基于单边选择链和样本分布密度融合机制的非平衡数据挖掘方法

翟 云[1,2],王树鹏[3],马 楠[4],杨炳儒[2],张德政[2]
(1.国家行政学院电子政务研究中心,北京 100089;2.北京科技大学计算机与通信工程学院,北京 100083;
3.中国科学院信息工程研究所,北京 100093;4.北京联合大学信息学院,北京 100101)

摘 要: 非平衡数据集分类问题是机器学习领域的重大挑战性难题.针对该难题,传统的少数类样本合成技术(Synthetic Minority Over-Sampling Technique,SMOTE)已成为一种有力手段并得到广泛采用.但在新样本生成过程中,SMOTE 利用所有少数类样本合成新样本,由此产生过拟合瓶颈.为更好地解决该问题,提出了一种基于单边选择链和样本分布密度的非平衡数据挖掘新方法(One-Sided Link & Distribution Density-SMOTE,OSLDD-SMOTE).OSLDD-SMOTE 通过单边选择链遴选出处于分类边界的少数类样本,根据这些样本的动态分布密度生成新样本.进而分析了样本合成度对节点数目和对少数类精度的影响;基于 G-mean、F-measure 和 AUC 三个指标综合比较了 OSLDD-SMOTE 与其他同类方法的分类性能.实验结果表明,OSLDD-SMOTE 有效提高了少数类样本的分类准确率.

关键词: 非平衡数据分类;单边选择链;分布密度;重采样
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2014)07-1311-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.07.011

A Data Mining Method for Imbalanced Datasets Based on One-Sided Link and Distribution Density of Instances

ZHAI Yun[1,2], WANG Shu-peng[3], MA Nan[4], YANG Bing-ru[2], ZHANG De-zheng[2]
(1. *E-Government Research Center*, *Chinese Academy of Governance*, *Beijing*, 100089, *China*;
2. *School of Computer and Communication Engineering*, *University of Science and Technology Beijing*, *Beijing* 100083, *China*;
3. *Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100093, *China*;
4. *College of Information Technology*, *Beijing Union University*, *Beijing* 100101, *China*)

Abstract: Classification in imbalanced datasets poses a great challenge to machine learning region, where the synthetic minority over-sampling technique(SMOTE)has become a powerful means and widely adopted as an effective method. But in generating new instances, SMOTE uses all instances in minority class such that it takes with over-generalization. To better solve the problem, a data mining method for imbalanced datasets based on one-sided link and distribution density of the minority(OSLDD-SMOTE)is proposed in this paper. OSLDD-SMOTE firstly selects the minority near the classification boundary using the one-sided link, then generates new instances with SMOTE based on the dynamic distribution density of these instances. Effects of synthetic degree on new generated instances and accuracy of the minority are respectively compared with the OSLDD-SMOTE, SMOTE, Borderline-SMOTE and Surrounding-SMOTE method. Furthermore, from the simulation results with 8 UCI datasets, our proposed method has the most accurate and robust performance on the G-mean, F-measure and AUC metrics.

Key words: classification in imbalanced datasets; one-sided link; distribution density; resample

1 引言

近年来,非平衡数据集分类问题开始受到机器学习领域的重视.2000 年人工智能研究进展国际会议(Association for the Advancement of Artificial Intelligence, AAAI)和 2003 年机器学习国际会议(International Conference on Machine Learning Workshop on Learning from Imbalanced Data Sets, ICML)分别设立论坛进行专题研讨.2005 年数据挖掘国际会议上(IEEE International Conference on Data Mining, ICDM),非平衡类数据分类被列为数据挖掘领域

收稿日期:2013-11-03;修回日期:2014-03-03;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61300078,No.61271275);国家行政学院科研招标课题(No.2012ZBKT016)

的十大挑战性难题之一.现实世界里,非平衡数据分类问题是常见的,如通过对不同病人检查形成的一系列乳房-射线数据集已在处理非平衡类数据算法中得到广泛应用.其中,癌变和健康的病例分别分到少数类和多数类样本.事实上,非癌变的病人数目要远远大于癌变的病人数目.其他应用如信用卡欺骗检测[1,2]、文本分类[3]、信息搜索及过滤[4]、市场行为分析[5]等,在这些应用中,人们主要关心的是数据集中的少数类样本,因为这些样本的错分代价异常大,甚至是不可估量的.

目前,关于非平衡数据集分类问题的研究一般从三个方面展开:重采样技术[6]、集成学习的分割投票方法[7]以及基于代价敏感的权重方法[8].重采样方法主要包括过采样和欠采样方法.本文拟从过采样方面展开讨论.尽管许多学者开始探讨该问题,以 SMOTE 方法为代表的过采样技术从一定程度上提高了少数类样本的分类精度[9~13],但仍存在一定问题,主要体现在:(1)有些方法利用 SMOTE 技术对所有少数类样本进行合成,提高了时间代价;(2)有些方法尽管试图找出边界少数类样本,但仍未考虑其实时分布状况;(3)没有考虑到噪音数据对 SMOTE 过程的重要影响.可见,如何利用 SMOTE 方法对边界少数类样本合成,从而使其更充分体现样本特征,进而提高分类器泛化能力仍需深入探讨.基于此,我们提出了一种基于单边选择链和样本分布密度的改进 SMOTE 技术,利用单边选择链有效识别出边界少数类样本,同时根据样本的动态分布密度进行少数类样本的合成,避免了原始方法中样本生成的盲目性,丰富了少数类样本的特征,从而提高了合成样本的质量,进而提高整体分类准确率.

