Journal of Sensor Technology and Application 传感器技术与应用, 2023, 11(6), 538-549
Published Online November 2023 in Hans. https://www.hanspub.org/journal/jsta
https://doi.org/10.12677/jsta.2023.116061

基于多准则决策的不平衡感知数据集成特征选择算法

王　刚[1]，任丽萍[1]，方　力[1]，徐维磊[2]

[1]南京航空航天大学自动化学院，江苏　南京
[2]南通思振电子科技有限公司，江苏　南通

收稿日期：2023年10月20日；录用日期：2023年11月17日；发布日期：2023年11月24日

摘　要

在数据挖掘领域，不平衡数据普遍存在。在许多情况下，这些数据通常具有高维性和类不平衡性。不平衡数据集特征属性分布失衡，会造成分类性能下降，数据的高维性则会导致学习算法非常耗时。针对这一问题，提出了一种基于组合采样和集成学习的特征选择方法。首先使用组合采样方法，处理类不平衡问题，重点合成少数类样本，在保证数据集达到平衡的前提下去除噪声样本，将集成特征选择建模为一个多准则决策过程，使用VIKOR方法得到特征重要性排序，然后在序列前向搜索特征的过程中，使用XGBoost算法的准确率作为评估特征子集优劣的指标，确定最优特征子集。选择AUC、G-mean和F-measure作为评价指标，通过在5组不平衡数据集进行实验，证实了所提算法具有更好的分类效果，且模型的鲁棒性更好。

关键词

不平衡数据分类，组合采样，多准则决策，VIKOR法，前向序列选择

Imbalance-Aware Data Based on Multi-Criteria Decision Making Integrated Feature Selection Algorithm

Gang Wang[1], Liping Ren[1], Li Fang[1], Weilei Xu[2]

[1]College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu
[2]Nantong Sizhen Electronic Technology Co., Nantong Jiangsu

Received: Oct. 20th, 2023; accepted: Nov. 17th, 2023; published: Nov. 24th, 2023

文章引用: 王刚, 任丽萍, 方力, 徐维磊. 基于多准则决策的不平衡感知数据集成特征选择算法[J]. 传感器技术与应用, 2023, 11(6): 538-549. DOI: 10.12677/jsta.2023.116061

Abstract

In the field of data mining, unbalanced data are prevalent. In many cases, these data are usually of high dimensionality and class imbalance. An unbalanced distribution of feature attributes in unbalanced datasets can cause degradation of classification performance, while the high dimensionality of the data can lead to very time-consuming learning algorithms. To address this problem, a feature selection method based on combinatorial sampling and integrated learning is proposed. Firstly, we use the combined sampling method to deal with the class imbalance problem, focus on synthesizing a few class samples, and remove the noise samples under the premise of ensuring that the dataset is balanced, model the integrated feature selection as a multi-criteria decision-making process, and use the VIKOR method to get the feature importance ranking, and then in the process of sequential forward searching for the features, we use the accuracy of the XGBoost algorithm as an indicator of the assessment of the feature subset's The optimal feature subset is determined by using the index of superiority and inferiority. AUC, G-mean, and F-measure are chosen as the evaluation indexes, and the proposed algorithm is confirmed to have a better classification effect and better robustness of the model through the experiments in five unbalanced datasets.

Keywords

Unbalanced Data Classification, Combined Sampling, Multi-Criteria Decision Making, VIKOR Method, Forward Sequence Selection

1. 引言

由于大数据时代的到来，各种领域都出现了大量的高维数据[1]，但在实际应用中，样本的分布通常是不均匀的。例如，在电信用户丢失数据识别中，丢失用户的比例通常远低于非丢失用户的比例。这种数据集通常被称为不平衡的数据集[2]。

