashtag recommendation system[C]// International Database Engineering & Applications Symposium. ACM,2014:330-333.

[13] GAO M,JIN C Q,QIAN W N,et al. Real-time and personalized recommendation on microblogging systems[J]. Chinese Journal of Computers,2014,37(4):963-975. (in Chinese)
高明,金澈清,钱卫宁,等. 面向微博系统的实时个性化推荐[J]. 计算机学报,2014,37(4):963-975.

[14] QIU H H,LIU Y,ZHANG Z J,et al. An Improved Collaborative Filtering Recommendation Algorithm for Microblog Based on Community Detection[C]// Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE,2014:876-879.

[15] CHEN X. A Hybrid Microblog Recommendation Model in Mobile Social Network[J]. Journal of Electronic Commerce in Organizations,2014,12(4):69-79.

[16] JIANG C. A Mircroblog Recommendation System Based on User Clustering and Semantic Dictionary[D]. Hangzhou: Zhejiang University,2013. (in Chinese)
蒋超. 基于用户聚类和语义词典的微博推荐系统[D]. 杭州:浙江大学,2013.

[17] CHEN L,JIANG C,WANG W. A Micro blog Recommendation System Based on User Clustering[C]// 2014 International Conference on Computer Science and Electronic Technology (ICCSET 2014). Atlantis Press,2015.

[18] XI Y,YANG J,TANG C H,et al. An Overlapping Semantic Community Detection Algorithm Based on Local Semantic Cluster[J]. Journal of Computer Research & Development,2015,52(7):1510-1521. (in Chinese)
辛宇,杨静,汤楚蘅,等. 基于局部语义聚类的语义重叠社区发现算法[J]. 计算机研究与发展,2015,52(7):1510-1521.

[19] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research,2003(3):993-1022.

[20] BERGROTH L,HAKONEN H,RAITA T. A survey of longest common subsequence algorithms[C]// International Symposium on String Processing and Information Retrieval, 2000 (Spire 2000). IEEE,2000:39-48.

[21] ZHANG J P,XIE J,YANG J,et al. A t-closeness privacy model based on sensitive attribute values semantics bucketization[J]. Journal of Computer Research & Development,2014,51(1):126-137. (in Chinese)
张健沛,谢静,杨静,等. 基于敏感属性值语义桶分组的 t-closeness 隐私模型[J]. 计算机研究与发展,2014,51(1):126-137.

[22] GAO C,MIAO D Q,ZHANG Z F,et al. A semi-supervised rough set model for classification based on active learning and co-training[J]. Pattern Recognition & Artificial Intelligence,2012,25(5):745-754. (in Chinese)
高灿,苗夺谦,张志飞,等. 主动协同半监督粗糙集分类模型[J]. 模式识别与人工智能,2012,25(5):745-754.