

引用格式:林静怀,刘治宇,李军良,等.面向不平衡数据分类的高维超球体过采样方法[J].微电子学与计算机,2021,38(5):65-72. [LIN J H, LIU Z Y, LI J L, et al. High-dimensional hypersphere oversampling method for imbalance classification[J]. Microelectronics & Computer, 2021, 38(5): 65-72.]

面向不平衡数据分类的高维超球体过采样方法

林静怀¹, 刘治宇², 李军良³, 高欣², 李泽科¹,
唐志军⁴, 余斯航⁴, 徐建航³

(1 国网福建省电力有限公司, 福建 福州 350003; 2 北京邮电大学 人工智能学院, 北京 100876;
3 南瑞集团(国网电力科学研究院)有限公司, 北京 100192;
4 国网福建省电力有限公司电力科学研究院, 福建 福州 350007)

摘 要: 在机器学习不平衡分类方法研究中, 由于多数类与少数类样本数量之间存在较大差异, 导致分类器易出现判定准确率低的问题. 以 SMOTE 为代表的一类过采样方法是处理该问题的一种有效手段. 该类方法在选定的线段中随机生成少数类新点来重新平衡数据集, 但存在忽略少数类样本在超维空间中分布多样性的缺陷. 本文提出一种面向不平衡数据分类的高维超球体过采样(HS-SMOTE)方法. 在少数类样本集上通过随机抽样获得需要平衡的样本数目, 在此基础上依次对每一样本通过欧氏距离选取其在少数类分布空间中的对应最近邻点, 以两点连线中点为球心在超维空间构建采样超球体, 在此区域内通过维度空间距离迭代随机生成所需的少数类新点, 在类别样本数据再平衡的基础上增加少数类样本的空间分布多样性. 在 15 组 KEEL 不平衡数据集上结合随机森林(RF)分类器开展了大量实验, 与 6 种典型过采样方法相比, 所提方法在 G-mean 以及 F1-score 指标上均有较好的表现, 并通过了 2 种统计学假设检验方法的有效性验证.

关键词: 机器学习; 不平衡分类; 过采样; 超球体

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2021)05—0065—08

High-dimensional hypersphere oversampling method for imbalance classification

LIN Jing-huai¹, LIU Zhi-yu², LI Jun-liang³, GAO Xin², LI Ze-ke¹,
TANG Zhi-jun⁴, YU Si-hang⁴, XU Jian-hang³

(1 State Grid Fujian Electric Power Corporation, Fuzhou 350003, China; 2 School of Artificial Intelligence,
Beijing University of Posts and Telecommunications, Beijing 100876, China;
3 NARI Group (State Grid Electric Power Research Institute) Corporation, Beijing 100192, China;
4 State Grid Fujian Electric Power Corporation Electric Power Research Institute, Fuzhou 350007, China)

Abstract: In the research of imbalance classification methods of machine learning, the classifier is prone to the problem of low judgment accuracy due to the large difference of the number between the majority class and the minority class. A class of oversampling methods represented by SMOTE are effective to deal with this problem. These types of methods randomly generate the minority new points in the selected line segment to rebalance the data set, but there is the defect of ignoring the diversity of minority samples in the super-dimensional space. A high-dimensional Hypersphere-SMOTE (HS-SMOTE) method is proposed for imbalanced data classification. On the minority sample set, the number of samples that need to be balanced is obtained by random sampling, and based on

this sampling, for each sample, its corresponding nearest neighbor is selected in turn through the Euclidean distance in the minority distribution space, and the midpoint of the two points is used for the center to construct a sampled hypersphere in the super-dimensional space. In this area, randomly generate the required minority new points through the dimensional space distance iteration, thus the spatial distribution diversity of the minority samples is increased on the basis of rebalancing the category sample data. A large number of experiments have been carried out on 15 sets of KEEL imbalanced data sets combining Random Forest (RF) classifiers. Compared with the 6 typical oversampling methods, the method proposed in the article has good performance on G-mean and F1-score indicators, and have passed the validity verification of two statistical hypothesis testing methods.