高维且不平衡的数据给数据挖掘工作带来了前所未有的挑战。在分类任务中，这些数据包含了不相关特征和冗余特征。这些特征的存在使得学习算法耗费时间，同时也影响了分类性能。传统的分类算法更注重多数类的准确性，并倾向于将样本类别分类为多数类。这样一来，数据集的总体准确率会非常高，但并没有将我们更关注的少数样本进行正确划分。因此，这些传统的算法在处理不平衡的数据时效率非常低。特征选择作为一种降维技术，选择有意义的特征，并从数据集中去除不相关和冗余的特征，对于高维性的不平衡数据集，特征选择有时比分类算法更重要[3]。

目前，解决不平衡数据问题的研究主要集中在数据级方法、算法级方法和数据与算法结合方法[4]。数据级方法通常被称为外部方法，因为它们通过减少数据集中多数类样本或增加少数类样本来平衡数据。数据级方法主要分为过采样和欠采样。过采样技术增加了少数类的数量，常用的过采样方法有 Chawla 等[5]提出的 SMOTE 算法，该算法通过寻找样本的近邻集，在样本点与其近邻集随机选择的样本连线上合成新的样本点；Haibo 等[6]提出了 ADASYN 算法，根据数据分布自动确定每个少数类样本需要生成的样本数；Fernández-Navarro 等[7]提出了一种动态过采样方法来平衡数据集。欠采样技术减少了多数类的

数量，Yu 等[8]提出了一种基于蚁群优化思想的启发式欠采样方法来解决类不平衡问题；用于识别边界线和噪声数据的 Tomek Link 欠采样方法。算法级的方法通常被称为内部方法[9]，因为它采用新的分类算法或增强现有算法来解决不平衡数据产生的偏差。代表性的算法有代价敏感学习[10]和集成学习[11]。混合方法是数据级方法和算法级组合，克服了数据级方法和算法级方法中存在的问题，并实现更好的分类精度，Yong 等[12]提出了基于 k-means 聚类和遗传算法的采样方法，以突出不平衡数据集中少数类的性能；Galar 等[13]提出增强算法和随机欠采样算法相结合的方法。

将特征选择应用到不平衡数据，在近年来已经引起了研究者的关注。根据算法与后续学习算法的结合方式可以将特征选择算法分为过滤式、包裹式和嵌入式[14]。针对同一个数据集，不同的特征选择算法会产生不同的最优子集。在这种情况下，特征选择过程被称为不稳定过程。为了提高所选特征子集的稳定性，集成学习近年来得到了发展和广泛的应用。集成学习的思想是多个模型的聚合结果可能会比使用单一模型得到更好的结果。最初的集成方法是在分类模型中引入的，但 Saeys 等人[15]提出了为特征选择方法构建集成框架的想法。

针对不平衡数据，本文提出了一种基于组合采样和集成学习的特征选择方法。使用 SMOTE-Tomek 组合采样算法处理类不平衡问题后，将集成特征选择建模为一个多准则决策过程，将特征作为可选择的方案，特征选择方法作为选择标准，使用 VIKOR 方法得到特征重要性排序，然后在序列前向搜索特征的过程中，使用 XGBoost 算法的准确率作为评估特征子集优劣的指标，确定最优特征子集。每个特征算法可以被认为是局部最优的特征子集，采用集成方法可以提高特征选择算法的鲁棒性和精确性。通过与其他集成算法和基础特征选择算法进行对比实验，验证了本文方法的优越性。

2. 模型介绍

2.1. 组合采样算法

SMOTE-Tomek 组合采样算法，充分利用过采样和欠采样两种采样方法的优势。首先，通过 SMOTE 过采样方法，按设定的采样比例合成少数类样本。随后，将合成的少数类样本进行合并得到总的新增样本集，并将总的新增样本集加到原始数据集中得到最终扩展后的新数据集。最后，在新数据集中计算样本间的距离，如果两个不同类别的样本互为最近邻，则这两个样本构成一个 Tomek Link 对，在 Tomek Link 对中的两个样本，要么一个样本使噪声样本，要么两个样本都在类边界附近，因此找出新数据集中存在的 Tomek Link 对，并删除对中的样本，得到最终的平衡数据集。其中，SMOTE 方法对少数类样本的合成流程如下。