2 相关工作

非平衡数据集具有鲜明的样本分布特征,下面先给出一些基本概念和术语(本文只讨论两分类问题).

定义 1 称 $\Omega(X,Y,A,D)$为非平衡数据集,其中,Ω 是有限样本集合,$X\cup Y=\Omega$,其中,X,Y 分别是少数类样本集和多数类样本集,$|X|<|Y|$,且满足 $X\cap Y=\varnothing$.对 $\forall x_i\in X(i=1,2,\cdots,m)$称为正样本(Positive Instance),$\forall y_j\in Y(j=1,2,\cdots,n)$称为负样本(Negative Instance).A 是样本属性集,$A=(a_1,a_2,\cdots,a_n)$,其中,n 为样本的属性维度.D 是 Ω 的非平衡度,$D=|Y|/|X|$;可见,反映了样本集 Ω 的非平衡状态:D 越大,样本集 Ω 非平衡程度越严重,反之亦然.一般情况下,$D>1$.

当对非平衡数据集采样时,原本是为了更准确地获得"有代表性的样本"[14],其训练算法通常对多数类样本会产生很高的预测准确率,由于样本数量非平衡,常见的分类算法对少数类样本分类精度不尽如人意.为解决非平衡数据集分类问题,Chawla 提出了少数类样本过采样技术(Synthetic Minority Over-sampling Technique,SMOTE),这种方法在少数类样本与其 k 个最近邻居连线上生成新样本,方法如下:

$$\text{Synthetic}[l][\text{attr}]=\text{instance}[i][attr]+gap*dif \quad (1)$$

其中,dif 为第 i 个少数类样本与其第 j 个近邻的全部属性值之差;gap 是一个随机数,$gap\in[0,1]$.

SMOTE 可使分类器对于少数类样本具有更大的泛化空间,允许分类器更好地预测未知少数类样本,进而提高分类器的整体分类准确率.然而,SMOTE 技术在生成新样本过程中,是基于所有少数类样本的,故存在一定盲目性,特别是在数据集极端非平衡时,新产生的少数类样本往往会造成严重的样本混叠现象,影响分类效果.针对这一问题,在 SMOTE 原有技术基础上,文献[10]提出了一种自适应 SMOTE 方法,根据样本集内部分布特性,自适应调整 SMOTE 方法中近邻选择策略,控制合成样本的质量;文献[11]则只对边界样本利用 STOME 方法合成新样本;文献[12]提出一种基于核 SMOTE(Synthetic Minority Over-sampling Technique)的分类方法来处理支持向量机(SVM)在非平衡数据集上的分类问题.其核心思想是首先在特征空间中采用核 SMOTE 方法对少数类样本进行上采样,然后通过输入空间和特征空间的距离关系寻找所合成样本在输入空间的原像,最后再采用 SVM 对其进行训练.文献[13]通过 SMOTE 技术人工增加少数类样本量,以具有较强分类性能和泛化性能的 SVM 作为弱分类器,并以 AdaBoost 算法构建集成分类器.文献[15]提出了一种基于大边界原理(Large Margin Principle)的过采样方法,在人工数据集和多个 UCI 数据集上均表现出较好的性能.

3 单边选择链和样本分布密度 SMOTE 方法

3.1 数据预处理

由第二节论述可知,SMOTE 方法本质上是由某一少数类样本与其 k 个近邻合成新样本,以此填充样本空间,实现数据集类间平衡.但是,如果该少数类样本的近邻是噪音数据,则不但不会生成更多有益的样本,反而会误导分类器,进而降低其分类精度.如图 1 所示,样本 X_1 与噪音数据 X'_1、X'_2 生成合成样本 XX'_1、XX'_2,非但没丰富少数类样本空间,为分类器提供更多的分类信息,反而进一步恶化了样本的分布状况,降低了分类器模型的泛化能力.可见,在 SMOTE 之前必须对数据集预处理,以此消除噪音数据,使分类边界更光滑.

噪音数据处理步骤如下:

首先,把训练集等分成十份,将其中的九份做训练集,剩余的一份做测试集.对训练集中任意样本,如果某一分类器判定其属于另一类,则该分类器把它标识为可疑样本.鉴于集成分类模型具有良好的分

类性能和较强的鲁棒性[16~19],我们在集成分类器上采取十交叉验证.集成时我们采用一致过滤原则[20],即所有基分类器均认为该样本可疑时,集成分类器才判定其可疑.

然后,对可疑样本进行进一步处理,常用方法既可将可疑样本删除,又可将其标识为异类样本[21,22].考虑到边界处少数类样本分布稀疏,故我们采取最近邻决策机制,如图 2 所示.