Key words: machine learning; imbalance classification; oversampling; hypersphere

1 引言

不平衡分类是一种在二类或多类数据集上的分类问题,具有类别之间存在明显分布差异的特点.该类问题通常出现在异常检测、故障诊断、人脸识别等领域中^[1-2],如何在数据集中存在类别样本数量不平衡的情况下开展分类器有效训练一直是机器学习中的一个重要且具有挑战性的问题.一般来说,在数据不平衡的情况下现有常规分类器所训练出的模型均表现不佳,分类结果往往偏向于多数类样本.这是由于分类器模型在训练时,因为少数样本的稀缺,模型并没有从少数样本获得足够的信息^[3].而在大多数不平衡问题中,所关注的却恰好是少数类样本的信息.比如在医疗诊断领域,患病人数相比健康人数可以被看作少数类,但从患者身上所获得的信息明显能给医疗人员正确识别患者提供更好的帮助.

一般来说,旨在处理不平衡分类问题的方法,可以分为两类:算法级方法和数据级方法^[4].不同于需要与特定分类器绑定的算法级方法,数据级方法在数据层面进行修改,生成的数据可以被普遍应用,因此在研究中,数据层面的方法更为通用^[5].而数据层面上,主要为重采样方法.重采样方法又主要分为欠采样方法和过采样方法^[6];欠采样是通过删除部分多数类样本使数据分布重新平衡,但如果少数类与多数类数量差距过大,强行使用欠采样方法会导致大量重要信息的丢失,可能使得模型效果反而降低^[7].而过采样方法则是针对少数类样本,通过增加少数类样本的数量使数据分布重新平衡,在遇到不平衡比率较大的数据时,过采样方法相比欠采样方法有着明显的优势^[3].

目前针对过采样方法国内外学者已经开展了深度的研究.2003年所提出的 SMOTE 方法^[8]作为数据级不平衡处理中的典型方法获得了较高的关注度,它通过在所选择的两个少数类连成的线段上创建新的实例来实现类别数据的重新平衡.目前,已经

提出了很多种基于 SMOTE 的变体方法,例如 Borderline-Smote^[9]、ADASYN^[10]等等,这些方法都对于原始 SMOTE 方法有了一些不同程度的改进.其中 Borderline-Smote 在采样过程中将少数类进一步划分,最后仅选择边界上的少数点来合成新样本.ADASYN 则是根据每个少数类 k 近邻中多数类的占比来给每个少数点赋予权重,并根据这个权重生成不同数量的新样本. Georgios 等人提出一种名为 Kmeans-SMOTE 的方法^[11],将聚类的思想与采样相结合,既关注了类别之间的不平衡,也对类别之内的不平衡做了处理.由于在少数群体区域内应用 SMOTE,这样更安全,不易产生噪声.尽管这些方法在预处理或是后处理阶段对 SMOTE 方法进行了不同的改进,但是它们仍旧是在少数类样本间的线段上随机生成新的少数点,并未打破直线的局限性,在处理高维数据时,可能会忽略少数类样本在高维空间分布的多样性,从而影响了后续分类结果的准确度.

针对以上问题,本文提出一种高维超球体过采样 (Hypersphere-SMOTE, HS-SMOTE) 方法,突破了线段的范围限制,通过在少数类高维空间里随机抽样出需平衡的样本数目,再依次寻找每个样本的最近邻的点用以构建超球体区域,并根据每个维度的空间距离在该区域内部生成所需的新少数点,最后完成对不平衡数据的重新平衡.为了验证了该方法的有效性,本文选取了 15 组 KEEL 上的不平衡数据集进行了一系列的实验分析.

2 HS-SMOTE 方法

2.1 SMOTE 算法

Chawla 等人提出的 SMOTE 算法可以说是使用最广泛、最经典的过采样方法之一^[8],在这之前所用的随机过采样算法只是简单的复制样本,很容易导致过拟合,降低模型的泛化能力. SMOTE 的主要思想基于 k 近邻算法,首先找到每个少数类样本 x_i

$\in S_{\min}$ ($S_{\min}$ 为少数类样本集) 所确定的 k 个最近邻少数点,并在这 k 个少数点中随机选择一点 x_i 用以生成新的少数类样本,即:

$$x_{\text{new}} = x_i + \text{rand}[0,1] \times (x_t - x_i) (t \in [1,k]) \quad (1)$$