(1) 对于少数类中每一个样本 x，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其 k 近邻。

(2) 根据样本不平衡比例设置一个采样比例以确定采样倍率 N，对于每一个少数类样本 x，从其 k 近邻中随机选择若干个样本，假设选择的近邻为 x_n。

(3) 对于每一个随机选出的近邻 x_n，分别与原样本 x 按照式(1)的方式构建新的样本。

$$x_{new} = x + rand(0,1) \times (x_n - x) \tag{1}$$

SMOTE-Tomek 组合采样算法的优势在于：充分考虑了经过 SMOTE 过采样后新数据集可能出现噪声样本和边界样本的问题，通过寻找 Tomek Link 对的方式继续处理新的数据集，消除数据集中存在的噪点和边界问题，得到更高质量的数据集，提高训练模型的精确度，达到更好分类测试数据的目的。

2.2. 多准则决策与 VIKOR 算法

多准则决策(Multiple Criteria Group Decision Making, MCDM)是指一群决策者，按各自的偏好对备选

目标进行评价，从中寻求最满意的目标。目前，多准则决策分析方法主要用于解决选择、排序、有序分类、描述问题。

常用的多准则决策方法有 TOPSIS 和 VIKOR 方法，TOPSIS 通过构造多属性问题的理想解和负理想解，以方案靠近理想解和远离负理想解两个基准作为方案排序的准则，来选择最满意方案。VIKOR 决策方法是一种折衷排序方法，通过最大化群效用和最小化个体遗憾值对有限决策方案进行折衷排序。与 TOPSIS 相比，VIKOR 方法既有前者的正负理想值概念，还具备妥协个体和群体利益的折衷解，更适合解决存在利益冲突和测度单位不同的多准则决策问题。本文采用 VIKOR 方法作为解决方法。

VIKOR 方法确定好各指标的权值以及决策矩阵 $X \in R^{m\times n}$，其中 m 表示备选方案的数量，n 表示准则的数量。第一步，根据规划后的决策矩阵确定各准则的正理想解 r_j^+ 和负理想解 r_j^-：

$$r_j^+ = \left\{\max\left(x_{ij}\right) \mid i=1,2,\cdots,m;\ j=1,2,\cdots,n\right\} \tag{2}$$

$$r_j^- = \left\{\min\left(x_{ij}\right) \mid i=1,2,\cdots,m;\ j=1,2,\cdots,n\right\} \tag{3}$$

第二步，计算各方案到正理想解和负理想解的距离比值：

$$S_i = \sum w_j\left(\frac{r_j^+ - x_{ij}}{r_j^+ - r_j^-}\right), R_i = \max_j\left\{w_j\left(\frac{r_j^+ - x_{ij}}{r_j^+ - r_j^-}\right)\right\} \tag{4}$$

其中，w_j 是第 j 个指标的权重。

第三步，计算各个方案产生的利益比率值 Q，以便对备选方案进行排序：

$$S^+ = \max_i S_i, S^- = \min_i S_i \tag{5}$$

$$R^+ = \max_i R_i, R^- = \min_i R_i \tag{6}$$

$$Q_i = v\frac{S_i - S^-}{S^+ - S^-} + (1-v)\frac{R_i - R^-}{R^+ - R^-} \tag{7}$$

v 为多准则决策的决策机制系数，v 值体现了准则的重要程度或决策者的偏好。$v > 0.5$ 时，表示根据大多数人的意见，属于风险偏好型；$v = 0.5$ 时，兼顾大多数群体利益和少数反对意见，属于风险中性型；$v < 0.5$ 时，根据少数人所持的反对意见，属于风险厌恶型。一般取 $v = 0.5$。

