ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 3,May 2021,pp. 509－518
DOI:10. 16337/j. 1004-9037. 2021. 03. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于属性选择和采样策略的不平衡数据动态分类方法

赵冬雪，王　昕，王利东

（大连海事大学理学院，大连 116026）

摘　要：不平衡数据分类是机器学习领域的重要研究方向之一，现有不平衡学习算法大多针对二分类而无法满足多分类需求。本文面向多类不平衡数据分类问题，通过结合粗糙集、重采样方法以及动态集成分类策略设计了一种新的多分类模型。该模型运用综合采样方式和粗糙集属性约简技术获得多个平衡数据子集，在此基础上实现动态集成分类模型的构建。真实数据集上的22组实验验证了该模型与两种经典算法相比对少数类样本具有更好的预测性能，可成为多类不平衡数据分类的可选策略。
关键词：不平衡数据；动态选择；粗糙集；属性约简；重采样
中图分类号：TP181　　**文献标志码：**A

Dynamic Classification for Multi-imbalanced Datasets via Attribute Selection and Sampling Strategy

ZHAO Dongxue, WANG Xin, WANG Lidong

(School of Science, Dalian Maritime University, Dalian 116026, China)

Abstract: The classification of imbalanced datasets is one of the important topics in machine learning. Most of the existing imbalance learning algorithms designed for dichotomies are insufficient to meet multi-class classification needs. To tackle multi-class imbalance classification problem, we design a new multi-classification model synthesizing rough sets, resampling methods and dynamic ensemble classification strategy in this study. The model utilizes the hybrid sampling and the rough set reduction algorithm to generate multiple balanced data subsets, on which the construction of the dynamic ensemble classification model is realized. The experiments on 22 real datasets demonstrate that the designed method has higher prediction performance on identifying minority samples compared with two previous algorithms, which can be an alternative selection strategy in multi-class imbalance classification.
Key words: imbalanced data; dynamic selection; rough sets; attribute reduction; resampling

引　　言

类失衡是现实生活中普遍存在的问题，正确识别少数类样本可以获取重要的信息。例如，银行欺诈行为预测（发生欺诈情况远小于正常交易的情况），医疗系统中识别癌症病变（癌变细胞相对于正常细胞的数量极少），工厂机器系统的异常检测（绝大多数时间运转是正常的，极少数时间是非正常的）等

基金项目：国家自然科学基金（61803065，61773352）资助项目；中央高校基本科研业务费专项基金（3132019602）资助项目。
收稿日期：2020-06-12；**修订日期：**2020-12-11

均涉及不平衡数据的处理。特别是在大数据背景下,数据比例严重失衡,给生产、生活实践带来巨大挑战。

分类任务中的数据不平衡问题[1]体现在样本类别分布呈偏斜状态,即数据集中一个或多个类(多数类)的样本数量远远超过其他类别(少数类)。而多种经典分类算法是以准确率为主要衡量指标,会侧重多数类的正确率而忽略识别少数类的重要价值,从而难以获得理想的分类结果。因此,数据不平衡问题引起了研究者的关注,并且提出了多种处理方法[2-4]。总体来看,解决方案主要采取数据预处理和新算法设计,如数据重采样技术和集成学习等技术。

重采样技术[5]旨在调整各类别间的不平衡率,常用的重采样技术有随机欠采样、SMOTE过采样[6]等。随机欠采样通过降低多数类样本数量达到平衡的目的,但会导致多数类样本信息的丢失;SMOTE过采样通过少数类样本生成新样本,使不平衡数据在一定程度上达到平衡状态,但不足之处在于盲目地扩充少数类容易模糊决策边界。因此许多研究者不断探索新的重采样技术,尤其对SMOTE方法进行改进,如Borderline-SMOTE[7]、MSMOTE[8]、SMOTE-RSB*[9]、SMOTE-IPF[10]以及三支决策理论的SMOTE算法[11]等。以往的研究工作大多集中于不平衡数据的二分类问题。与二分类相比,多类不平衡数据集内部结构复杂,表现出以下特点:(1)少数类分布较稀疏;(2)局部甚至整体出现多类重叠;(3)多数类稠密区域存在少数类噪声点;(4)少数类样本聚集在多数类稠密区域。由此可见多类不平衡数据集决策边界涉及更多的两类区分,特别是多类重叠的情况给分类任务带来更大困难。