公式(1)中 $\text{rand}[0,1]$ 生成一个介于0和1之间的随机数。

但作为经典方法,SMOTE 算法仍旧存在许多不足^[3]:不能精确控制合成新样本数量;不能对少数类样本进行区别性选择;样本混叠现象严重;同时使用 SMOTE 扩充样本集合时,由于生成点的范围比较局限,有可能生成的点相比原来的点并未有显著差异,从而使得分类器分类效果提升不大。

2.2 超球体过采样(HS-SMOTE)

本文提出一种 HS-SMOTE 方法,主要分为三个步骤:取点、划分区域、采样。首先,根据少数类与多数类的差值在少数类群集中通过随机抽样取得所需的样本;然后依次将所取样本点与离它欧氏距离最近的点用直线连接,将这些连线作为直径,连线的中点作为圆心,从而在空间中划分出多个超球体;同时由于在高维空间中,维数越高,超球体的体积在其外接超立方体中所占比例越小,如果每个维度上独立生成随机数,将会导致很难生成位于超球体内部的新少数点,为了保证所取点在超球体的内部,最后通过维度迭代调整每次取值的范围,从而生成新的少数点,最终完成不平衡数据集的重新平衡。图1为所提过采样方法的原理示意图。

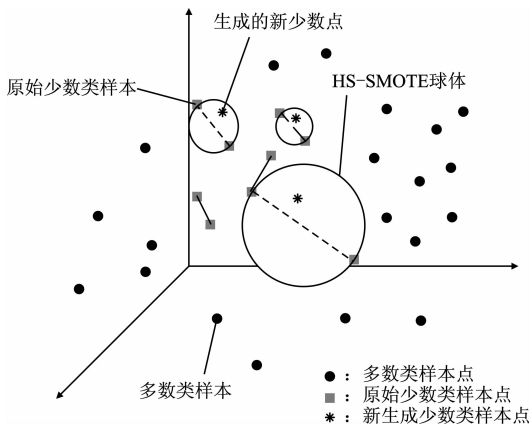


图1 超球体过采样方法原理示意图

本文所提出的高维超球体采样方法相比于在线段上采样的算法有着明显的优势。首先,构建的高维超球体涵盖了少数类样本数据的所有维度,相比于传统 SMOTE 算法取 k 近邻的点作线段,本方法仅

使用最近邻的点,减少了其他点所造成的影响;同时,对于构建的每个超球体来讲,其直径都与传统 SMOTE 方法所取到的直线对等,这样就相当于将 SMOTE 直线中每个少数点所对应的最短直线取出,并且将这条直线扩展到高维球体,这样新生成少数点的范围就不仅限于在一维直线上,而是在整个高维空间内,大大拓展了最近邻点与初始点之间采样取点的范围,突破了 SMOTE 算法在线段上的限制。其次,采样范围变大,同样也意味着减少了样本堆叠的概率,使新生成点分布更可能符合少数类在空间中的分布。最后,相比随机过采样方法,本文方法是通过生成新样本而不是单纯的复制样本,从而可以抵消过拟合的干扰。在后续的实验中,同样也证明了本文方法的有效性以及相比其他现有方法的优势。

超球体 SMOTE(HS-SMOTE)方法生成新样本过程如伪代码2所示。

伪代码2 HS-SMOTE 方法

输入:多数类样本集 S_{maj} ,少数类样本集 $S_{\min}$;

输出:过采样生成的少数类样本集 S_{omin}

1. For $i = 1, 2, \dots, k$: # k 为少数类与多数类样本数量差值
2. 在少数类实例中随机取一点 $x_i \in S_{\min}$,取离 x_i 欧氏距离最短的另一个点记为 $y_i (y_i \in S_{\min})$;
3. 取 x_i, y_i 两点连线的中心,记为球心 Q_i ;计算 x_i, y_i 两点之间欧氏距离的一半,记为半径 r_i ;
4. For $j = 1, 2, \dots, N$: # N 为输入少数类数据的维数

$$l_{ij} = \text{rand}[-1,1] \times \sqrt{r_i^2 - \sum_{n=0}^{j-1} l_{in}^2}$$

5. 式中 l_{ij} 为新生成的点在第 j 维度上离球心 Q_i 的欧氏距离,初始值 $l_{i0} = 0$; $\text{rand}[-1,1]$ 则生成一个介于-1和1之间的随机数。