第四步，根据 Q_i 值对被评价对象进行排序。Q_i 值越小，第 i 个被评价对象排序越靠前。

3. 基于多准则决策的不平衡数据集成特征选择算法

集成特征选择是借鉴了集成学习的思想，通过结合多个单一特征选择模型的输出得到最终的特征子集。各种研究表明，采用集成方法不仅可以避免关键信息的丢失，还可以提高特征选择算法的鲁棒性以及分类精度。根据所使用的基选择器是否为同一种类型可将集成特征选择方法分为同构集成特征选择方法和异构集成特征选择方法。同构集成特征选择方法利用数据的多样性，数据被均匀地划分为 K 份，对每一份数据都采用同一种特征选择方法进行特征选择，最后，将每一份数据特征选择后的特征合并为新的特征集合。异构集成方法利用了特征选择方法的多样性，采用了不同的特征选择方法对相同的数据进行处理。集成特征选择方法的关键是如何结合部分特征选择结果来获得最终特征子集的输出，主要的方法有集成特征子集和集成特征排序。集成特征子集最典型的方法是计算特征子集的交集或并集，如果一个特征被所有选择选择方法选择，那么它必定是一个高度相关的特征，但极端情况下可能会出现空集。每个特征采用不同的特征选择方法会得到不同的排名，集成特征排名的最简单的方法是通过计算每个特

征排名的中位数或平均值得到最终的排名。

本文的算法基于多准则决策方法构建的，将集成特征选择过程建模为一个多准则决策问题，采用VIKOR 方法对特征进行排序，即将数据集中的特征作为备选方案，特征选择方法作为标准，在图 1 中给出了具体流程。使用 SMOTE-Tomek 组合采样算法平衡数据，对平衡后的数据集用多种过滤式特征选择方法进行特征选择，根据每个特征选择方法得到特征权重构建决策矩阵，采用 VIKOR 方法得到最终的特征重要性排序，然后在序列前向搜索特征的过程中，使用 XGBoost 算法的准确率作为评估特征子集优劣的指标，确定最终的最优特征子集。