基于代价敏感的分类算法和集成学习算法是两种处理多类不平衡数据分类的主要策略。在分类过程中,基于代价敏感的学习算法[12]主要关注当不同分类误差导致不同惩罚程度时,如何更好地训练分类器,但合理地构造代价矩阵和函数是研究此类问题的难点。近年来,集成学习与数据预处理方法相融合成为解决不平衡数据分类问题的新模式[13]。例如,García等[14]提出动态集成分类算法,将SMOTE算法嵌入集成过程中,通过获取不同的平衡数据子集来训练基分类器,并选择分类能力排名前40%的基分类器来构建集成分类模型,一定程度上提高了对多类不平衡数据的分类性能。但该算法在实现过程中存在以下问题:(1)获取平衡子集时未对子集的规模作细致规划,容易产生较小规模的训练数据集,不利于分类模型的学习;(2)子集规模的不确定性使得训练集中的每一类都要进行采样处理,增加了算法复杂度;(3)不平衡率较大时SMOTE平衡过程将产生大量合成样本,增加数据的不确定性,不利于提升基分类器的预测能力。

本文在数据预处理阶段结合属性选择和数据平衡算法来减少不相关属性对分类器的影响;在分类器集成阶段,利用分类能力的评价函数选择能力较优的分类器构造集成分类模型,以提高分类器对少数类样本的识别能力。

1 不平衡数据动态集成分类方法的设计过程

粗糙集理论在处理含糊性和不确定性问题中发挥显著性作用,利用粗糙集可实现信息系统的属性约简,即剔除冗余属性挖掘数据的核心信息。研究表明,利用不同粗糙集模型可从多个视角获得约简子集,并在其基础上构建集成算法有助于促进分类器的差异性[15-17]。受此启发,本文基于邻域粗糙集约简方法[18]以及模糊粗糙集约简方法[19-24]获得多个属性子集,将其与重采样方法和动态集成分类策略融合来处理多类不平衡数据,其主要流程如图1所示,整体包含3个部分:数据预处理、分类池构建以及动态选择实现。

1.1 数据预处理

对数据进行平衡处理是提高分类器对少数类样本识别能力的重要手段。本文对García等[14]提出的数据平衡算法进行改进,首先明确平衡子集的容量,既保证子集大小具有充分多样性,同时避免由于

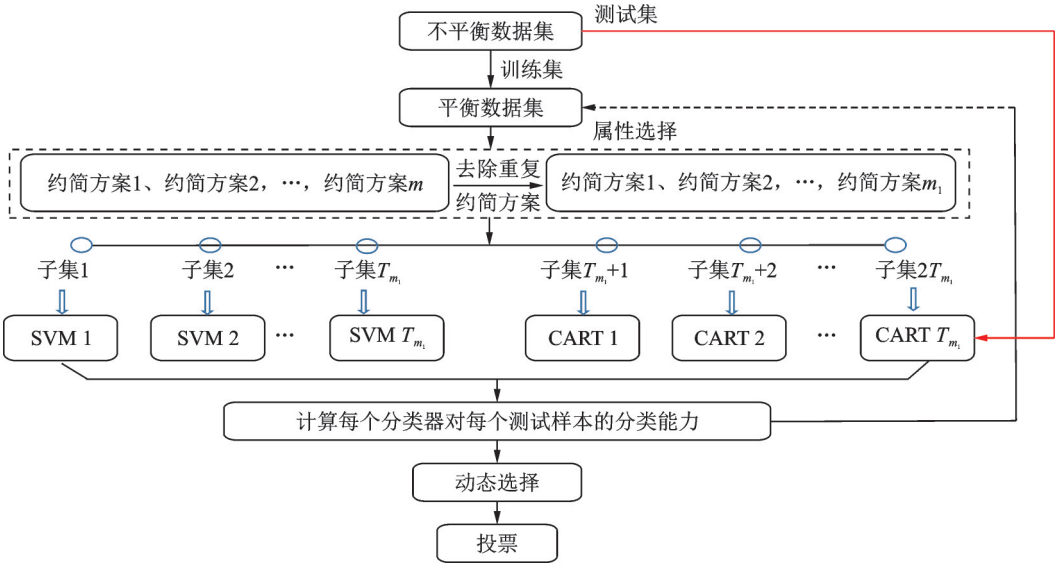

图1 基于粗糙集属性约简的动态集成分类方法流程图

Fig.1 Flow chart of dynamic ensemble classification method based on rough set attribute reduction