6. end for

7. 循环后可得 $\vec{l}_i = (l_{i1}, l_{i2}, \dots, l_{iN})$, $\vec{l}_i$ 为球心指向新生成点的 N 维向量。由此可得新生成点 s_i 在高维空间每个维度的坐标,即:

$$\vec{s}_i = \vec{Q}_i + \vec{l}_i$$

8. 将 s_i 放入 S_{omin} 中

9. end for

10. 输出 S_{omin} :

$$S_{\text{omin}} = [s_1, s_2, \dots, s_k]$$

代码结束

3 实验部分

3.1 数据集的选择

为了测试 HS-SMOTE 方法的性能,本文使用了来自公开数据集网站 KEEL(<https://sci2s.ugr.es/keel/datasets.php>)上的 15 组二类不平衡数据集,这些数据集具有不同的特征,其中不平衡度从 1.86~195.96,覆盖广泛,样本数量也有多有少,从而保证了实验数据的多样性.表 1 按照不平衡度由低到高列出了这 15 个数据集:

表 1 数据集介绍

数据集名称	维数	样本总数	不平衡度
wisconsin	9	546	1.86
glass	9	171	2.05
haberman	3	275	2.82
vehicle	18	761	3.25
new-thyroid	5	193	5.22
ecoli	7	302	8.44
vowel	13	780	9.97
car	6	1 555	24.08
yeast-2vs4	8	1 336	28.04
winequality-red-4	11	1 279	29.45
poker-9vs7	10	195	31.5
abalone-21vs8	7	464	41.18
winequality-red-8vs6-7	8	684	47.9
poker-8-9vs6	10	1 188	58.4
page-blocks	10	4 924	195.96

3.2 对比方法选择

首先为了验证 HS-SMOTE 方法针对不平衡数据集的平衡效果,在实验过程中建立了两个模型,第一个模型使用原始未处理数据集进行训练,而第二个模型使用 HS-SMOTE 方法处理后的数据集进行.

实验中,将本文方法与一些典型的过采样方法放到一起做比较,对比方法分为两类,一类是 SMOTE 方法以及其 3 种改进方法: SMOTE^[8]、Borderline-SMOTE^[9]、ADASYN^[10] 以及 Kmeans-SMOTE(2018)^[11];还有一类是未使用 SMOTE 算法思路的典型过采样方法如: SOMO(2017)^[5]、AMSCO(2018)^[12]. 其中 SOMO 算法是基于人工神经网络模型自组织映射(SOM)的一种过采样方法,通过该算法划分数据集,最后生成新的少数点.而 AMSCO 方法同时使用了过采样与欠采样方法,并

将他们结合来平衡数据集.通过与这些方法的对比,能进一步证明本文所提出方法的有效性.

3.3 实验过程

为了评估 HS-SMOTE 方法的有效性,选取了机器学习分类领域典型分类器随机森林(RF)^[13]开展后续算法性能实验验证.随机森林是一个包含多个决策树的分类器,它可以处理大量的输入变数,并且拥有极高的学习效率.分类器的实现基于 Python 的 scikit-learn 库.

为了更好的评估算法性能,在实验过程采用了十折交叉验证来测试算法有效性.十折交叉验证是常用的测试方法,将数据集分成十份,轮流将其中 9 份作为训练数据,1 份作为测试数据,进行实验.每次实验都会得出一组结果,将 10 次结果的平均值作为对算法精度的估计.本文在每组数据上均进行了 5 次 10 折交叉验证,再求其均值,作为对算法准确性的估计.

同时为了公平起见,在与其他 SMOTE 方法以及改进方法的对比中,所取的 k 近邻值始终保持一致,分类器的参数在实验过程中也不进行二次更改.并且由于在对不平衡数据的预测过程中,用准确率来做评价指标不是很妥当,当少数类样本过少时,就算全部预测为多数类,也能有较高的准确率,所以实验使用针对少数类的 2 种评价指标: G -mean、 $F1$ -score^[14],将这 2 种指标作为评估本文所提方法及多种对比方法分类性能的度量参数. G -mean、 $F1$ -score 指标定义分别如公式(2)、(3)、(4)所示:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN},$$
$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2}$$

$$F1\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$
$$G\text{-mean} = \sqrt{\text{recall} \times \text{specificity}} \tag{3}$$

式中,TP 代表被正确预测为少数类的少数类样本数目,FP 代表被错误预测为多数类的少数类样本数目,FN 则代表被错误预测为少数类的多数类样本数目,TN 代表被正确预测为多数类的多数类样本数目.

