# 高维类不平衡冠心病数据的统计建模

宗敏洁 江玉婷 吴愿交

**摘要：**医疗数据具有较高的不平衡性，应用研究亟需解决的问题主要是提高医疗不平衡数据的分类性能。选取五种常用分类器，处理冠心病不平衡数据，得出以下结论：在不平衡比增加时，传统的分类器会忽略少数类对整个数据集的影响，分类精度明显下降；对冠心病不平衡数据进行再平衡后，分类结果均有所改善，数据再平衡可以提高分类性能；SVM算法在针对样本容量较小的数据集时，其分类性能最好。

**关键词：**不平衡数据分类；冠心病；不平衡比；再平衡策略

## 一、研究背景

（一）问题背景

1. 冠心病概述

冠状动脉粥样硬化性心脏病是冠状动脉血管发生动脉粥样硬化病变而引起血管腔狭窄或阻塞，造成心肌缺血、缺氧或坏死而导致的心脏病[1]，常常被称为“冠心病”。我国1987～1993年多省市35～64岁人群调查（中国MONICA）中发现，冠心病的最高发病率为108.7/10万（山东青岛），最低为3.3/10万（安徽滁州），冠心病患病率数据有较显著的地区差异，往往北方省市会高于南方省市。其中，城市人口冠心病患病率为1.59%，农村为0.48%，合计为0.77%，均呈现上升趋势[2]。近年来，冠心病发病率随着我国社会的发展和生活水平的提高也不断上升，是严重影响人们生活质量和身体健康的主要疾病之一[3]。因此，对于冠心病及其并发症数据的研究是非常重要的。

2. 不平衡数据概述

近年来，机器学习与数据挖掘主要的研究热点和难点之一是不平衡数据的处理。随着大数据概念的不断升温，学术界及产业界对不平衡数据处理问题的研究呈现出了逐渐增长的趋势。在众多领域中，广泛存在着不平衡数据的分类，例如：医学诊断、文本分类和生物信息学等等。

当罕见的样本很少出现时，很可能被预测为稀有事件。稀有事件常常未被发现或被忽视，或被假设为噪音或异常值，从而导致对少数类的错误分类比多数类更多。这时某一类不能得到充分展现，数据集中就会出现不平衡的类分布，这意味着少数类要比多数类的样本量少得多。

不平衡数据的主要特征是：变量维度高、数据共线性严重、样本少、数据的不平衡度高。在医疗不平衡数据中，二类不平衡问题在实际应用中最为常见。在医学诊断中，疾病患者的数量往往远小于健康人的数量，如何能准确检测疾病患者尤为重要。诊断过程中，若将疾病患者误诊为健康人，必然会耽误治疗时间，对患者的身心健康十分不利，后果很严重；若将健康人误诊为疾病患者，就会给健康人增加不必要的心理压力。医疗数据往往是不平衡的，其分类研究在医学数据中具有重要意义。针对不平衡医疗数据集，较为传统的机器学习算法往往只考虑总体的分类准确率，会忽略少数类的准确率，若分类器将少数类预测为多数类（即将疾病患者误诊为健康人），往往会造成不可估量的损失，对少数类样本的信息研究，提高少数类的分类精度是文章的研究重点。

（二）国内外研究情况

目前，不平衡医疗数据的分类问题在国内外有不少的研究内容，Guang-Hui Fu等人在代谢组学数据分析中运用了一种新的算法（*SRS SVM*）进行生物标志物筛选，生物标志物识别是基于稀疏正则化的变量选择结合二次抽样（*SRS*），然后在选择变量空间中使用线性支持向量机（*SVM*）分类器进行分类，以获得最大的分类精度[4]。Yong-Huan Yun等人将一种基于排序聚合的变量重要分析法运用在代谢数据（儿童超重数据集和肾小管间质病变数据集）分析中，发现排序聚合的性能与使用所有变量相比有了较大的提高，预测精度更高[5]。在国外，Ashish Anand等人利用支持向量成功地实现了基于6个数据集的癌症子类型预测的类智能特征选择和多类分类框架[6]。Moloud Abdar等人采用了一种新的基于决策树的算法，考虑了更多的一般因素，对于预测肝病有较高的预测精度[7]。

研究类不平衡数据问题的方法除算法层面的特征选择外，也常从数据层面和判别准则层面进行研究。从数据层面来看，它的解决策略是对数据的采样，也是对数据预处理的过程。通过采样技术，使得不平衡数据达到平衡。