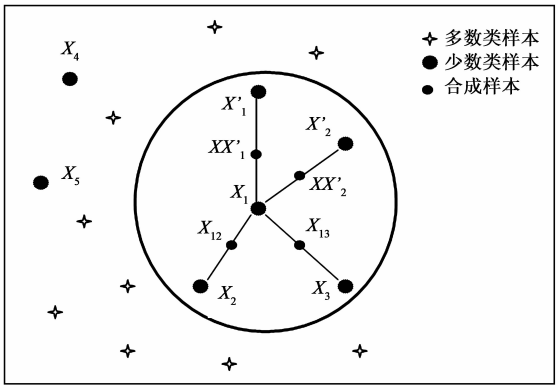

图1　带噪音数据的SMOTE

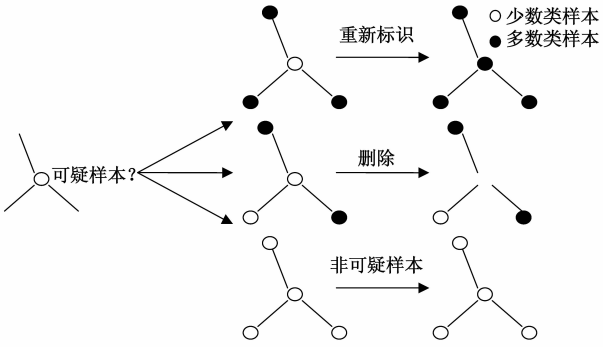

图2　最近邻决策机制

3.2　选择链生成

已有研究证明[21],分类边界处的样本对分类精度起到至关重要的作用,而远离边界点的样本对分类精度影响甚微.基于此,为了更有效地提高分类速度,在不影响分类精度的前提下,我们提出单边选择链方法(One-sided Link,OSL),力图找到靠近分类边界的少数类样本,然后利用 SMOTE 方法进行过采样,达到数据类间平衡进而提高分类性能之目的.

定义 2　对任意少数类样本 x_i,最近邻居 $NB_{i1}=neigbour(x_i)$是其异类最近邻,而 $NB_{i2}=neigbour(x_i)$为同类最近邻.$d_{ij}=\| x_i, NB_{ij}\|$ 为 x_i 与 NB_{ij}的欧式距离.则由样本序列$\{x_i, NB_{i1}, NB_{i2}, \cdots, NB_{ik}\}$以及距离序列$\{d_{i1}, d_{i2}, \cdots, d_{ik}\}$交叉构成的链称为少数类样本单边选择链.

OSL 算法描述如下:

算法 1.

输入:样本集 Ω　//$X\cup Y=\Omega$, $X\cap Y=\varnothing$,其中,X 与 Y 分别表示少数类样本与多数类样本集

输出:少数类边界样本集合 Ω'

Algorithm: OSL

FOR($i=1, i\leqslant n, i++$)

　Findlink(x_i)　　//x_i 是 X 中第 i 个样本

ENDFOR

Algorithm: *Findlink*(x_i)

(1)初始化:$L_1=x_i$;//$X\cup Y=\Omega$, $X\cap Y=\varnothing$,其中,$\forall x_i\in X$ 与 $\forall y_j\in Y$ 分别表示少数类样本与多数类样本

(2)找到 x_i 的异类最近邻居 $neigbour(x_i)$

(3)$y_{++j}=neigbour(x_i)$

(4)$L_{++m}=y_j$

(5)找到 y_j 的异类最近邻居 $neigbour(y_j)$

(6)$x_{++i}=neigbour(y_j)$

(7)$L_{++m}=x_i$

(8)重复(2)-(7),直到 $neigbour(x_i)=neigbour(x_{i-1})$

少数类样本选择链中 d_{ij}以非递增次序排列,终止于 x_i,由所有满足此条件的 x_i 构成了少数类边界数据集Ω'.

3.3　基于样本分布密度的过采样算法

容易推知,在样本空间中,如某一样本周围的邻居样本很多,则其周围样本的分布密度较大.对于决策树分类器而言,产生的叶子节点(规则)就较多,而越多的叶子节点容易产生过拟合,故该节点与其近邻新生成的样本不宜过量.基于此,本文利用 OSL 和少数类样本分布密度确定利用 SMOTE 方法产生新样本数量的新方法 OSLDD-SMOTE.

首先给出几个定义:

定义 3　在数据集 Ω' 中,对 $\forall x\in\Omega'$,如 $y_i\in\Omega'$ 是 x 的第 i 个近邻,$1\leqslant i\leqslant k$,则所有 x 与其 k 个最近邻居构成的密度矩阵可表示为 $DM=(d_{ij})_{t*k}$,其中,t 为样本个数,k 为近邻数.令 $z_i=\sum_{j=1}^{k}\frac{1}{d_{ij}}$,其中,$k$ 为近邻个数.对 z_i 归一化,则 $density(z_i)=\frac{z_i}{\sum_{i=1}^{n'}z_i}$,其中,$n'$为$\Omega'$的样本个数.

定义 4　在数据集 Ω' 中,由所有的 $density(x)$构成的序列 L 称作密度序列.如果不存在这样的 x_i, x_j,对 $\forall x_i, x_j\in\Omega'$, $i\leqslant j$,满足 $density(x_i)>density(x_j)$,则称 L 为有序密度序列(Sequential Density Link,SDL).

性质 1　L 中所有值之和等于 1.

性质 2　L 中所有元素的次序体现了其密度大小程度.

性质 3　在数据集 Ω' 中,每次 SMOTE 生成新样本数量约为 $k*T$,k 为近邻数,T 为 L 中样本数目.