整个实验在 Python 3.7.0 环境中运行.

4 实验结果分析

表 2 为本文所提的超球体过采样方法(HS-SMOTE)在 15 种不平衡数据集以及 RF 分类器上

与其他 6 种典型过采样方法之间的对比实验结果，缩写。
表中 G 、 F 分别为指标 G -mean 以及 $F1$ -score 的

表 2 在 RF 分类器上的实验结果

Datasets	指标	RF	SMOTE	Borderline-SMOTE	ADASYN	Kmeans-SMOTE	SOMO	AMSCO	HS-SMOTE
Wisconsin	G	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.97
	F	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.96
Glass	G	0.82	0.84	0.84	0.84	0.84	0.85	0.85	0.87
	F	0.77	0.78	0.78	0.77	0.79	0.81	0.79	0.82
Haberman	G	0.50	0.57	0.54	0.56	0.52	0.47	0.55	0.58
	F	0.34	0.42	0.38	0.41	0.37	0.32	0.40	0.42
Vehicle	G	0.95	0.97	0.96	0.96	0.96	0.96	0.96	0.97
	F	0.93	0.94	0.93	0.93	0.93	0.93	0.93	0.93
New-thyroid	G	0.97	0.95	0.95	0.94	0.96	0.95	0.94	0.97
	F	0.96	0.93	0.94	0.93	0.96	0.95	0.93	0.96
Ecoli	G	0.63	0.73	0.73	0.74	0.72	0.67	0.68	0.74
	F	0.52	0.56	0.57	0.57	0.57	0.59	0.55	0.61
Vowel	G	0.95	0.97	0.96	0.97	0.97	0.95	0.96	0.98
	F	0.93	0.96	0.95	0.95	0.95	0.94	0.95	0.96
Car	G	0.81	0.88	0.87	0.86	0.85	0.88	0.92	0.80
	F	0.74	0.78	0.76	0.76	0.74	0.82	0.79	0.73
Yeast-2vs4	G	0.81	0.87	0.84	0.86	0.84	0.80	0.85	0.84
	F	0.74	0.75	0.70	0.74	0.76	0.72	0.74	0.75
Winequality-red-4	G	0.00	0.32	0.22	0.34	0.34	0.00	0.06	0.46
	F	0.00	0.16	0.10	0.17	0.17	0.00	0.03	0.22
Poker-9vs7	G	0.14	0.34	0.34	0.34	0.27	0.14	0.18	0.43
	F	0.13	0.34	0.32	0.33	0.33	0.13	0.17	0.40
Abalone-21vs8	G	0.42	0.83	0.63	0.82	0.78	0.41	0.45	0.88
	F	0.33	0.58	0.51	0.56	0.55	0.36	0.37	0.65
Winequality-red-8_vs_6-7	G	0.20	0.35	0.20	0.35	0.33	0.00	0.20	0.37
	F	0.16	0.17	0.12	0.16	0.15	0.00	0.15	0.14
Poker-8-9vs6	G	0.00	0.62	0.53	0.64	0.56	0.30	0.61	0.61
	F	0.00	0.55	0.50	0.58	0.50	0.24	0.57	0.58
Page-blocks	G	0.77	0.91	0.88	0.89	0.87	0.76	0.92	1.00
	F	0.71	0.78	0.77	0.75	0.75	0.71	0.78	0.83
平均值	G	0.596 0	0.741 3	0.697 3	0.738 7	0.718 7	0.606 7	0.673 3	0.764 7
	F	0.548 0	0.644 0	0.613 6	0.638 0	0.632 0	0.564 7	0.606 7	0.664 0
Friedman	G	6.24	2.80	4.74	3.70	4.61	6.78	4.89	2.24
平均值等级	F	5.78	2.98	5.11	4.59	4.24	6.00	4.93	2.37

通过观察表 2,可以清楚的看到在大多数数据集上,HS-SMOTE 在 G -mean 以及 $F1$ -score 指标上有着明显的优势. 具体来说,在随机森林(RF)分类器上,本文方法在 G -mean 以及 $F1$ -score 指标上,分