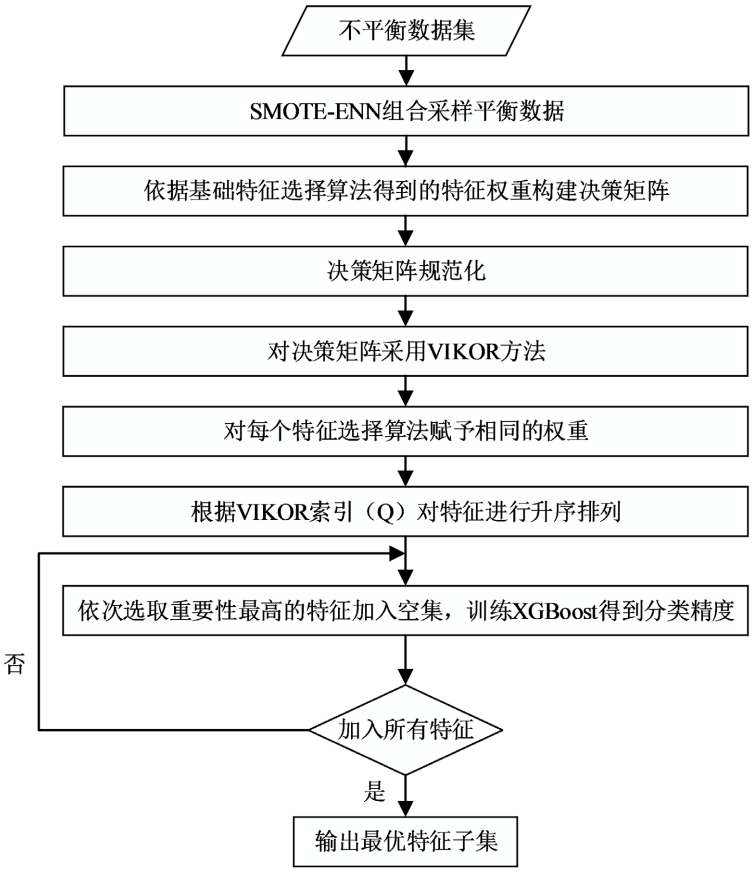

Figure 1. Flow chart of feature selection method in this paper
图 1. 本文特征选择方法流程图

对不平衡数据集采用 SMOTE-Tomek 组合采样方法做平衡化处理，对处理后的数据集采用不同特征选择算法计算特征权重，根据特征权重构建决策矩阵，因为不同特征选择算法得到的特征权重量纲不同，所以采用极差法对决策矩阵进行归一化处理，为了得到更加公正的结果，为每个特征选择算法的重要性赋以相同的权重，对决策矩阵采用 VIKOR 方法，根据 VIKOR 索引(Q)对特征进行升序排序，Q 值越小，特征越重要。然后引入包裹式思想，采用 XGBoost 的分类精度作为序列前向搜索策略的评价函数，初始特征集合为空集，从特征排序集合中选择重要性得分最高的特征放入初始集合中并计算 XGBoost 的分类精度，接下来将排名第二的特征放入选特征集合并计算 XGBoost 的分类精度，每次从余下的特征集合中选择一个特征，将其放入已选特征集合中并计算其分类精度直至遍历整个特征空间，通过对每组特征集合对应的分类精度进行比较，得到最终的最优特征子集。

通过一个示例展示本文所提算法的步骤。例如，有一个 7 个特征的数据集，采用 4 个特征选择算法，每个特征选择算法都会得到一个特征权重向量，基于此构造决策矩阵如下：

$$R=\begin{bmatrix}0.9377 & 0.2664 & 1.7410 & 212.5192\\ 0.4702 & 0.1088 & 1.7753 & 25.9614\\ 0.2262 & 0.0356 & 1.7066 & 4.1388\\ 0.7552 & 0.1396 & 1.7897 & 36.2948\\ 0.6832 & 0.0681 & 1.7897 & 19.4997\\ 0.6652 & 0.1301 & 1.7897 & 28.4255\\ 0.8359 & 0.2535 & 1.7897 & 219.3343\end{bmatrix} \tag{8}$$

其中，行数为特征个数，列数为使用特征选择算法的个数，每一列表示一个特征选择算法得到的特征权重向量，特征权重越大，特征重要性越高。

为了消除不同量纲带来的不可公度性，采用极差法对决策矩阵进行规范化，规范化后的矩阵如下：

$$E=\begin{bmatrix}0 & 0 & 0.5869 & 0.0316\\ 0.6569 & 0.6826 & 0.1738 & 0.8986\\ 1 & 1 & 1 & 1\\ 0.2564 & 0.5492 & 0 & 0.8506\\ 0.3576 & 0.8593 & 0 & 0.9286\\ 0.3829 & 0.5902 & 0 & 0.8871\\ 0.1431 & 0.9558 & 0 & 0\end{bmatrix} \tag{9}$$

根据式(2)、(3)得到正理想解 r_j^+ 和负理想解 r_j^-：

$$r_j^+=[1,1,1,1], r_j^-=[0,0,0,0] \tag{10}$$

MCDM 函数需要标准的权重，所以我们为每个特征选择方法设置相等的权重(1/n)，n 为采用特征选择算法的个数，这里 $w=[0.25,0.25,0.25,0.25]$。

根据式(4)计算各方案的群体效益值 S_i 和个体遗憾值 R_i：

$$S_i=[0.1546,0.60291,1.0,0.4140,0.5364,0.4651,0.0497] \tag{11}$$

$$R_i=[0.1467,0.2246,0.25,0.2126,0.2321,0.2218,0.03577] \tag{12}$$

为了兼顾大多数群体利益和少数反对意见 v 取 0.5，根据式(7)计算各个方案产生的利益比率值 Q_i，以便对备选方案进行排序：

$$Q_i=[0.3141,0.73191.0,0.6045,0.7144,0.65268,0.0] \tag{13}$$

其中 i 为特征数。

最后，根据特征在 Q_i 中的值按升序进行特征重要性排序，即 $f_7>f_1>f_4>f_6>f_5>f_2>f_3$。

4. 实验结果与分析

本文采用 ReliefF [16] [17]、MIC [18]、mRMR [19]、fisher-score [18]这四种过滤式特征选择算法构建基于 VIKOR 的集成特征选择模型，通过组合采样和集成特征选择算法，有针对性的解决不平衡数据种存在的高维性和类不平衡性问题，得到更好地分类效果。为了验证所提方法的性能，将本文的算法与其他特征选择算法比较。