随机性而出现训练样本数目过少情况;其次对多数类样本随机欠采样,并将所得样本集内部所有小于类平均样本数的类别视为少数类,减少需要采样的类别数量。

以多类重叠情况为例给出数据平衡处理的步骤。

第1步:统计最大类的样本数、类平均样本数(图2(a)中分别为17、9.33)。

第2步:随机选取类平均样本数与最大类之间的整数作为重采样标准S(在本例中假设$S=10$)。

第3步:按照标准S对所有样本数大于类平均样本数的类别进行随机欠采样(图2(b)中的菱形样本是欠采样后的结果)。

第4步:重新计算欠采样后数据集的类平均样本数,当前数据集中小于该值的类别记为少数类(图2(b)中新数据集的类平均样本数为7,少数类为三角形)。

第5步:对少数类进行复制(图2(c))后执行SMOTE过采样,得到相对平衡子集如图2(d)所示。

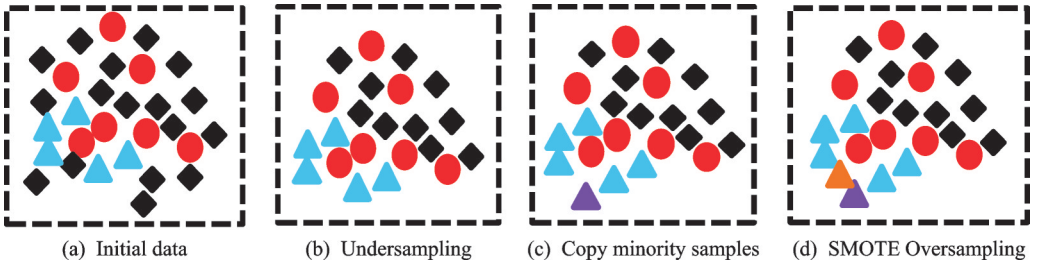

图2 数据平衡处理过程

Fig.2 Data balancing process

在平衡数据的基础上,对数据进行属性约简,以提高分类器的差异性。本文选用的属性约简算法包括:(1)前向搜索邻域粗糙集属性约简快速算法[18];(2)基于模糊邻域粗糙集的启发式约简算法(Fuzzy neighborhood rough sets, FNRS[19]);(3)基于拟合模糊粗糙集的启发式约简算法(Fitting fuzzy

rough sets, FFRS[20]);(4)基于K近邻的约简算法(K-nearest neighborhood rough sets, NNRS[21]);(5)基于邻域辨别指数的启发式约简算法(Heuristic algorithm based on neighborhood discrimination index, HANDI[22]);(6)基于邻域模型的前向属性约简算法(Forward attribute reduction based on neighborhood model, FARNeM[23]);(7)基于变精度模糊粗糙模型的前向属性约简算法(Forward attribute reduction based on variable precision fuzzy-rough sets, FAR-VPFRS[24])。多种约简算法综合使用,并去除具有包含或相等关系的子集,以获取互补性的约简子集。

1.2 基础分类池

为提高基分类器的准确性和多样性,本文综合考虑属性约简和候选分类器之间的协同作用。经过数据平衡和属性约简预处理过程后,在多样化数据子集上训练基分类器。算法1中的伪代码描述了生成候选分类池的过程。

算法1 生成候选分类池

输入:原始数据集$\boldsymbol{D}$,基础分类器分类回归树(Classification and regression tree, CART)和支持向量机(Support vector machine, SVM)、循环次数T

输出:候选分类池$\boldsymbol{P}$

(1) 将$\boldsymbol{D}$通过5折交叉验证划分出训练集$\boldsymbol{D}_1$和测试集$\boldsymbol{D}_2$

(2) for $i=1:T$

(3) 平衡训练集$\boldsymbol{D}_1$,得到$\boldsymbol{D}_{1\text{new}}$

(4) 对$\boldsymbol{D}_{1\text{new}}$进行属性约简($m$种方法),经筛选后得到$T_{m_1}+1$个具有属性差异性的数据子集$\boldsymbol{DR}_j$,$j=1,2,\cdots,T_{m_1}+1$

(5) for $j=1:T_{m_1}+1$

(6) 将$\boldsymbol{DR}_j$通过5折交叉验证法划分为二次训练集和二次测试集

(7) 分类器CART_{ij}与SVM_{ij}←二次训练集

(8) end for

(9) end for

(10) $2T_{m_1}$个分类器合并为候选分类池$\boldsymbol{P}$,输出候选分类池$\boldsymbol{P}$

1.3 动态选择实现

动态集成分类器由候选分类池中的基础分类器确定,受García等[14]的研究启发,在候选分类池中选择分类能力最优的分类器组来构造动态集成分类器。在此过程中,通过构造分类能力的评价函数来评估候选分类器的分类能力,计算方法如下。