别在 12、10 种数据集上排名第一,且平均值相比 6 种对比方法中性能最好方法分别提升了 3.15%、3.11%。
但仅从性能评估指标下的多组实验数据及其平

均值上的分析并不能充分说明所提方法的优劣,还需要进一步通过统计学上的校验.首先,基于随机森林(RF)分类器,分别对所提方法及 6 种对比过采样方法在 15 组不平衡数据集上的实验结果使用了 Friedman 检验方法来计算其平均值等级,Friedman 检验是一种针对多个样本的统计检验,它能够充分利用相关样本中的全部信息,一般用于推断多组相关样本是否有差别^[15].在多次比较下,所计算出的平均值等级越低,代表该方法在该性能指标上的表现越好.从表 2 最后一栏中可以清楚的看到,在由 Friedman 检验所计算出的平均值等级方面,本文所提 HS-SMOTE 方法在 G-mean 以及 F1-score 指标的检验结果分别为 2.24 和 2.37,是所有方法中的最小值,从统计学意义上表明其在少数类的分类性

能上要优于其他对比方法.

同时,为了进一步验证本文方法的有效性,还使用了统计学上的另一种检验手段—Wilcoxon 检验方法,该方法是一种针对两个样本的统计学方法,可更深入的验证本文方法与其他每种对比方法的差异性^[16].通常将测试阈值 α 设定为 0.05,并在每组检验中获得 p 值,当 p 值小于 α (0.05)时,则表明假设被拒绝,两种方法有显著差异, p 值越小,则差异越大;同时还能根据计算出的 $R+$ 、 $R-$ 值判断两种方法的好坏,当 $R+>R-$ 时,则表明前方法相比后方法更具有优势.当两种方法在多组数据集上表现有好有坏时,使用该检验方法可以比较快速的得出两者的对比结果.

Wilcoxon 检验实验结果如表 3 所示.

表 3 Wilcoxon 检验实验结果(p 值)

对比方法	G-mean				F1-score			
	$R+$	$R-$	p 值	$\alpha=0.05$	$R+$	$R-$	p 值	$\alpha=0.05$
HS-SMOTE vs 原始分类器	65	1	0.004	reject	51	5	0.019	reject
HS-SMOTE vs SMOTE	51	16	0.119	not reject	36	9	0.041	reject
HS-SMOTE vs Borderline-SMOTE	59	7	0.021	reject	62	4	0.01	reject
HS-SMOTE vs ADASYN	54	13	0.067	not reject	57	10	0.036	reject
HS-SMOTE vs Kmeans-SMOTE	59	7	0.021	reject	48	8	0.04	reject
HS-SMOTE vs SOMO	83	8	0.009	reject	71	7	0.012	reject
HS-SMOTE vs AMSCO	69	10	0.021	reject	67	11	0.027	reject

在表 3 由 Wilcoxon 检验所得的实验结果中,可以观察到大部分的假设都被拒绝了.这意味着本文所提 HS-SMOTE 方法与其他典型对比方法存在差异,并且 $R+$ 值大于 $R-$ 值,这表明 HS-SMOTE 方法相比其他方法具有优势.具体来说,在 92.86% 的组合中,本文方法相比对比方法的分类性能都获得提升,且通过了 Wilcoxon 检验.

因为本文方法的主要目标是通过对于少数类进行过采样来提高分类器对少数类的识别率,这意味着针对少数类的识别准确率将会得到显著提升,而经过过采样后,由于少数类的数量大幅度增加,可能会影响到模型对多数类的识别,从而导致有一些多数类会被错认为少数类,影响到多数类的识别准确率.针对这一情况,设计了一组实验用以验证本文方法对于多数类的识别效果.实验同样也在随机森林分类器上进行,实验参数和前文实验保持一致.实验结果如表 4 所示,表中第二列(RF)为使用原始数据直接对模型进行训练所得多数类识别结果,其后各列为经过各种过采样方法处理后再训练所得的多数类识别结果.

实验结果显示,相比于采用原始数据训练得到的分类结果,几乎所有采用过采样方法再平衡后的分类结果在多数类召回率指标上的分类效果均有不同程度的降低,这一结果表明过采样方法在提升了对少数类识别精度的同时,也会影响到对多数类的准确判断.但同时,从指标的数学意义上来看,G-mean 指标表示的是少数类的召回率以及多数类的召回率的几何平均值,前文中已经表明,本文方法在 G-mean 指标上表现良好,也就是说,所提 HS-SMOTE 方法不仅提升了模型对于少数类的识别准确率,而且正如表 4 所示,在针对多数类的召回率指标(Specificity)上,相比于其他过采样方法也有着一定的优势.