证明 在有序密度序列 L 中,每次 SMOTE 过程中,对 $\forall x_i \in L$,经过 SMOTE 生成新样本为$\lfloor(1-density(x_i))*k\rfloor$,其中,$k$ 为近邻数.则在 L 中所有样本共生成新样本个数为

$$Total=\sum_{i=1}^{T}\lfloor(1-density(x_i))*k\rfloor<k*\sum_{i=1}^{T}(1-density(x_i))<k*(T-1)<k*T.$$ 证毕.

OSLDD-SMOTE 算法如下:

算法 2

输入:边界少数类样本集 Ω',近邻数目 k

输出:合成新少数类样本

步骤 1 计算样本分布密度 *density*

利用 *cosine* 计算公式得到样本 x 与 y 距离

$$sim(x,y)=\frac{\sum_{i=1}^{l}w_{x_i}\cdot w_{y_i}}{\sqrt{\sum_{i=1}^{l}(w_{x_i})^2\cdot\sum_{i=1}^{l}(w_{y_i})^2}} \tag{2}$$

$$d_{xy}=dist(x,y)=1-sim(x,y) \tag{3}$$

其中,l 是特征向量维度,w_{x_i}和w_{y_j}分别是样本 x 和 y 的第 i 维和第 j 维特征向量的权值.

步骤 2 利用 SDL-SMOTE 生成新样本

输入:SDL 长度 L,最近邻数 k

输出:合成样本

1. 计算每个样本密度 $density(x_i)$
2. 计算 L 中每个样本的 SMOTE 因子,即生成新样本数目:
   $SMOTEfactor(x_i)=\lfloor(1-density(x_i))*k\rfloor$
3. WHILE $SMOTEfactor(x_i)\neq 0$
4. FOR $attr\leftarrow 1$ to numattrs
5. Compute: $dif=Instance[nnarray[nn]][attr]-Instance[i][attr]$
6. Compute: gap = random number between 0 and 1
7. $Synthetic[newindex][attr]=Instance[i][attr]+gap*dif$
8. ENDFOR
9. $nn++$
10. newindex++
11. $SMOTEfactor(x_i)=SMOTEfactor(x_i)-1$
12. ENDWHILE

3.4 OSLDD-SMOTE 特点

与 SMOTE 及其一系列改进算法相比,OSLDD-SMOTE 的特点主要表现在:

(1)噪音数据预处理方案不同.不是将可疑样本一概删除,而是区别对待,提出最近邻决策机制对可疑样本进行处理.

(2)重采样样本选择方法不同.不是对所有少数类样本利用 SMOTE 方法合成新样本,本文利用单边选择链 OSL 选择出位于分类边界的少数类样本,然后利用 SMOTE 合成新样本.该策略既提高了 SMOTE 过程的时间效率,又合理选择了 SMOTE 对象,从而丰富和提升了经典少数类样本的重采样方法.

(3)SMOTE 决策机制不同.每次 SMOTE 过程后,重新形成样本密度矩阵,使得样本分布密度具有动态特性,较好地反映了少数类样本实时分布状况,从而为 SMOTE 决策提供强有力支持.

(4)新样本生成数量不同.根据样本分布密度决定 SMOTE 过程和产生新样本的数量,样本分布密度越大,生成新样本越少,反之亦然.

由(2)(3)(4)形成的样本合成粒度递进原则,既充分利用了 SMOTE 技术生成适量少数类样本,使样本类间数量趋于平衡,又考虑到了少数类内部样本的分类特征,从而最大程度避免了过拟合,增强了分类器的泛化能力.

4 实验验证

4.1 数据集

为验证本文方法的有效性与普适性,我们选择了属性维度、样本数量及非平衡度差异较大的 8 个 UCI 数据集,详见表 1.其中,Attributes 是样本属性个数及类型;Size 为样本的总数量;Concept 是目标少数类,其余样本统归为多数类;Positive instances 和 Negative instances 分别是少数类和多数类样本的数量,Imbalance ratio 为非平衡度.实验前均去掉了丢失部分信息的样本.

4.2 性能指标

非平衡数据集两分类问题中,分类精度性能指标主要有:

$$TP\text{-}rate=Recall=\frac{TP}{TP+FN} \tag{4}$$

$$FP\text{-}rate=\frac{FP}{FP+TN} \tag{5}$$

$$TN-rate=\frac{TN}{TN+FP} \tag{6}$$

$$G\text{-}mean=\sqrt{TP\text{-}rate*TN\text{-}rate} \tag{7}$$

$$Precision=\frac{TP}{TP+FP} \tag{8}$$

$$F-measure=\frac{k*Precision*Recall}{Precision+Recall} \tag{9}$$

其中,TP 和 TN 分别为预测正确的正样本和负样本的数量,FP 表示原本属于负类的样本错误地预测为正类的样本数,FN 表示原本属于正类的样本错误地预测为负类的样本数.