假设h为分类器,$\boldsymbol{x}_i$为待测样本,待测样本的邻居个数为k,则h对$\boldsymbol{x}_i$的分类能力函数为:$F_{h|x_i}=\sum_{t=1}^{k}I(\boldsymbol{x}_{it})\times w_{it}$,其中$I(h(\boldsymbol{x}_{it})=y_{it})=\begin{cases}0 & h(\boldsymbol{x}_{it})\neq y_{it}\\ 1 & h(\boldsymbol{x}_{it})=y_{it}\end{cases}$,即当样本的预测标签与真实标签相同时为1,否则为0;$y_{it}(t=1,2,\cdots,k)$为第t个邻居样本的类标签;$w_{it}=1\Big/\left(\exp\left(\left|\oplus y_{it}\right|\right)\right)$为第$t$个邻居的权重,$\left|\oplus y_{it}\right|$表示$k$邻域中,与该邻居样本类别相同的样本个数。显然$k$邻域中同类别样本越多,其权重越低,该定义方式间接提高了少数类的权重。

当各基分类器对测试样本确定分类能力后,针对每个测试样本均选择排名前N的基分类器来构建集成分类模型,实现过程见算法2。

算法2 动态选择构建集成分类模型

输入:测试样本 $\mathbf{test}_i \in \boldsymbol{D}_2$ $(i=1,2,\cdots,s)$,候选分类池 $\boldsymbol{P}(h_j(j=1,2,\cdots,2T_{m_1})$为 $\boldsymbol{P}$ 中的第 j 个分类器),二次测试集,最近邻个数 k,动态选择参数 N,权重参数 a

输出:集成分类模型 M

(1) for $i=1:s$

(2) for $j=1:2T_{m_1}$

(3) 预测 $\mathbf{test}_i$ 的类标签←$h_j(\mathbf{test}_i)$

(4) 标记二次测试集中 test_i 的 k 个最近邻样本

(5) 计算每个邻居样本的权重并标准化

(6) 计算 h_j 对样本 $\mathbf{test}_i$ 的分类能力 $F_{h_j|\mathbf{test}_i}$

(7) end for

(8) 对分类能力排序,选择能力排名前 N 的基分类器合成 $\mathbf{test}_i$ 的集成分类模型 M

(9) end for

(10) 输出集成分类模型 M

2 实验设计与结果分析

根据上述研究思路与内容,将对所提模型步骤的合理性和性能进行实验分析与比较。

2.1 实验参数与数据

本文以CART和SVM为基分类器,为便于实验结果比较,实验中部分参数指标与文献[14]中相同,具体如下:算法1中循环次数 $T=25$,$m=7$;算法2中最近邻个数 $k=9$,动态选择参数 $N=9$;结果评价指标如下。

(1) 类平均准确率

$$\text{MAvA}=\sum_{i=1}^{C}\text{Acc}_i \Big/ C \tag{1}$$

式中:C 为样本集的类数,Acc_i 代表第 i 类的准确率。

(2) 几何平均准确率

$$\text{G-mean}=\sqrt[C]{\prod_{i=1}^{C}\text{Acc}_i} \tag{2}$$

(3) 精确率

$$\text{Precision}=\frac{1}{C}\cdot\sum_{i=1}^{C}\frac{TP_i}{TP_i+FP_i} \tag{3}$$

式中:TP_i 表示第 i 类中预测正确的样本个数,FP_i 表示第 i 类中预测错误的样本个数。

(4) F 值

$$F\text{-measure}=\frac{1}{C}\cdot\sum_{i=1}^{C}\frac{\text{Precision}_i\cdot\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i} \tag{4}$$

式中:Recall_i 表示第 i 类样本被正确预测的比例。

本文实验数据选自KEEL数据库[25]中22个不平衡数据集(如表1所示),其不平衡率介于1.023 6到163.67之间,类数介于2到10之间,属性个数介于2到34之间。

表1 不平衡数据集描述

Table 1 Description of the imbalanced data sets

数据集名称	不平衡率	类数	属性个数	类分布情况
auto_205	22.33	6	25	54/32/27/67/22/3
balance	5.88	3	4	288/49/288
car	18.62	4	6	384/69/1210/65
contraceptive	1.89	3	9	629/333/511
dermatology	5.60	6	34	111/60/71/48/48/20
ecoli	71.50	8	7	143/77/2/2/35/20/5/52
glass	8.44	6	9	70/76/17/29/13/9
hayesroth	1.70	3	4	65/64/31
heart	1.25	2	13	150/120
lymphography	28.50	8	18	57/37/18/10/8/8/8/2
newthyroid	5.00	3	5	150/35/30
pageblocks	163.67	5	10	491/33/3/9/12
penbased	1.10	10	16	114/114/106/114/106/105/115/105/106/115
redwinequality	68.10	2	11	681/10
transfusion	3.20	2	4	178/570
vehicle	1.11	4	18	217/217/216/196
wan_2	2.50	4	2	100/150/250/250
wine	1.48	3	13	59/71/48
yeast	125.20	7	8	288/480/626/35/30/20/5
zoo	10.25	6	16	41/20/5/13/4/8/10
sick	15.37	2	27	2629/171
wdbc	1.68	2	30	212/357