5 结束语

在本文中,为了突破 SMOTE 算法在线段上采样的限制,提出了一种超球体过采样(HS-SMOTE)方法.为了验证该方法的有效性,以 G-mean 以及 F1-score 指标作为评价指标,在 KEEL 上的 15 种不平衡数据集以及 RF 分类器上进行实验对比,并使用了 Wilcoxon 检

验以及 Friedman 检验对实验结果进行了分析. 实验结果表明 HS-SMOTE 方法能够有效地提升对不平衡数

据中少数类样本的识别效果,并且能够在分类器上相对稳定的运行,可以应用于多种场合.

表 4 多数类的召回率(Specificity)指标上的实验结果

Datasets	RF	SMOTE	Borderline-SMOTE	ADASYN	Kmeans-SMOTE	SOMO	AMSCO	HS-SMOTE
Wisconsin	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.97
Glass	0.92	0.86	0.80	0.80	0.85	0.85	0.82	0.88
Haberman	0.83	0.71	0.74	0.71	0.87	0.85	0.73	0.79
Vehicle	0.97	0.96	0.97	0.96	0.96	0.97	0.97	0.97
New-thyroid	1.00	0.98	0.98	0.99	0.99	0.99	0.99	0.98
Ecoli	0.98	0.90	0.89	0.92	0.92	0.95	0.94	0.96
Vowel	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Car	0.99	0.94	0.95	0.95	0.95	0.95	0.96	0.99
Yeast-2_vs_4	0.99	0.95	0.95	0.95	0.97	0.97	0.96	0.98
Winequality-red-4	0.99	0.84	0.93	0.84	0.85	0.95	0.96	0.97
Poker-9_vs_7	1.00	0.99	0.99	0.98	0.99	0.99	0.98	0.99
Abalone-21vs8	0.99	0.97	0.98	0.96	0.97	0.98	0.99	0.99
Winequality-red-8_vs_6-7	1.00	0.94	0.96	0.94	0.94	0.94	0.95	0.96
Poker-8-9_vs_6	1.00	0.87	0.90	0.87	0.87	0.98	0.89	0.99
Page-blocks	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
平均值	0.991 3	0.940 0	0.955 0	0.941 3	0.947 5	0.965 0	0.966 3	0.978 8

在本文算法对不平衡样本进行过采样的过程中,虽说扩展了在高维空间中的采样范围,使得新少数点的分布不再局限于一条线段上,但同时这也可能产生一些噪声,从而影响到分类的结果. 如何在采样过程中降低噪声的影响,进一步提升 HS-SMOTE 方法的性能,将会是接下来的工作重点.

参考文献:

[1] 贾统,李影,吴中海. 基于日志数据的分布式软件系统故障诊断综述[J]. 软件学报, 2020, 31(7): 1997-2018. DOI:10.13328/j.cnki.jos.006045.

JIA T, LI Y, WU Z H. Survey of state-of-the-art log-based failure diagnosis[J]. Journal of Software, 2020, 31(7): 1997-2018. DOI:10.13328/j.cnki.jos.006045.

[2] 李新鹏,高欣,何杨,等. 不平衡数据集下基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型[J]. 微电子学与计算机, 2020, 37(3): 14-19. DOI:10.19304/j.cnki.issn1000-7180.2020.03.004.

LI X P, GAO X, HE Y, et al. Prediction model of disk failure based on adaptive weighted bagging-GBDT algorithm under imbalanced dataset[J]. Microelectronics & Computer, 2020, 37(3): 14-19. DOI:10.19304/j.cnki.issn1000-7180.2020.03.004.

[3] LIANG X W, JIANG A P, LI T, et al. LR-SMOTE-An

improved unbalanced data set oversampling based on K-means and SVM[J]. Knowledge-Based Systems, 2020, 196: 105845. DOI: 10.1016/j.knosys.2020.105845.

[4] ELREEDY D, ATIYA A F. A comprehensive analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance[J]. Information Sciences, 2019, 505: 32-64. DOI:10.1016/j.ins.2019.07.070.