一般情况下,常用准确率指标评价分类器的性能.但非平衡数据分布环境下,传统的准确率指标已不适用,以 Balance 数据集为例,如把所有样本均判为多数类,准确率为 92.16%,但此时对少数类样本而言毫无意义.因此,为综合考察 $TP-rate$ 和 $FP-rate$,采用受试者操作特征(Receiver Operating Characteristic,ROC)和 AUC[23]作为衡量少数类样本分类精度的指标.ROC 曲

线独立于数据集类间样本分布,对数据集的非平衡状态有很好的鲁棒性,因此可用于非平衡数据集分类器性能评价.ROC 曲线反映了当分类器参数变化时,正确率(*TP* – *rate*)与确定率(*FP* – *rate*)之间的关系.AUC 是 ROC 曲线下方区域的面积,通常,AUC 的值介于 0.5 到 1.0 之间,较大的 AUC 代表了较好的 performance,已被证明在非平衡数据集中,利用 AUC 更能有效地衡量分类器的性能[23].

表 1　实验所用 UCI 数据集

数据集	属性	样本数量	正样本类/负样本类	正样本数量	负样本数量	非平衡度
Abalone	1 N, 7 C	4177	Class 10/ Remainder	634	3543	5.59
Balance	4 C	625	Class 2/Remainder	49	576	11.75
Car	6 N	1728	Acc/Remainder	384	1344	3.50
Glass	10 C	214	Class 1/ Class 0	51	163	3.19
Prima	8 C	768	Class 1/ Class 0	268	500	1.86
Satimage	36 C	6435	4/Remainder	626	5809	9.28
Spect	23 N	267	Class 0/ Class 1	55	212	3.85
Vehicle	18 C	946	Van/Remainder	226	720	3.18

4.3　算法性能评估

实验运行配置如下:Intel Pentium 2.4GHz CPU,2.0GB内存,500GB 硬盘,操作系统为 Windows 2000 (Server),分类器采用 C4.5,利用十交叉验证方法.为研究 OSLDD-SMOTE 在不同样本分布下的运行时间,在 4.1 节介绍的 8 个数据集上当合成因子(N%)分别取自集合{50%,100%,200%,300%,400%,500%,600%,700%,800%,900%,1000%}时,进行十次十交叉验证计算平均值,结果如表 2 所示.其中,RANK(ORIGINAL)代表合成因子为 50% 时按运行时间排位位次,RANK(TERMINAL)代表合成因子为 1000% 时按时间复杂度排位位次,RANK(AVG)代表运行时间均值的排位位次.

表 2 表明,随着测试样本特征数量和样本规模的增加,OSLDD-SMOTE 的运行时间不断提高,具体而言,表现出三个特征:

第一,OSLDD-SMOTE 的运行时间随合成因子增大而增加:当合成因子为 50% 时,OSLDD-SMOTE 的运行时间在各个数据集上均为最小,当合成因子为 1000% 时,运行时间在各个数据集上均为最大.

第二,OSLDD-SMOTE 的运行时间随样本特征数量增大而增加:当合成因子为 50% 时,在 Glass 和 Spect 数据集上运行时间分别是 1.96s 和 1.58s,但由于 Spect 数据集的样本特征数量远高于 Glass 数据集,故随着合成度不断提高,OSLDD-SMOTE 在 Spect 数据集的运行时间提高程度更快,当合成因子为 1000% 时,在 Spect 和 Glass 数据集的运行时间分别是 5.4s 和 8.55s.

第三,OSLDD-SMOTE 运行时间与样本规模同向增加.总体而言,样本规模越大,运行时间越长.

表 2　运行时间(Seconds)

	Abalone	Balance	Car	Glass	Prima	Satimage	Spect	Vehicle
50%	54.72	3.83	8.51	1.96	4.14	305.28	1.58	9.36
100%	61.74	4.41	9.05	2.03	4.73	315.99	2.66	11.75
200%	62.95	4.59	10.26	2.12	5.67	325.62	3.65	12.92
300%	64.22	5.4	11.30	2.52	6.975	343.40	3.83	13.46
400%	71.01	5.81	12.96	3.15	8.37	360.86	4.05	14.45
500%	71.96	6.35	13.86	3.69	8.96	381.02	4.28	15.62
600%	76.55	7.16	14.86	3.96	10.17	401.85	4.41	16.11
700%	80.82	7.65	15.75	4.32	11.52	443.30	4.73	16.29
800%	86.81	8.06	16.56	4.73	12.51	482.85	6.39	17.42
900%	108.05	8.46	18.99	5.04	13.32	540.09	7.43	17.96
1000%	116.87	8.69	20.03	5.4	14.76	604.35	8.55	20.75
AVG	77.76	6.39	13.82	3.51	9.18	409.50	4.68	15.08
RANK(ORIGINAL)	7	3	5	2	4	8	1	6
RANK(TERMINAL)	7	3	5	1	4	8	2	6
RANK(AVG)	7	3	5	1	4	8	2	6

4.4　样本合成因子影响

为揭示不同的过采样率对分类器产生节点数目和 TP 性能指标的影响,我们在 Satimage 数据集上在保持原样本集分布不变前提下择取了 5000 个样本,综合比

较了 OSLDD-SMOTE 与 SMOTE[9]、Borderline-SMOTE[11]及 Surrounding-SMOTE(S-SM OTE)[24]. 对比实验采用 C4.5 分类器,最近邻居数 k 分别取 1,3,5,7,9. 利用十交叉验证方法,将数据集分成十份,轮流将其中 9 份做训练 1 份做测试,共进行 10 次 10 交叉验证求得均值,作为对算法精度的估计,结果如图 3 和图 4 所示.