2.2 实验结果

由于部分不平衡数据集中少数类样本数目过小,实验中将数据集划分为过多折数将导致更大程度的不平衡,本文实验使用5折交叉验证。

2.2.1 属性约简与平衡先后顺序对比实验

为探究粗糙集属性约简与数据平衡先后顺序对不平衡数据集的分类性能的影响,本文分别将属性约简步骤置于数据平衡过程之前和之后,以SVM算法为基分类器来构建集成分类模型,并以16个数据集进行类平均准确率(MAvA)的对比。结果如图3所示,约93.75%的数据集在平衡之后再进行属性约简的分类结果更优,故后续实验步骤将采取“先平衡后约简”的顺序。

2.2.2 实验步骤

基于上组实验,本文为讨论数据平衡、属性约简和动态选择3个步骤的内在影响,将采用去除其中1个步骤进行实验对比。实验步骤对结果的影响如图4所示。图4表明,去掉数据平衡过程后,MAvA最低;仅使用数据平衡和属性约简两个步骤或者只涉及平衡处理和动态选择时,MAvA略高;当实验同时包含3个步骤,结果最优。综上所述,本文方法的3个步骤在不平衡学习中具有协同促进作用。

2.2.3 不同集成算法对比实验

为进一步衡量本文所设计实验方法的性能,与文献中的两种集成方法进行实验比较。“SC”

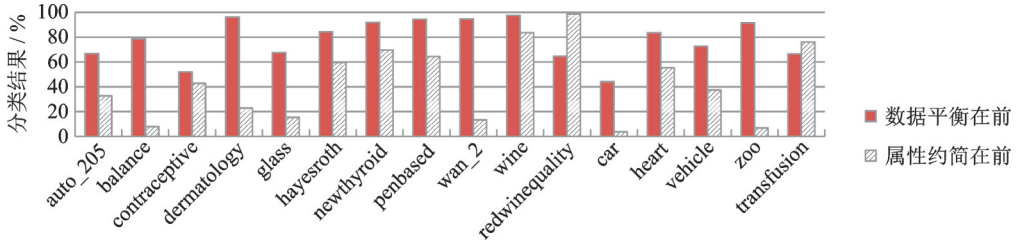

图3 属性约简与数据平衡不同顺序的对比实验结果

Fig.3 Comparison of experimental results with different order between attribute reduction and data balance

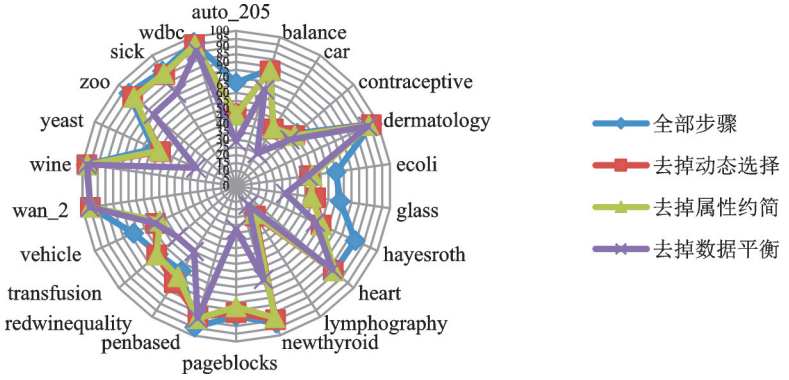

图4 实验步骤对结果的影响(单位:%)

Fig.4 Influence of experimental steps on the results (Unit: %)

(SMOTE+CART)模型是García等[14]提出的多类不平衡动态集成方案;"SS"(SMOTE+SVM)是全体基分类器参与投票的SVM集成模型;"SSCRS"(SMOTE+SVM+CART+ROUGH SET)为本文所设计的模型。10次5折交叉实验的类平均准确率(MAvA)、几何平均准确率(G-mean)、精确度(Precision)以及F值(F-measure)结果分别如表2~5所示,其中括号中数字表示同1个数据集在不同方法下准确率排序。