[5] DOUZAS G, BACAO F. Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning [J]. Expert Systems with Applications, 2017(82): 40-52. DOI:10.1016/j.eswa.2017.03.073.

[6] 李睿峰,许爱强,孙伟超,等. 基于样本重采样的电路非平衡数据预处理方法[J]. 系统工程与电子技术, 2020, 42(11): 2654-2660. DOI:10.3969/j.issn.1001-506X.2020.11.30.

LI R F, XU A Q, SUN W C, et al. Preprocessing method based on sample resampling for imbalanced data of electronic circuits[J]. Systems Engineering and Electronics, 2020, 42(11): 2654-2660. DOI: 10.3969/j.issn.1001-506X.2020.11.30.

[7] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches[J]. IEEE Transactions on Systems, Man,

- and Cybernetics, Part C (Applications and Reviews), 2012, 42 (4): 463-484. DOI: 10. 1109/TSMCC. 2011. 2161285.
- [8] 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(1): 321-357. DOI:10. 1613/jair. 953.
- [9] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//Proceedings of 2005 International Conference on Intelligent Computing. Hefei, China: Springer, 2005: 878-887.
- [10] HE H B, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong, China: IEEE, 2008. DOI: 10. 1109/IJCNN. 2008. 4633969.
- [11] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information-Sciences, 2018 (465): 1-20. DOI: 10. 1016/j. ins. 2018. 06. 056.
- [12] LI J Y, FONG S, WONG R K, et al. Adaptive multi-objective swarm fusion for imbalanced data classification[J]. Information Fusion, 2018(39): 1-24. DOI: 10. 1016/j. infus. 2017. 03. 007.
- [13] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32. DOI:10. 1023/A:1010933404324.
- [14] RAGHUWANSHI B S, SHUKLA S. Class imbalance learning using UnderBagging based kernelized extreme learning machine[J]. Neurocomputing, 2019 (329): 172-187.
- [15] GARCÍA S, FERNÁNDEZ A, LUENGO J, et al. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power[J]. Information Sciences, 2010, 180 (10): 2044-2064. DOI:10. 1016/j. ins. 2009. 12. 010.
- [16] DEWAN I, RAO B L S P. Wilcoxon-signed rank test for associated sequences[J]. Statistics & Probability Letters, 2005, 71(2): 131-142. DOI:10. 1016/j. spl. 2004. 10. 034.

作者简介:

林静怀 男,(1971-),硕士,教授级高级工程师.研究方向为电力调度自动化研发和管理、电力市场、电力大数据分析.

刘治宇 男,(1997-),硕士.研究方向为机器学习.

李军良 男,(1981-),硕士,高级工程师.研究方向为电力系统自动化、软件工程.

高欣(通讯作者) 男,(1974-),博士,副教授.研究方向为机器学习、数据挖掘. E-mail: xlhhh74@bupt. edu. cn.

李泽科 男,(1977-),硕士,高级工程师.研究方向为电力调度自动化系统建设、电力大数据分析.

唐志军 男,(1973-),硕士,高级工程师.研究方向为电力系统自动化、电力大数据分析.

余斯航 男,(1988-),硕士,工程师.研究方向为电力系统自动化.

徐建航 男,(1988-).研究方向为电力系统自动化.

(上接第 64 页)

- [9] 李娟,张海波,李健,等.一种带有温漂修调的二阶曲率补偿基准源电路[J].电子与封装,2019,19(11): 26-30. DOI:10. 16257/j. cnki. 1681-1070. 2019. 1106.
- LI J, ZHANG H B, LI J, et al. Second-order curvature-compensated band-gap reference with trimming resistive circuits[J]. Electronics and Packaging, 2019, 19(11): 26-30. DOI: 10. 16257/j. cnki. 1681-1070. 2019. 1106.

作者简介:

韦崇敏 男,(1993-),硕士.研究方向为模拟集成电路. E-mail: chongminwei@163. com.

宋树祥 男,(1970-),博士,教授.研究方向为模拟集成电路、智能检测系统.

蒋品群 男,(1970-),博士,副教授.研究方向为集成电路设计与复杂网络理论及其应用.

蔡超波 男,(1986-),硕士,讲师.研究方向为模拟集成电路.

吕芸芸 女,(1995-),硕士.研究方向为模拟集成电路.

陈佳佳 女,(1995-),硕士.研究方向为模拟集成电路.