

基于改进的 SMOTE 和 RST 的新型混合重取样算法

谷 琼, 袁 磊, 宁 彬, 吴 钊, 华 丽, 李文新

(湖北文理学院 数学与计算机科学学院, 湖北 襄阳 441053)

摘 要: 分类预测是数据挖掘和机器学习的重要任务之一, 非均衡数据广泛存在于真实世界的分类问题中. 本文提出一种新的解决非均衡数据集的预处理方法(ImSMOTE-RSTR.). 通过改进的 SMOTE 方法创建新的人工合成少数类实例, 并在此基础上应用基于粗糙集理论的子集下近似技术对训练集进行清理, 该算法被验证得到较理想的结果.

关键词: 分类; 非均衡数据集; 混合重取样; SMOTE; 粗糙集理论

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2012)09-0083-04

A Novel Hybrid Resampling Algorithm Based on Improved SMOTE and Rough Sets Theory

GU Qiong, YUAN Lei, NING Bin, WU Zhao, HUA Li, LI Wen-xin

(School of Mathematics and Computer Science, Hubei University of Arts and Science, Xiangyang 441053, China)

Abstract: Classification forecasting is an important work of data mining and machine learning, imbalanced data is a common problem in classification in most real domains. This paper proposes a new hybrid method for preprocessing imbalanced datasets (ImSMOTE-RSTR.) through the construction of new samples, using the improved Synthetic Minority Oversampling Technique together with the application of an editing technique based on the Rough Set Theory and the lower approximation of a subset. The proposed method has been validated by an experimental study showing good results as the learning algorithm.

Key words: classification; imbalanced dataset; hybrid resampling; SMOTE; rough set theory

1 引言

真实世界广泛存在非均衡数据的分类预测, 如欺诈检测、故障监测、医疗诊断等, 传统分类算法倾向于多数类实例而忽视少数类样本, 从而造成少数类样本分类效果较差. 国内外学者提出对数据预处理和改进分类算法等多种解决方案来解决非均衡数据集分类问题^[1]. 数据预处理通过对训练数据重取样改变数据类分布从而降低其非均衡程度, 改进分类算法是通过已有的算法进行改进使其更有利于少数类分类, 如对各类样本分设不同的权值或是调节样本之间的代价函数等措施来改进. 本文提出一

种新的数据预处理重取样方式, 通过改进的 SMOTE 方法创建新的人工合成少数类实例, 并在此基础上应用粗糙集理论子集的下近似的采样得到最终的训练数据集, 选取 9 个典型的非均衡数据集进行仿真对比实验. 验证结果显示, 本文算法在不降低总体分类性能的前提下, 提高少数类分类性能上具有明显优势.

2 ImSMOTE-RSTR. 算法

2.1 SMOTE 方法及其存在的问题

SMOTE^[2]方法是 2002 年由 Chawla 提出的一种“人工合成少数类样本”的过取样方法. 该方法基

收稿日期: 2012-02-08; **修回日期:** 2012-03-20

基金项目: 国家自然科学基金项目(61172084); 国家“八六三”计划项目(2009AA12Z117); 湖北省自然科学基金项目(2010CDB05201); 湖北省教育厅中青年基金项目(Q20112604)

于少数类实例的近邻也是少数类这一前提下设计实现的,没有考虑数据集中少数类样本附近可能存在多数类样本的真实分布特性,其近邻的选择策略存在一定盲目性,并且数据集有时既包含离散的类型型的文本属性,又包含连续的数值型属性,但 SMOTE 算法只能对数值型数据进行再造生成新的合成样本插值,而对于含有类别属性的混合型数据集则无能为力,原有的办法无法创造新的插值。

2.2 ImSMOTE 算法基本思想

为改进 SMOTE 算法在人工生成少数类实例的过程中,存在一定的盲目性和只能人工生成单一数值型属性的缺陷,文中设计了根据不同样本集内部的真实分布特性,自适应选择调整少数类的近邻,在近邻候选集合中选择距离少数类实例较近的少数类实例参与生成新样本,去除部分距离较远的少数类样本,允许部分距离较近的多数类样本参与到新样本的生成过程中,并根据近邻的不同情况分别采取不同的生成方法,最终达到控制少数类样本近邻区域、提高合成少数类样本质量的目的。