表2 类平均准确率结果

Table 2 Results of MAvA %

数据集	SC	SSCRS	SS	数据集	SC	SSCRS	SS
auto_205	69.48(1)	66.76(2)	44.89(3)	newthyroid	68.59(3)	91.98(1)	88.50(2)
balance	37.71(3)	78.62(1)	77.72(2)	pageblocks	67.99(3)	84.09(1)	81.12(2)
car	37.75(3)	43.99(2)	46.24(1)	penbased	84.42(3)	94.50(1)	88.76(2)
contraceptive	48.25(3)	51.99(1)	50.21(2)	redwinequality	50.00(3)	64.87(1)	64.64(2)
dermatology	96.96(1)	96.23(2)	93.17(3)	transfusion	50.00(3)	66.52(2)	68.06(1)
ecoli	70.60(1)	64.90(2)	49.62(3)	vehicle	57.14(2)	72.75(1)	56.19(3)
glass	66.84(2)	67.55(1)	48.94(3)	wan_2	72.56(3)	94.78(2)	94.83(1)
hayesroth	65.22(2)	84.34(1)	57.51(3)	wine	74.93(3)	97.57(1)	96.93(2)
heart	50.00(3)	83.58(1)	83.08(2)	yeast	53.75(3)	58.30(1)	54.51(2)
lymphography	19.45(3)	21.51(2)	23.37(1)	zoo	84.72(3)	91.57(1)	85.71(2)
sick	51.85(3)	88.04(1)	80.52(2)	wdbc	50.62(3)	96.63(1)	94.17(2)
平均排名	SC:2.55 SSCRS:1.32 SS:2.09						

表3 几何平均准确率结果

Table 3 Results of G-mean %

数据集	SC	SSCRS	SS	数据集	SC	SSCRS	SS
auto_205	0.00(3)	51.35(1)	8.81(2)	newthyroid	24.27(3)	91.40(1)	87.58(2)
balance	0.00(3)	77.37(1)	76.00(2)	pageblocks	0.00(3)	64.19(2)	71.16(1)
car	5.43(3)	39.35(2)	45.04(1)	penbased	0.00(3)	94.28(1)	87.65(2)
contraceptive	1.18(3)	51.20(1)	46.94(2)	redwinequality	0.00(3)	42.68(2)	46.87(1)
dermatology	96.83(1)	95.88(2)	91.35(3)	transfusion	0.00(3)	66.03(2)	67.33(1)
ecoli	22.56(1)	12.86(2)	0.00(3)	vehicle	0.00(3)	68.87(1)	51.32(2)
glass	11.03(2)	49.15(1)	0.00(3)	wan_2	0.00(3)	94.67(2)	94.69(1)
hayesroth	10.21(3)	82.36(1)	40.58(2)	wine	68.62(3)	97.48(1)	96.79(2)
heart	0.00(3)	83.39(1)	82.90(2)	yeast	0.00(3)	40.69(1)	39.79(2)
lymphography	0.00(1)	0.00(1)	0.00(1)	zoo	51.86(2)	58.64(1)	46.78(3)
sick	0.00(3)	87.97(1)	80.37(2)	wdbc	0.00(3)	96.62(1)	94.09(2)
平均排名	SC:2.64 SSCRS:1.32 SS:1.91						

表4 精确率结果

Table 4 Results of precision %

数据集	SC	SSCRS	SS	数据集	SC	SSCRS	SS
auto_205	17.93(2)	27.25(1)	11.72(3)	newthyroid	33.06(3)	89.32(1)	86.68(2)
balance	15.89(3)	53.61(1)	51.35(2)	pageblocks	2.26(3)	38.59(1)	30.82(2)
car	6.50(3)	21.94(1)	21.93(2)	penbased	33.95(3)	65.97(1)	46.98(2)
contraceptive	23.10(3)	33.92(1)	31.09(2)	redwinequality	1.45(3)	53.48(1)	51.11(2)
dermatology	82.60(2)	84.24(1)	73.14(3)	transfusion	23.80(3)	62.27(2)	63.14(1)
ecoli	35.61(1)	26.53(2)	12.78(3)	vehicle	26.46(3)	46.33(1)	28.25(2)
glass	15.26(2)	22.54(1)	9.22(3)	wan_2	27.67(3)	81.30(2)	81.32(1)
hayesroth	42.89(2)	71.86(1)	35.01(3)	wine	62.27(3)	94.62(1)	93.00(2)
heart	55.56(2)	83.85(1)	83.50(3)	yeast	8.66(3)	13.93(1)	9.53(2)
lymphography	6.41(2)	6.77(1)	5.97(3)	zoo	54.32(3)	76.22(1)	64.27(2)
sick	6.11(3)	68.19(1)	58.82(2)	wdbc	37.26(3)	96.62(1)	95.32(2)
平均排名	SC:2.64 SSCRS:1.14 SS:2.23						