常用的采样技术是过采样、欠采样和混合采样方法。孟军[8]从不平衡数据集的采样法入手，设计出约束型欠采样法，提高了模糊分类系统的精确性。Kevin W Bowyer等人[9]提出的SMOTE算法，通过在少数类中加入新的合成样本或随机去除噪声，一定程度上解决了过拟合的问题。虽然过采样少数类的样本可以平衡类的分布，但是其他问题（数据集中存在倾斜或者不平衡比较大）并没有得到解决。通常，类集群的定义不是很好，因为一些多数类样本可能会占用少数类样本空间。对类不平衡数据进行再平衡预处理后，分类精度往往会有所提高。

文章以不平衡冠心病数据为研究目对象，对不平衡数据进行统计建模与处理，以此来改善不平衡数据的分类效果，提高少数类的分类准确率。从中探讨不平衡数据处理在冠状动脉粥样硬化性心脏病数据分析中的应用价值，为冠状动脉粥样硬化性心脏病防治工作提供理论依据，使其能采取有效的防治措施，从整体上降低冠状动脉粥样硬化性心脏病的发病率。

## 二、数据来源

数据集包括21例冠心病（CHD）患者和51例健康志愿者。所有患者均来自中国云南省第一人民医院。另外，健康对照组51例健康成人均来自同一城市，无血缘关系。采用超高效液相色谱－高分辨质谱联用技术（UPLC-HRMS）检测了50种代谢产物。临床特征包括年龄、收缩压、舒张压、空腹血糖等。一般情况下，健康人样本比冠心病患者的样本更容易获得，所以这里的多数类代表的是健康人样本类，少数类代表的是冠心病患者样本类。本数据集无缺失数据。

## 三、方案设计

高维不平衡数据的主要特征是：变量维度高、样本少、数据共线性严重、数据的不平衡度高。文章从数据层面、算法层面和评价标准3个不同层面对高维不平衡数据进行统计建模与分析。数据层面上采用SMOTE等再平衡策略和变量选择对数据进行预处理；从算法层面上，采用支持向量机，人工神经网络等分类算法；从评价标准层面，采用召回率 *recall* 、查准率 *precision* 、 $F-measure$ 值、 $G-mean$ 值、 *AUC* [5-6]等标准来度量不平衡数据的分类性能。

文章针对不平衡冠心病及其并发症数据，从三个层面进行分析，提高分类精度。具体流程如下图1。

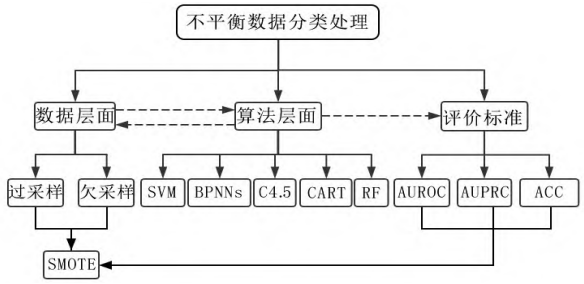


图1 方案设计流程图

## 四、高维不平衡数据的统计建模与分析

文章以冠心病数据为例，从数据、算法和评价准则的角度，按照图1的实验设计方案，对高维类不平衡数据的统计建模进行如下统计分析。

（一）类间不平衡比对于分类的影响

在理想情况下，当样本中不存在不平衡和其他干扰因素时，分类器对于数据集能够精确无误的进行分类。但是，本数据集为不平衡数据集，其不平衡比为51:21，约为2.5:1，变量个数为50。文章选择了五种分类器，支持向量机（*SVM*）、神经网络（*NN*）、决策树（*CART*）、C4.5和随机森林（*RF*）分别进行分类处理。为了对比，人为剔除少数类（即冠心病患者样本）的个数，使其不平衡比为51:10，约为5:1，变量个数为50。文章仍选取上述五种分类器进行分类处理，采用的评价标准为 *ACC*， *AUROC* 和 *AUPRC*，结果如表1。

表1 不平衡比为51:21和51:10时，五种分类器分类结果

| 平衡比 | 评价指标 | SVM | BPNNs | CART | C4.5 | RF |
|---|---|---|---|---|---|---|
| 51:21 | AUROC | 0.732 | 0.888 | 0.500 | 0.500 | 0.683 |
| | AUPRC | 0.579 | 0.818 | 0.304 | 0.304 | 0.555 |
| | ACC | 0.695 | 0.622 | 0.695 | 0.695 | 0.782 |
| 51:10 | AUROC | 0.593 | 0.775 | 0.500 | 0.500 | 0.500 |
| | AUPRC | 0.183 | 0.749 | 0.157 | 0.157 | 0.157 |
| | ACC | 0.842 | 0.836 | 0.842 | 0.852 | 0.842 |