图 1 说明 ImSMOTE 算法的处理过程,在确定少数类样本 x_1 的 5 个近邻过程中,按照 ImSMOTE 思想,确定 x_1 候选近邻为 $\{x_2, x_3, x_4, y_1, y_2\}$,再对每个候选近邻分别产生新的少数类合成样本,生成新的合成样本时,对于 x_1 的少数类近邻,随机数选择范围在 $[0, 1]$ 之间,对于 x_1 的多数类近邻,随机数选择在 $[0, 0.5]$ 之间,这样可使生成的少数类合成新样本大多分布在 x_1 周围,基本保持了原始样本分布。

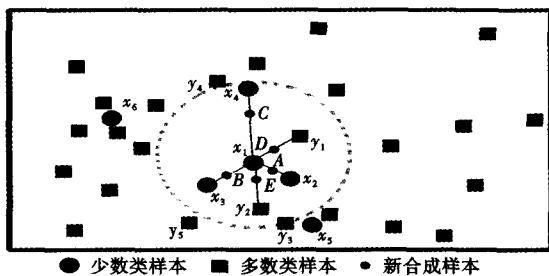


图 1 ImSMOTE 算法的处理过程

2.3 粗糙集理论

Pawlak 教授提出的粗糙集理论 (Rough Set Theory, RST)^[3] 能定量分析处理不精确、非完整信息,且其概念简单、计算简便已经在机器学习等领域广泛应用。由于 RST 无需提供所需处理数据之外的先验信息,对问题不确定性的描述和处理较客观,与概率论、模糊数学等其他处理不确定问题的方法具

有很强的互补性^[4-6]。

经典的 RST 创建一对近似算子,给定一个论域 U, R 为 U 上的一组等效关系,将 U 划分为互不相交的基本等效类,二元对 $K=(U, R)$ 构成一个近似空间. X 为 U 的一个子集, a 为 U 的一个对象, $[a]_R$ 表示所有与 a 不可分辨的对象所组成的集合. 当 X 表示成基本等效类组成的并集时,集合 X 关于 R 的上、下近似的定义分别见式(1)和(2),集合 X 的边界区 $BN(X)$ 见式(3),如果上下近似相等,则是精确集合,否则是粗糙集,利用上下近似及边界区可刻画一个不能精确定义的集合的近似特性^[7]。

$$R^*(X) = \{a \in U : [a]_R \cap X \neq \emptyset\} \quad (1)$$

$$R_*(X) = \{a \in U : [a]_R \subseteq X\} \quad (2)$$

$$BN(X) = R^*(X) - R_*(X) \quad (3)$$

2.4 ImSMOTE-RSTR.

为解决非均衡数据集存在的问题,文中设计一种混合重取样算法 ImSMOTE-RSTR., 该算法分两个阶段执行:首先用我们改进的 SMOTE 算法创建新的少数类实例,其次在人工合成少数类实例的基础上,应用 RST 技术下近似(R_*)对所合成的新的少数类实例的数据进行清理,去除冗余和噪声数据。

算法的具体步骤如下:

① 用改进的 SMOTE 算法自适应人工重取样,生成少数类实例;

② 将所生成的少数类实例加入到原始训练集;

③ 创建一个包含所有数据集的相似矩阵;

用 SimilarityMatrix 来确定两个实例 x_i 和 x_j 的相似程度,定义如下:

$$\text{SimilarityMatrix}(i, j) = \frac{\sum_{k=1}^n w_k * \delta_k(x_{ik}, x_{jk})}{M} \quad (4)$$

其中, n 是实例的属性数, w_k 表示 k 属性的权值, x_{ik} 和 x_{jk} 分别代表 i 和 j 实例属性 k 的值, δ_k 代表属性 k 的比较函数, M 表示具有等效关系的属性个数, B 表示具有等效关系的属性集. 某个属性权值定义如下:

$$w_k = \begin{cases} 1 & \text{如果 } k \in B \\ 0 & \text{否则} \end{cases} \quad (5)$$

δ_k 计算离散属性的方式:

$$\delta_k(x_{ik}, x_{jk}) = \begin{cases} 1 & \text{如果 } x_{ik} = x_{jk} \\ 0 & \text{否则} \end{cases} \quad (6)$$

δ_k 计算连续属性的方式如式(7), $\max A_k$ 和

$\min A_k$ 分别代表属性 k 的极值。

$$\delta_k(x_{ik}, x_{jk}) = 1 - \frac{|x_{ik} - x_{jk}|}{\max A_k - \min A_k} \quad (7)$$

④ 分析人工新合成的数据是否属于下近似,如果相似值低于所给定的阈值,意味着数据集里没有近似的元素,可以将它们添加到最后的“训练集”;