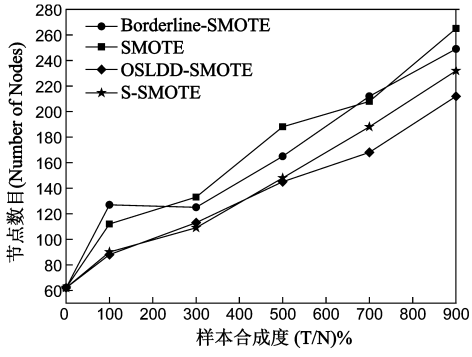

图3 样本合成度对节点数目影响

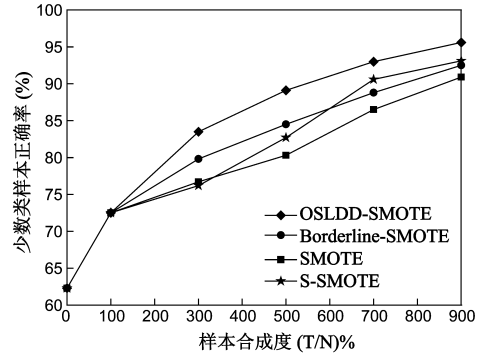

图4 样本合成度对少数类精度影响

由图 3 可知,随着合成少数类样本数量不断增多,分类器产生的节点数目也在不断增多;OSLDD-SMOTE 在四种方法中产生节点数目最少,最大可能避免了过拟合,进而提高分类器的泛化能力. 图 4 中,当横坐标为 0 时,表示没有合成样本,此刻四种方法得到相同结果. 随着样本合成度增大,SMOTE 方法在一定程度上提高了少数类样本精度,但性能最差;Borderline-SMOTE 与 S-SMOTE 性能互有优劣,整体优于 SMOTE 方法;OSLDD-SMOTE 则使少数类样本精度随着合成样本增多不断提升,且性能最优.

4.5 性能比较

为进一步衡量 OSLDD-SMOTE 的性能,我们在该节扩大了比较范围,使 OSLDD-SMOTE 与随机过采样(Random Oversampling, ROSMP)、SMOTE[9]、Borderline-SMOTE[11]、S-SMOTE[24]、MSYN[15]、ADASYN[25] 和 D-SMOTE[26]等一系列过采样算法在 8 个数据集上对 G-mean、F-measure 和 AUC 等多个性能进行了比较,结果如表 3 ~ 5 所示,表中数据均为进行十次十交叉验证计算的均值.

表 3 ~ 5 中,各个单元格值为利用 Wilcoxon 符号等级检验标准对 OSLDD-SMOTE 与其它方法的性能比较结果. 例如,某值用下划线标识,表示 OSLDD-SMOTE 方法优于该值对应的方法;如果某值用星号标识,表示 OSLDD-SMOTE 方法逊于该值对应的方法;如某值正常标识,表示根据 Wilcoxon 符号等级检验标准 OSLDD-SMOTE 方法与该值对应的方法相比,无明显差别. W/D/L 行对应值则表示 OSLDD-SMOTE 方法优于/等于/逊于其它各种方法的次数. 由以上分析可知,OSLDD-SMOTE 方法在多数数据集尤其在 Balance、Prima 和 Vehicle 几个较难分类数据集上[7]优势明显.

表 3 各种方法在 8 个数据集上的 G-mean 结果

Data set	OSLDD-SMOTE	Borderline-SMOTE	D-SMOTE	S-SMOTE	MSYN	ROSMP	SMOTE	ADASYN
Abalone	0.7408	0.7416	0.7323	0.7544 *	0.7412	0.3763	0.7207	0.7529 *
Balance	0.5845	0.5370	0.5165	0.5501	0.5619	0.2236	0.4219	0.5417
Car	0.9519	0.9452	0.9132	0.9263	0.9743 *	0.6999	0.9255	0.9280
Glass	0.7319	0.7047	0.7204	0.7255	0.7429 *	0.3873	0.6908	0.7003
Prima	0.7221	0.6933	0.6875	0.6937	0.7003	0.3873	0.6671	0.6809
Satimage	0.7419	0.7547 *	0.7514 *	0.7555 *	0.7528 *	0.3873	0.7208	0.7434
Spect	0.7083	0.7068	0.7053	0.6760	0.7091	0.3937	0.7068	0.7049
Vehicle	0.7749	0.7270	0.7219	0.7470	0.7344	0.6481	0.7239	0.7332
W/D/L	N/A	5/2/1	6/1/1	6/0/2	3/2/3	8/0/0	7/1/0	5/2/1

表 4　各种方法在 8 个数据集上的 F-measure 结果

Data set	OSLDD-SMOTE	Borderline-SMOTE	D-SMOTE	S-SMOTE	MSYN	ROSMP	SMOTE	ADASYN
Abalone	0.3820	0.3774	0.3756	0.3817	0.3823	0.2540	0.3790	0.3810
Balance	0.2101	0.1611	0.1509	0.1701	0.1523	0.0017	0.1569	0.1498
Car	0.9210	0.9110	0.9109	0.9101	0.9130	0.5010	0.9057	0.9198
Glass	0.8090	0.7802	0.7900	0.7960	0.8006	0.6709	0.7450	0.7520
Prima	0.6501	0.6302	0.6400	0.6460	0.6516	0.5009	0.6245	0.6320
Satimage	0.8090	0.8142 *	0.8150 *	0.8156 *	0.8144 *	0.6709	0.7450	0.7520
Spect	0.4802	0.4568	0.4689	0.4757	0.4750	0.4445	0.4365	0.4556
Vehicle	0.6307	0.6205	0.6063	0.6257	0.6301	0.5547	0.6110	0.6166
W/D/L	N/A	7/0/1	7/0/1	6/1/1	4/3/1	8/0/0	7/1/0	6/2/0