表2～5可以得出：

(1)22组数据中有15组数据在本文模型下获得了最优MAvA值，并且所有数据的MAvA排名均在前两位。这表明本文模型较之其他两种方法对少数类样本的识别能力有所增强。

(2)受数据偏斜分布的影响，“SC”模型在多个数据集上表现出对数据某一类别全部分错的情况，使得G-mean指标值为0。而本文构建的“SSCRS”模型仅在lymphography一个数据集表现出类似情况。这说明，“SSCRS”模型能够较好地实现对少数类样本的分类。同时，与“SS”模型相比，“SSCRS”模型在G-mean指标下的平均排名提升了0.59，即该模型整体性能有一定程度的提升。

(3)“SSCRS”模型在Precision指标下的平均排序结果为1.14，与“SC”“SS”分类模型相比分别提高了1.50、1.09，该结果表明本文模型对少数类样本具有较高的预测性能。

(4)“SSCRS”模型在91%的数据上取得最优F-measure值，结果表明本文构建的“SSCRS”模型对

表5 F值结果

Table 5 Results of F-measure %

数据集	SC	SSCRS	SS	数据集	SC	SSCRS	SS
auto_205	33.66(2)	38.47(1)	21.86(3)	newthyroid	29.09(3)	90.19(1)	87.08(2)
balance	38.51(3)	58.09(1)	55.63(2)	pageblocks	5.34(3)	49.26(1)	36.52(2)
car	12.00(3)	23.70(1)	21.56(2)	penbased	53.75(3)	77.40(1)	61.05(2)
contraceptive	45.20(1)	39.90(2)	36.86(3)	redwinequality	2.85(3)	75.88(1)	65.61(2)
dermatology	88.50(2)	88.99(1)	78.91(3)	transfusion	38.44(3)	60.48(1)	60.35(2)
ecoli	51.67(1)	39.63(2)	26.16(3)	vehicle	47.99(2)	55.20(1)	36.89(3)
glass	28.17(2)	30.72(1)	22.45(3)	wan_2	52.78(3)	86.93(1)	86.90(2)
hayesroth	55.75(2)	74.88(1)	43.71(3)	wine	64.60(3)	95.85(1)	94.60(2)
heart	71.43(3)	83.44(1)	83.07(2)	yeast	16.63(2)	18.24(1)	12.82(3)
lymphography	21.65(3)	27.14(1)	25.68(2)	zoo	66.11(3)	84.99(1)	76.34(2)
sick	11.51(3)	73.32(1)	58.47(2)	wdbc	54.29(3)	96.62(1)	94.67(2)
平均排名			SC:2.55 SSCRS:1.09 SS:2.36				

于解决多类不平衡数据分类任务具有一定优势。

综上所述,本文所提出的分类模型在MAvA、G-mean、Precision以及F-measure四种评价指标下的平均排名均高于SC和SS模型,即SSCRS较之SC和SS对大多数不平衡数据具有更强适用性且模型对少数类样本的分类能力略有提升。

3 结束语

本文针对多类不平衡数据集的分类问题,提出基于粗糙集属性选择和采样策略的动态集成分类方案,其主要优势在于:(1)采样策略和粗糙集属性选择分别从数量和属性角度对数据进行切割,增加数据多样性;(2)动态选择过程为每一测试样本均构建了最优集成分类模型,使得基分类器的差异性得到有效利用。通过对22组数据集进行实验验证,可得出结论:本文模型一定程度上提高了对不平衡数据的预测能力。考虑到不平衡数据集分类算法遇到的主要困难与不平衡率有关,研究不平衡率与分类器动态选择之间的关联性将成为下一步研究的方向。

参考文献:

[1] KRAWCZYK B. Learning from imbalanced data: Open challenges and future directions[J]. Progress in Artificial Intelligence, 2016, 5: 221-232.

[2] GUO Haixiang, LI Yijing, GU Mingyun, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Systerm Application, 2017, 73: 220-239.

[3] 李瑞, 袁小玲. 半动态集成选择分类方法[J]. 计算机与现代化, 2015, 2: 48-51.
LI Rui, YUAN Xiaoling. Semi-dynamic integration selection classification method[J]. Computer and Modernization, 2017, 73: 220-239.

[4] KHAN S H, HAYAT M, BENNAMOUN M, et al. Cost sensitive learning of deep feature representations from imbalanced data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 29 (8): 3573-3587.