由表1可以得出，对于不平衡比为51:21，约为2.5:1的冠心病数据集，支持向量机分类器和 *BPNNs* 分类器的 *AUROC* 和 *AUPRC* 的值均大于其他三种分类器，但是，*BPNNs* 分类器的 *ACC* 值小于其他四种分类器且这四种分类器结果相同，依据分类器得出的混淆矩阵可知，五种分类器在数据出现不平衡时，对于少数类均出现错分的情况。

对于不平衡比约为5:1的冠心病数据集，支持向量机分类器和 *BPNNs* 分类器的 *AUROC* 和 *AUPRC* 的值均大于其他三种分类器，但是， *BPNNs* 分类器的 *ACC* 值小于其他四种分类器，且这四种分类器结果相同，依据分类器得出的混淆矩阵可知，五种分类器在数据出现不平衡

比增大时，对于少数类也出现错分的情况。依据 *ACC* 的性质可知，对于同一数据集，由于少数类样本减少，不平衡比增大，*ACC* 值也会增大，极大地忽略了少数类对整个数据集的影响。

原始冠心病数据集存在不平衡，当不平衡比增加时，各分类器对于数据集的分类能力均在降低，均会忽略少数类样本对整个数据集的影响，这也说明不平衡比越高对于冠心病数据的分类结果影响越大，尤其是对于少数类分类结果影响甚深。因此，为了提高传统分类方法在类不平衡数据上的分类准确性，从数据层面，将不平衡数据进行修正，得到再平衡数据，以提高分类精度。

（二）再平衡对于分类的影响

基于再平衡思想，将使用 *SMOTE* 算法来修正不平衡数据，得到再平衡数据：利用最简单的 K 近邻算法，在两个同类 ( 少数类 ) 中进行线性插值，按照一定的规则在原有少数类样本的领域空间中插入新的样本，通过控制新生成的少数类样本数目来实现平衡样本集的目的。可以有效地解决传统过采样方法因决策区间过小而引起分类器过拟合的问题。

为了评估经过数据再平衡处理后的分类性能是否有提高，同样选取前述五种分类器进行处理，结果如表 2。

表 2 不平衡比为 51:21 的数据再平衡后，五种分类器分类结果

| 平衡比 | 评价指标 | SVM | BPNNs | CART | C4.5 | RF |
|---|---|---|---|---|---|---|
| 40:40 | AUROC | 0.861 | 0.750 | 0.500 | 0.500 | 0.583 |
| | AUPRC | 0.865 | 0.563 | 0.500 | 0.500 | 0.567 |
| | ACC | 0.750 | 0.500 | 0.500 | 0.500 | 0.583 |
| 30:30 | AUROC | 0.765 | 0.718 | 0.555 | 0.500 | 0.722 |
| | AUPRC | 0.810 | 0.535 | 0.555 | 0.500 | 0.673 |
| | ACC | 0.555 | 0.500 | 0.545 | 0.500 | 0.722 |

在实际计算中，文章得到两组再平衡数据，其平衡比为 40:40 和 30:30。为了能够直观看出数据再平衡后各个分类器性能的变化，对不平衡比为 51:21 的数据集通过算法再平衡前后在五种分类器分类结果进行可视化（如图 2）。

由图 2 可知，支持向量机（*SVM*）对于 40:40 这一个数据集，其 *AUROC*、*AUPRC* 和 *ACC* 值均比其他四种分类器分类精度高，分类效果好。对于 30:30 的数据集，支持向量机（*SVM*）分类器的 *AUROC* 和 *AUPRC* 值均比其他四种分类器分类精度高，分类效果好，而 *ACC* 值，随机森林结果比支持向量机结果好。C4.5 分类效果最差。

不平衡比为 51:21 的数据与再平衡后数据在五种分类器下分类精度如图 2 所示。再平衡后 *ACC* 的值只有 *SVM* 的结果有所提高，其他四个分类器的结果都比原数据的 *SVM* 值低。对于 *AUPRC* 的值，再平衡数据除了 *BPNNs* 的值比原数据的值低以外，其他四个分类器的值都比原数据的值高。而 *AUPRC* 的值，再平衡后数据 *SVM* 的值都有提高，*BPNNs* 的值比原数据的值低，*CART* 和 C4.5 的值与原数据几乎没有变化。*RF* 对于再平衡后比为 30:30 的数据值比原数据的值高，而对于再平衡后比为 40:40 的数据值比原数据的值低。