4.1. 数据集

本文实验采用的是 NASA Metrics Data Program (MDP)数据库中的数据集，MDP 数据库是由美国国家航空航天局提供的用于软件研究的开放数据库，数据集基本信息如表 1 所示，其中不平衡度由多数类个数除以少数类个数得到。

Table 1. Basic information of the data set
表 1. 数据集的基本信息

数据集	特征数	少数类样本数	多数类样本数	不平衡度
CM1	38	42	302	7.19
MW1	38	27	236	8.74
PC1	38	61	674	11.05
KC3	40	36	164	4.56
JM1	22	1759	7832	4.45

4.2. 评价指标

由于不平衡数据集的类不平衡性，准确率这一指标没有意义。例如，样本数为 100 的数据集中，有 99 个好的产品和 1 个坏的产品，通过简单地预测好的产品，会获得 99%的准确率，然而我们真正关心的坏的产品却被忽略了。所以，本文采用 AUC、F-measure 和 G-mean 作为评价指标，本文定义少数类为正类，多数类为负类，利用混淆矩阵表示分类结果，如表 2 所示。

Table 2. Confusion matrix
表 2. 混淆矩阵

	预测为正类	预测为负类
实际为正类	TP (True Positive)	FN (False Negative)
实际为负类	FP (False Positive)	TN (True Negative)

精确率(Precision)，是指所有预测为正类样本中，正确预测为正类的样本所占的比率。计算公式如式(14)所示：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

召回率(Recall)，是指所有实际为正类的样本中，正确预测为正类的样本所占的比率。计算公式如式(15)所示：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

特异度(Specificity)是指所有负类样本中，被正确预测为负类的样本所占的比率。计算公式如式(16)所示：

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{TP}} \tag{16}$$

F-值(F-measure)是精确率和召回率的调和平均数。因为精确率和召回率是反比关系，所以采取折中的指标来评价。计算公式如式(17)所示：

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

F-mean 是召回率与特异度的综合指标，可用来评价处理不平衡数据的模型的表现情况，计算公式为：

$$\text{G-mean} = \sqrt{\text{recall} * \text{specificity}} \tag{18}$$

AUC (Area Under the Curve)指的是 ROC 曲线下方面积。ROC 曲线(receiver operating characteristic curve)，全称为受试者工作特征曲线。它以假正率 $\frac{FP}{TN+FP}$ 为横坐标，假正率 $\frac{TP}{TP+FN}$ 为横坐标绘制的曲线。

对于一个数据集，其预测结果对应 ROC 曲线上的一个点。通过调整阈值，可得到一条经过(0, 0)和(1, 1)的曲线，曲线下方的面积即为 AUC 的值。AUC 的取值范围为 0~1，当 AUC 为 0.5 时，跟随机猜测的一样，代表模型无意义，小于 0.5 则表示还不如随机猜测的结果。AUC 值越大，说明该模型的性能越好。因此，越靠近坐标系的左上角，表示模型的性能越好。

4.3. SMOTE-Tomek 算法数据不平衡处理

为了验证组合采样方法的有效性，将原始数据经 SMOTE、ADASYN、SMOTE-Tomek 进行处理，分别使用随机森林分类算法对平衡后的数据集进行训练和实验。具体实验结果如下。