⑤ 检查由 SMOTE 算法生成的所有的新的人工生成的少数类实例与原始训练集之间的相似性,直到全部检查为止。

设置 similarityValue 参数值时要考虑它用来确定所包含的实例上下近似的相似程度。通过去除训练集中部分人工生成的不一致的少数类实例,在最后的训练集中,加入与多数类实例并无相似之处的属于下近似的人工合成的少数类实例。

ImSMOTE-RSTR* 算法的伪代码如图 2 所示。

Algorithm Steps

```
(1) Create syntheticInstance[] // 用改进的 SMOTE 算法创建一个合成的少数类实例集直到训练集基本均衡
(2) Create resultSet // 创建包括初始实例的结果集
(3) Construct the similarityMatrix // 创建相似矩阵,人工合成的少数类实例放在行,多数类实例放在列,用公式(4)可以计算实例  $i$  和  $j$  的相似程度
(4) Count the number of similar examples in the negative class
// 对每个人工新合成少数实例计算其在负类相似实例的个数、nmin-syn: 人工生成少数类实例的个数、nmax: 多数类实例个数 similarityValue: 0.4
While(resultSet is empty) & (similarityValue ≤ 0.9)
    for  $i \rightarrow 1$  to nmin-syn
        for  $j \rightarrow 1$  to nmax
            if(similarityMatrix( $i, j$ ) > similarityValue)
                then cont[ $i$ ]++
            endfor
        if cont[ $i$ ] = 0 then insert syntheticInstance[ $i$ ] in resultSet
        // 实例是下近似,就插入人工合成实例到最后的训练集
    endfor
    similarityValue = similarityValue + 0.05
endwhile
(5) obtain last resultSet
// 如果不再有了下近似实例,即所有的合成实例同其他的少数类实例类似,将剩下的所有人工合成的实例加入 resultSet
```

图2 ImSMOTE-RSTR* 算法伪代码

3 仿真实验

3.1 实验数据集和参数设置

为验证 ImSMOTE-RSTR* 算法的有效性,从 UCI 标准数据集中选择 9 个非均衡数据集 Pima Indians Diabetes、German Credit Data、Haberman's Survival、Vehicle Silhouettes、new-thyroid、Ecoli、

Landsat Satellite、Flags、Glass Identification 分别与文献[8]中的 CNN、OSS、Tomek Links、NCL、Smote、SMOTE-TomekLinks、SMOTE-ENN 等经典的重取样算法分类结果进行比较,上述重取样算法的详细介绍及 9 个非均衡数据集的相关信息见文献[8]。

实验中,ImSMOTE-RSTR* 算法的参数设置如下:近邻数设为 5,近邻函数选取欧氏距离,多数类与少数类类分布调整为 1:1。对原始数据集的所有重取样算法均在 Weka 平台上选用 J48 算法为基准算法,算法参数选择默认参数。选用 AUC 做为算法评价指标,所有数据集均采用十折交叉验证进行分类性能的评估,整个实验重复 10 次,将此基础上 AUC 平均值作为最终结果进行分类性能比较。

3.2 实验结果与分析

将文中算法 ImSMOTE-RSTR* 分别对 9 个数据集进行重取样之后,所得到的少数类 AUC 值分别为 84.46、83.68、86.21、98.96、98.91、96.32、95.43、87.45、93.58、91.67。将 ImSmote-RSTR* 重取样后的分类 AUC 值与文献[8]的重取样算法进行比较,其中有 6 个数据的分类结果处于领先的位置,有 5 次排在首位,1 次同其他 2 个算法排在并列第一位,跟其他方法相比获得较好的性能,并且基本经过重取样后的分类结果普遍优于未经过处理的原始数据集的分类结果。

表 1 显示了每种算法同其他算法比较的结果,获得最高 AUC 值的次数,Test 行每个单元格有两个数字。第一个数字代表该算法绝对胜过其他算法的次数,第二个数字表示该算法与排第一的算法并列的次数(相同的 AUC 值)。表 2 为每个选定的数据集分别就原始数据集和经过八种不同类型的重取样算法之后的 AUC 值的排名。通过排名,可以看到,ImSmote-RSTR* 算法有 5 次排在首位,1 次同其他 2 个算法排在并列第一位,1 次排在第 3 位,1 次排在第 4 位。

为了直观地显示实验结果,将 9 个非均衡数据

表1 重取样算法比较

	Original	CNN	OSS
Test	0/0	0/0	0/0
	Tomek Links	NCL	Smote
Test	0/0	0/0	1/1
	Smote-Tomek Links	Smote-ENN	ImSMOTE-RSTR*.
Test	0/1	2/0	5/1