同样，人为剔除少数类（即冠心病患者样本）的个数，使其不平衡比为 51:10，约为 5:1，变量个数为 50。将数据进行再平衡处理，仍选取五种分类器，支持向量机（*SVM*）、神经网络（*ANNs*）、决策树（*CART*）、C4.5 和随机森林（*RF*）分别进行分类处理，结果如下（如表 3）。

表 3 不平衡比为 5:1 的数据再平衡后，五种分类器分类结果

| 平衡比 | 评价指标 | SVM | BPNNs | CART | C4.5 | RF |
|---|---|---|---|---|---|---|
| 40:40 | AUROC | 0.965 | 0.781 | 0.833 | 0.500 | 0.916 |
| | AUPRC | 0.967 | 0.636 | 0.887 | 0.500 | 0.857 |
| | ACC | 0.875 | 0.500 | 0.833 | 0.500 | 0.916 |
| 30:30 | AUROC | 0.802 | 0.871 | 0.777 | 0.500 | 0.666 |
| | AUPRC | 0.750 | 0.597 | 0.777 | 0.500 | 0.618 |
| | ACC | 0.500 | 0.500 | 0.692 | 0.500 | 0.666 |

得到两组再平衡数据，其平衡比为 40:40 和 30:30。同样的，为了能够直观看出数据再平衡后各个分类器性能的变化，对不平衡比为 51:10 的数据集通过算法再平衡前后在五种分类器分类结果进行可视化（如图 3）。