Table 3. Experimental results of different sampling algorithms
表 3. 不同采样算法实验结果

评价指标	输入数据	CM1	MW1	PC1	KC3	JM1	Average
AUC	原始数据	0.5304	0.5665	0.6410	0.6735	0.6260	0.60748
	SMOTE	0.7643	0.6729	0.8123	0.7903	0.6839	0.74474
	ADASYN	0.7117	0.6604	0.825	0.7814	0.6754	0.73078
	SMOTE-Tomek	**0.7951**	**0.6812**	**0.8320**	**0.8258**	**0.7085**	**0.76852**
F1-measure	原始数据	0.1600	0.1653	0.2537	0.3670	0.2383	0.23686
	SMOTE	0.2273	**0.4000**	0.2800	0.4400	0.3497	0.34222
	ADASYN	0.1403	0.3362	0.3300	0.3670	0.3497	0.30464
	SMOTE-Tomek	**0.2532**	**0.4000**	**0.3496**	**0.4541**	**0.3649**	**0.36154**
G-mean	原始数据	0.5000	0.5448	0.6102	0.6618	0.6105	0.58546
	SMOTE	0.7586	**0.6731**	0.7821	0.7731	0.6679	0.73096
	ADASYN	0.6317	0.6000	0.8126	0.8119	0.6735	0.70594
	SMOTE-Tomek	**0.7948**	**0.6731**	**0.8279**	**0.8218**	**0.7082**	**0.76516**

由表 3 可以看出，经 SMOTE-Tomek 组合采样方法处理后的数据，在三个评价指标上的评分基本高于采用其它采样方法处理数据的评分。对比未经处理的原始数据，经过 SMOTE-Tomek 组合采样方法处理后的数据，AUC 值提高了 16%，F1-measure 值提高了 13%，G-mean 提高了 18%。充分说明了对不平衡数据进行平衡处理的重要性。并且对比其他采样方法，SMOTE-Tomek 组合采样方法在各方面的表现都是最好的。

4.4. 与基础特征选择方法比较

4.3 节验证了 SMOTE-Tomek 组合采样方法的有效性，现进一步验证本文所提集成特征选择方法对分

类模型的提升。在采用组合采样方法平衡数据集后，将本文提出的基于多准则决策的集成特征选择算法与经典的 ReliefF、MIC、mRMR、fisher-score 算法进行对比。实验使用随机森林分类算法对经过特征选择处理后的数据集进行训练和测试。具体实验结果如下。

Table 4. Compares the results with the basic feature selection method
表 4. 与基础特征选择方法对比结果

评价指标	算法	CM1	MW1	PC1	KC3	JM1	Average
AUC	ReliefF	0.7920	0.7541	0.7820	0.8448	0.7471	0.7840
	MIC	0.8155	0.7017	0.7972	0.7502	0.7236	0.7576
	mRMR	0.8127	0.6770	0.8216	0.7359	0.7063	0.7507
	fisher-score	0.8243	0.7367	0.8083	0.8312	0.7489	0.7899
	本文算法	**0.8573**	**0.8417**	**0.8714**	**0.8835**	**0.7517**	**0.8411**
F1-measure	ReliefF	0.3153	0.4000	0.3538	0.5718	0.4197	0.4121
	MIC	0.3865	0.3448	0.4060	0.4623	0.4215	0.4042
	mRMR	0.3285	0.2625	0.4175	0.3734	0.4215	0.3607
	fisher-score	0.3670	0.4400	0.4444	0.4623	0.4559	0.4339
	本文算法	**0.4400**	**0.4823**	**0.4931**	**0.4966**	**0.4695**	**0.4763**
G-mean	ReliefF	0.7731	0.6731	0.7844	0.8216	0.6381	0.7381
	MIC	0.8099	0.7266	0.7748	0.7451	0.6429	0.7399
	mRMR	0.8449	0.6640	0.8162	0.7203	0.6304	0.7352
	fisher-score	0.8106	0.7412	0.7909	0.8255	0.6592	0.7655
	本文算法	**0.8597**	**0.8332**	**0.8691**	**0.8713**	**0.6997**	**0.8266**