表 5　各种方法在 8 个数据集上的 AUC 结果

Data set	OSLDD-SMOTE	Borderline-SMOTE	D-SMOTE	S-SMOTE	MSYN	ROSMP	SMOTE	ADASYN
Abalone	0.8091	0.7806	0.7904	0.7968	0.8193 *	0.6710	0.7450	0.8155 *
Balance	0.6210	0.5717	0.5719	0.5500	0.5777	0.4392	0.5487	0.5543
Car	0.9916	0.9914	0.9909	0.9920	0.9913	0.7149	0.9750	0.9917
Glass	0.8100	0.8126	0.8237 *	0.8006	0.8224 *	0.7298	0.7630	0.8114
Prima	0.7558	0.7581	0.7428	0.7432	0.7500	0.7208	0.744	0.7555
Satimage	0.8901	0.8679	0.8880	0.8910	0.9108 *	0.7380	0.8970 *	0.8649
Spect	0.7526	0.7481	0.7225	0.7500	0.7430	0.6885	0.7531	0.7229
Vehicle	0.8670	0.8634	0.8634	0.8616	0.8601	0.8224	0.8561	0.8610
W/D/L	N/A	4/4/0	5/2/1	5/3/0	4/1/3	8/0/0	6/1/1	4/3/1

5　结论

非平衡数据挖掘已成为机器学习领域的前沿课题,引起了越来越多学者的关注与研究.为解决“提高正样本分类精度”这一非平衡数据分类中的瓶颈问题,在我们提出的单边选择链和样本分布密度基础上,本文提出了改进 SMOTE 非平衡数据挖掘方法 OSLDD-SMOTE.OSLDD-SMOTE 基于少数类样本空间分布状况,核心思想是把分类模型视为密度分布空间的知识发现系统,研究其在非平衡样本分布下的潜在规律(机理),改变噪音数据处理策略、SMOTE 技术的固有样本选择方案、SMOTE 决策机制,形成了全新的基于非平衡数据分布环境下的分类模型,从而丰富和提升了经典重采样方法.通过多性能指标的测试及与其它算法对比实验,说明把样本遴选机制和分布密度融入到非平衡数据挖掘过程中去,在解决“提高正样本分类精度”问题上确实有很好的效果.我们认为,基于单边选择链和样本分布密度重采样方法为非平衡数据分类进行理论分析和实验论证提供了强有力的工具,同时对非平衡数据挖掘的理论研究具有重要意义,有可能对新一代重采样技术的发展起到重要的推动作用.

未来的工作包括进一步研究 OSLDD-SMOTE 在高维大规模数据集[27,28]上如何更有效地解决运行时间和内存消耗的瓶颈问题;本文研究主要着眼于两分类问题,但真实数据集很多是多类别的,如何以最小代价把本文方法移植到多分类数据集分类过程需要进一步分析与研究.

参考文献

[1] Chan P K, Stolfo S J. Toward scalable learning with nonuniform class and cost distributions: A case study in credit card fraud detection[A]. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining[C]. New York: AAAI, 1998.164-168.

[2] Phua C, Alahakoon D, Lee V. Minority report in fraud detection: Classification of skewed data[J]. SIGKDD Explore, 2004, 6(1): 50 – 59.

[3] Lewis D, Gale W. A sequential algorithm for training text classifiers[A]. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Dublin: ACM, 1994. 3 – 12.

[4] Turney P D. Learning algorithms for keyphrase extraction[J]. Information Retrieval, 2000, 2(4): 303 – 336.

[5] Ling C X, Li C. Data mining for direct marketing: Problems and solutions[A]. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining[C]. New York: AAAI, 1998. 73 – 79.

[6] Japkowicz N. The class imbalance problem: Significance and strategies[A]. Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning[C]. Las Vegas: AAAI, 2000. 111 – 117.

[7] Liu Xu-Ying, Wu Jian-xin, Zhou Zhi-Hua. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man and Cybernetics, 2009, 39(2): 539 – 550.

[8] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63 – 77.

[9] Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE-synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321 – 357.

[10] 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(12A):22 – 26.
Yang Zhi-Ming, Qiao Li-Yan, Peng Xi-Yuan. Research on datamining method for imbalanced dataset based on improved SMOTE[J]. Acta Electronica Sinica, 2007, 35(12A): 22 – 26. (in Chinese)

[11] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[A]. Proceedings of the 2005 international conference on Advances in Intelligent Computing[C]. Berlin, Heidelberg: Springer-Verlag, 2005, 3644. 878 – 887.