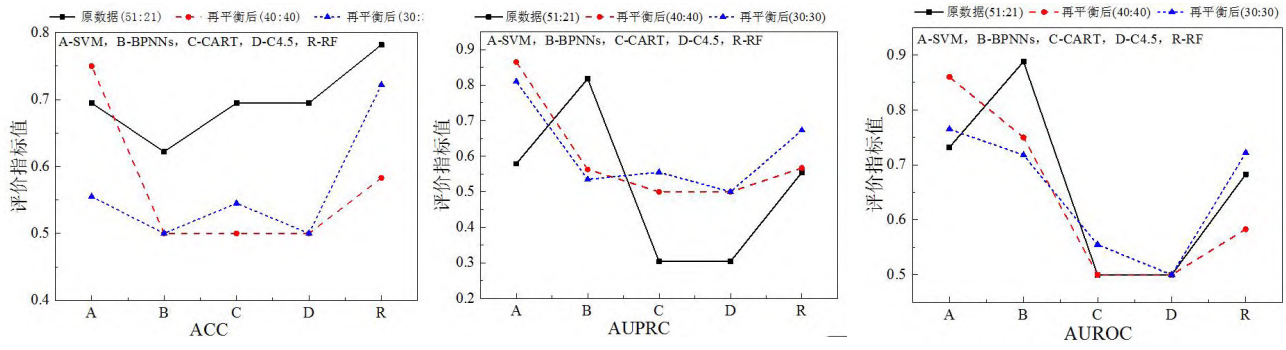


图 2 不平衡比为 51:21 的数据与再平衡后各分类器的结果

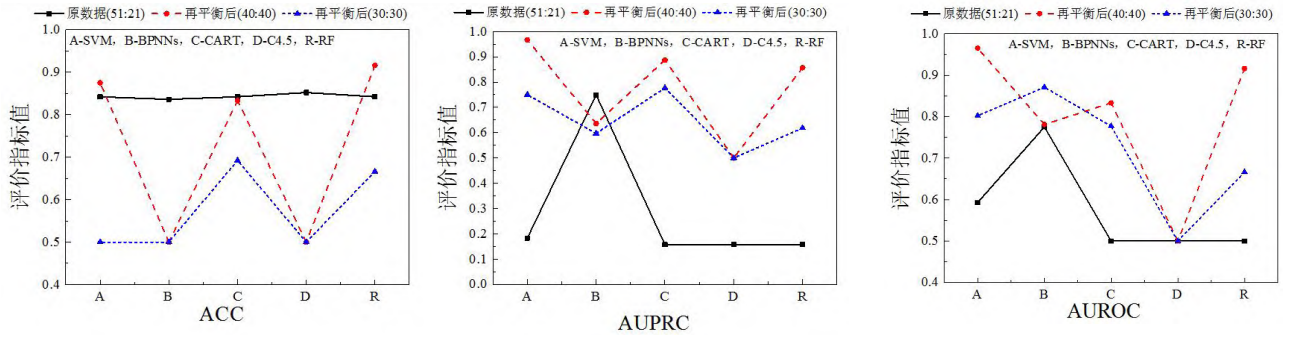


图 3 不平衡比为 51:10 的数据集与再平衡后各分类器的结果

A—*SVM*，B—*BPNNs*，C—*CART*，D—C4.5，R—*RF*

由图 3 可知，支持向量机（*SVM*）对于平衡比为 30:30 和 40:40 两个数据集，其 *AUROC* 和 *AUPRC* 值均比其他三种分类器分类精度高，分类效果好，而 *ACC* 值，随机森林结果比支持向量机结果好，只有在平衡比为 30:30 时，*BPNNs* 的 *AUROC* 比支持向量机结果好，C4.5 分类效果最差。与原数据再平衡相比，增加不平衡比再平衡，40:40 数据除 C4.5 之外，其他分类器分类结果均有提高，30:30 数据决策树算法分类效果有提高。

不平衡比为 51:10 的数据与再平衡后的数据在五种分类器下的结果如图 3 所示。对于 *ACC* 的值，再平衡后的数据只有 *SVM* 和 *RF* 的值比原数据高，其他三种分类器均比原数据低。对于 *AUPRC* 的值，再平衡后的数据只有 *BPNNs* 的值比原数据低，其他四个分类器的值都有所提高，尤其是 *SVM* 的值提高最多。而对于 *AUROC* 的值，再平衡后数据除了 C4.5 的值没有变化以外，其他四个分类器的值都得到了提高，同样的，*SVM* 的值提高最多。

与不平衡数据集相比，数据再平衡之后，支持向量机、决策树和随机森林分类结果均有改善，而 *BPNNs* 的分类结果反而呈现下降趋势。可以得出，*BPNNs* 分类器不适合处理再平衡数据，对其他三种分类器而言，数据在再平衡之后，分类性能有所提高，少数类的分类精度有所改善，并且对于该样本量较少的数据集，支持向量机算法的分类性能优于其他四种分类器的分类性能。

## 五、结语

类不平衡数据广泛存在于许多科学领域，如代谢组学。然而，一般分类器是针对类平衡或均匀分布的数据。传统的分类器的设计对平衡数据的分类具有一定的优势，如 *SVM*、随机森林、决策树等机器学习算法，文章研究的是冠心病不平衡数据，直接用这些分类器对数据集进行分类，可以得出在不平衡比增加时，传统的分类器会忽略少数类对整个数据集的影响，分类精度明显下降。

为了提高各分类算法在不平衡数据上的分类效果，对不平衡数据的分类研究现状进行了调研，常用的方法包括数据层的预处理方法和对算法本身进行改进的方法。文章主要从数据层面对不平衡数据进行再平衡处理，再平衡后的数据，经分类算法建模后，分类结果均有所提高。尤其是 SVM 算法，对于样本容量较小的数据集，其分类性能优于其他四种分类器。

文章为医疗不平衡数据提供了一种可行的处理手段，体现了不平衡数据对医疗数据分类的重要性。同时，也为其他领域不平衡数据处理理论增加一种可能的实现依据。

### 参考文献

[1] 徐玲，尹婷婷．冠心病冠状动脉粥样硬化发生的危险因素多因素 Logistic 分析 [J]. 临床和实验医学杂志，2019,18(06):626-629.

[2] 张建勇，丘志超．冠状动脉狭窄程度与冠心病危险因素的相关性 [J]. 哈尔滨医药，2019,39(02):110-111.

[3] 路航．早发冠心病的危险因素及冠脉病变特点分析 [J]. 中国疗养医学，2019,28(04):348-351.

[4] Fu Guanghui, Zhang Bingyang, KOU Hedan ,et al. Stable biomarker screening and classification by subsampling-based sparse regularization coupled with support vector machines in metabolomics[J]. Chemometrics and Intelligent Laboratory Systems,2017,160:22-31.

[5] Yang Ridong, LI Lin, CHEN Qiuyue,et al Prediction of disease-free survival in patients with hepatocellular carcinoma based on imbalance classification [J].Journal of Biomedical Engineering Research,2019,38(1):27-31.

[6] Ashish Anand, P.N. Suganthan. Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates[J].Journal of Theoretical Biology,2009,259:533-540.

[7] 彭驿茹．人工神经网络发展历史与训练算法概述 [J]. 科技传播，2018,10(21):129-130.

[8] 孟军．不平衡数据集分类算法的研究 [D]. 南京理工大学，2014.

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

**（作者单位：宗敏洁、江玉婷，黄河交通学院；吴愿交，西南交通大学希望学院）**