[5] 毕华, 梁洪力, 王珏. 重采样方法与机器学习[J]. 计算机学报, 2009, 5: 862-877.
BI Hua, LIANG Hongli, WANG Jue. Resampling methods and machine learning[J]. Chinese Journal of Computers, 2009, 5: 862-877.

[6] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligent Research, 2002 (16): 321-357.

[7] HAN Hui, WANG Wenyuan, MAO Binghuan. Borderline-SMOTE: A new over-sampling method in imbalanced data sets

learning[C]// Proceedings of Advances in Intelligent Computing. Berlin, Heidelberg: Springer, 2005: 878-887.

[8] HU Shengguo, LIANG Yanfeng, MA Lintao, et al. MSMOTE: Improving classification performance when training data is imbalanced[C]//Proceedings of 2009 Second International Workshop on Computer Science and Engineering. Qingdao, China: IEEE, 2009, 2: 13-17.

[9] RAMENTOL E, CABALLERO Y, BELLO R, 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, 2012, 33 (2): 245-265.

[10] SÁEZ J A, LUENGO J, STEFANOWSKI J, et al. SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering[J]. Information Sciences, 2015, 291: 184-203.

[11] 胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报, 2018, 46(1): 135-144.
HU Feng, WANG Lei, ZHOU Yao. Oversampling method of imbalanced data based on three-way decision[J]. Acta Electronica Sinica, 2018, 46(1): 135-144.

[12] ZHOU Zhihua, LIU Xuying. On multi-class cost-sensitive learning[J]. Computational Intelligence, 2010, 26(3): 232-257.

[13] GUO Haixiang, LI Yijing, LI Yanan, et al. BPSO-adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification[J]. Engineering Applications, 2016, 49: 176-193.

[14] GARCÍA S, ZHANG Zhongliang, ALTALHI A, et al. Dynamic ensemble selection for multi-class imbalanced datasets[J]. Information Sciences, 2018, 445: 22-37.

[15] 杨传振, 朱玉全, 陈耿. 一种基于粗糙集属性约简的多分类器集成方法[J]. 计算机应用研究, 2012, 29 (5): 1648-1650.
YANG Chuanzhen, ZHU Yuquan, CHEN Geng. A multi-classifier integration method based on rough set attribute reduction [J]. Application Research of Computers, 2012, 29 (5): 1648-1650.

[16] XU Hang, WANG Wenjian, QIAN Yuhua. Fusing complete monotonic decision trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(10): 2223-2235.

[17] QIAN Yuhua, XU Hang, LIANG Jiye. Fusing monotonic decision trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(10): 2717-2728.

[18] 胡清华, 赵辉, 于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能, 2008, 21(6): 732-738.
HU Qinghua, ZHAO Hui, YU Daren. Fast reduction algorithm for symbolic and numerical attributes based on neighborhood rough sets[J]. Pattern Recognition and Artificial Intelligence, 2008, 21(6): 732-738.

[19] WANG Changzhong, SHAO Mingwen, HE Qiang, et al. Feature subset selection based on fuzzy neighborhood rough sets[J]. Knowledge-Based Systems, 2016, 111: 173-179.

[20] WANG Changzhong, QI Yali, SHAO Mingwen, et al. A fitting model for feature selection with fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2016, 25(4): 741-753.

[21] WANG Changzhong, SHI Yunpeng, FAN Xiaodong, et al. Attribute reduction based on k-nearest neighborhood rough sets[J]. International Journal of Approximate Reasoning, 2019, 106: 18-31.

[22] WANG Changzhong, HU Qinghua, WANG Xizhao, et al. Feature selection based on neighborhood discrimination index[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(7): 2986-2999.

[23] HU Qinghua, YU Daren, XIE Zongxia. Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2): 866-876.

[24] HU Qinghua, XIE Zongxia, YU Daren. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521.

[25] ISAAC T, SERGIO G, SALVADOR G L, et al. KEEL 3.0: An open source software for multi-stage analysis in data mining [J]. International Journal of Computational Intelligence Systems, 2017, 10(1): 1238-1249.

作者简介:

赵冬雪(1996-),女,硕士研究生,研究方向:数据挖掘、模糊数学,E-mail:1741720293@qq.com。

王昕(1978-),女,副教授,研究方向:粒计算、模糊分类,E-mail:xenawang@dlmu.edu.cn。

王利东(1979-),**通信作者**,男,教授,研究方向:粒计算、预测与决策,E-mail:ldwang@dlmu.edu.cn。

(编辑:刘彦东)