[12] 曾志强,吴群,廖备水,高济.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.
Zeng Zhi-Qiang, Wu Qun, Liao Bei-Shui, Gao Ji. A classfication method for imbalance data set based on kernel SMOTE[J]. Acta Electronica Sinica, 2009, 37(11): 2489 – 2495. (in Chinese)

[13] 李正欣,赵林度.基于 SMOTEBoost 的非均衡数据集 SVM 分类器[J].系统工程,2008,26(5):116-119.
Li Zheng-Xin, Zhao Lin-Du. A SVM classifier for imbalanced datasets based on SMOTEBoost[J]. Systems Engineering, 2008, 26(5): 116-119. (in Chinese)

[14] 毕华,梁洪力,王珏.重采样方法与机器学习[J].计算机学报,2009,32(5):862 – 877.
Bi Hua, Liang Hong-Li, Wang Yu. Resamplingmethod and machine learning[J]. Chinese Journal of Computers, 2009, 32(5): 862 – 877. (in Chinese)

[15] Fan X N, Tang K, Weise T. Margin-based over-sampling method for learning from imbalanced datasets[A]. Proceedings of the 15th Pacific – Asia Conference on Knowledge Discovery and Data Mining[C]. Berlin: Springer, 2011. 24 – 27.

[16] 欧阳震诤,罗建书,胡东敏,吴泉源.一种不平衡数据流集成分类模型[J].电子学报,2010,38(1):184 – 189.
OUYANG Zhen-zheng, LUO Jian-shu, HU Dong-min, WU Quan-yuan. an ensemble classifier framework for mining imbalanced data streams[J]. Acta Electronica Sinica, 2010, 38(1): 184 – 189. (in Chinese)

[17] 周志华,陈世福.神经网络集成[J].计算机学报,2002,25(1):1 – 8.
Zhou Zhi-Hua, Chen Shi-Fu. Neural network ensemble[J]. Chinese Journal of Computers, 2002, 25(1): 1 – 8. (in Chinese)

[18] Zhou Z H, Jiang Y. MeV4dical diagnosis with C4.5 rule preceded by artificial neural network ensemble[J]. IEEE Transactions on Information Technology in Biomedicine, 2003, 7(1): 37 – 42.

[19] Zhou Zhi-Hua, Jiang Yuan, Chen Shi-fu. Extracting symbolic rules from trained neural network ensembles[J]. AI Communications, 2003, 16(1): 3 – 15.

[20] Brodley C E, Friedl M A. Identifying mislabeled training data[J]. Journal of Artificial Intelligence Research, 1999, 11(1): 131 – 167.

[21] Muhlenbach F, Lallich S, Zighed D. Identifying and handling mislabelled instances[J]. Journal of Intelligent Information Systems, 2004, 22(1): 89 – 109.

[22] Gamberger D, Lavrac N, Dzeroski S. Noise elimination in inductive concept learning: A case study in medical diagnosis[A]. Proceedings of the 7th International Workshop on Algorithmic Learning Theory[C]. Berlin, Heidelberg: Springer-Verlag, 1996, 1160. 199 – 212.

[23] Fawcett T. ROC graphs: Notes and practical considerations for data mining researchers[R]. USA: Technical Report HP Labs, 2003.

[24] Garcha V, Sanchez J S, Mollineda R A. On the use of surrounding neighbors for synthetic over-sampling of the minority class[A]. Proceedings of 8th WSEAS International Conference on Simulation, Modeling and Optimization[C]. Santander: WSEAS Press, 2008. 23-25.

[25] He H, Bai Y, Garcia E A, Li S. ADASYN: Adaptive synthetic

sampling approach for imbalanced learning[A]. Proceedings of 2008 IEEE International Joint Conference on Neural Networks[C]. Hong Kong: IEEE Press, 2008. 1322 – 1328.

[26] Calleja J D L, Fuentes O. A distance-based over-sampling method for learning from imbalanced data sets[A]. Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference[C]. Florida: AAAI Press, 2007. 634 – 635.

[27] 杨炳儒,谢永红,侯伟,周谆.基于复合金字塔模型的蛋白质二级结构预测系统的研究[J].科学通报,2009,54(21):3311 – 3319.
Yang Bing-Ru, Xie Yong-Hong, Hou Wei, Zhou Zhun. A novel protein secondary structure prediction system based on compound pyramid model[J]. Chinese Science Bulletin, 2009, 54(21): 3311 – 3319. (in Chinese)

[28] Yang B R, Hou W, Zhou Z, Quan HB. KAAPRO: An approach of protein secondary structure prediction based on KDD* in the compound pyramid prediction model[J]. Expert Systems With Applications, 2009, 36(1): 9000 – 9006.

作者简介

翟　云　男,1979 年生,山东青州人.博士,讲师,主要研究领域为知识发现、政务智能.
E-mail: yunfei_2001_1@aliyun.com

王树鹏(通信作者)　男,1980 年生,山东济南人.博士后,研究方向为海量数据存储、数据灾备.
E-mail: wangshupeng@iie.ac.cn

马　楠　女,1978 年生,北京市人.讲师,研究方向为知识发现与模糊认知图.

杨炳儒　男,1943 年生,天津市人.教授,博士生导师,CCF 高级会员,主要研究方向为知识发现与智能系统、柔性建模与集成技术.

张德政　男,1964 年生,山东青岛人.教授,博士生导师,主要研究方向为知识发现.