表 2 各个数据集重取样算法测试性能排序

Data set	第一	第二	第三	第四	第五	第六	第七	第八	第九
Pima Indians Diabetes	Smote	Smote-Tomek Links	ImSMOTE-RSTR*	Smote-ENN	Tomek Links	NCL	Original	CNN	OSS
German Credit Data	ImSMOTE-RSTR*	Smote-Tomek Links	Smote-ENN	Smote	CNN	OSS	Original	Tomek Links	NCL
Haberman's Survival	ImSMOTE-RSTR*	Smote-ENN	Smote-Tomek Links	Smote	NCL	Tomek Links	OSS	CNN	Original
Vehicle Silhouettes	Smote	Smote-Tomek Links	ImSMOTE-RSTR*	OSS	CNN	Original	Tomek Links	NCL	Smote-ENN
new-thyroid	Smote-ENN	Smote	Smote-Tomek Links	ImSMOTE-RSTR*	CNN	Original	Tomek Links	NCL	OSS
Ecoli	ImSMOTE-RSTR*	Smote-Tomek Links	Smote	Smote-ENN	NCL	Tomek Links	Original	OSS	CNN
Landsat Satellite	Smote-ENN	Smote	Smote-Tomek Links	ImSMOTE-RSTR*	NCL	Tomek Links	Original	OSS	CNN
Flags	ImSmote+RSTR*	Smote-ENN	Smote-Tomek Links	Smote	CNN	Original	OSS	Tomek Links	NCL
Glass Identification	ImSmote+RSTR*	Smote-ENN	NCL	Smote-Tomek Links	Smote	Original	Tomek Links	OSS	CNN

集的各种重取样方法的 AUC 值汇总后用图 3 显示. 可以看到 IMSmote-RSTR* 方法综合了 9 个数据集之后, 在图中占据最高的位置, 具有较明显的优势.

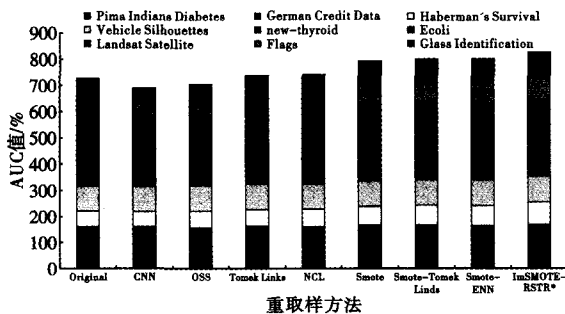


图 3 各种重取样方法在 9 个数据集上的总体性能比较

5 结束语

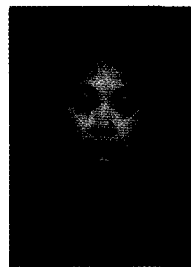
重取样方法是解决非均衡数据分类的一个有效途径. 文中提出一种解决非均衡数据分类的数据预处理方法. 该方法混合了过去非均衡数据集的过取样和欠取样技术, 其创新之处在于对新人工合成的少数类实例从质量上进行把关, 对新合成实例使用 RST 理论对少数类的下近似的一组实例进行评估, 只允许属于少数类下近似的人工实例加入最终的训练集. 通过标准数据集的仿真实验, 验证了本文算法具有可行性和有效性.

参考文献:

- [1] 赵会, 黄景涛, 谈书才. 最小二乘支持向量机的一种非均衡数据分类算法[J]. 微电子学与计算机, 2010, 27(4): 33-37.

- [2] 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(3): 321-357.
- [3] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science, 1982, 11(5): 314-356.
- [4] An A J, Stefanowski, Ramanna S, et al. Rough sets, fuzzy sets, data mining and granular computing[C]// RSFDGrC 2007. Toronto: Springer, 2007.
- [5] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009(7): 1229-1246.
- [6] 韩祯祥, 张琦, 文福拴. 粗糙集理论及其应用综述[J]. 控制理论与应用, 1999(2): 153-157.
- [7] Enislay R, Yailé C, Rafael B, et al. SMOTE-RSB*: a Hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory[J]. Knowledge and Information Systems, 2011(12): 1-21.
- [8] 谷琼. 面向非均衡数据集的机器学习及在地学数据处理中的应用[D]. 武汉: 中国地质大学 2009.

作者简介:



谷琼 女, 博士, 副教授, CCF 会员. 研究方向为数据挖掘和演化计算.