由表 4 可以看出，本文算法在五个数据集的三个评价指标中评分都是最高的，其中在 AUC 和 G-mean 指标上的均值都超过了 0.8，说明集成特征选择方法得到的特征子集具有比单一特征选择方法更好的性能，解决了单一特征选择方法在面对复杂多样数据时的局限性。

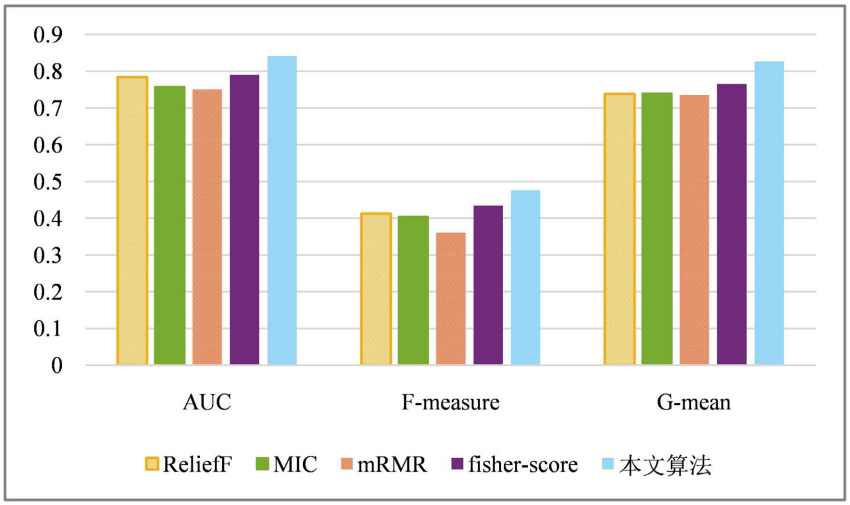

Figure 2. Compares with the basic feature selection method
图 2. 与基础特征选择方法对比

从图 2 则更直观地看出，本文算法明显优于 ReliefF 方法、MIC 方法、mRMR 方法和 fisher-score 方法，原因在于本文方法综合考虑了特征选择方法的优劣。在四种单一特征选择算法中，mRMR 方法算法表现较差，在 G-mean 均值上，ReliefF 方法、MIC 方法、mRMR 方法和 fisher-score 方法的结果较为接近。

4.5. 与集成特征选择方法比较

将本文方法与目前较新的集成特征选择算法进行比较，包括 E-max [20]、E-mean [20]、HFS。实验结果如表 5 所示。

Table 5. Compares the results with the integrated feature selection method
表 5. 与集成特征选择方法对比结果

评价指标	算法	CM1	MW1	PC1	JM1	KC3	Average
AUC	E-max	0.8094	0.8217	0.8319	0.7345	0.8376	0.8070
	E-mean	0.8155	0.8229	0.8432	0.7388	0.8407	0.8122
	HFS	0.8377	0.7637	0.8014	0.6956	0.8167	0.7830
	本文算法	**0.8573**	**0.8417**	**0.8714**	**0.7517**	**0.8835**	**0.8411**
F1-measure	E-max	0.3333	0.4375	0.3333	0.3897	0.4709	0.3929
	Es-mean	0.3857	0.4145	0.3902	0.4134	0.4718	0.4151
	HFS	**0.4965**	**0.5739**	0.4216	0.3121	0.418	0.4444
	本文算法	0.4400	0.4823	**0.4931**	**0.4695**	**0.4966**	**0.4763**
G-mean	E-max	0.7903	0.8124	0.8229	0.6515	0.8349	0.7824
	Es-mean	0.7998	0.8195	0.8372	0.6657	0.8391	0.7923
	HFS	0.8264	0.7469	0.7933	0.6229	0.8126	0.7604
	本文算法	**0.8597**	**0.8332**	**0.8691**	**0.6997**	**0.8713**	**0.8266**

由表 5 可以看出，本文提出的基于多准则决策的集成特征选择方法，在五个数据集的三个评价指标的得分大多都高于其他集成特征选择方法。虽然在 CM1 和 MW1 数据集中，本文算法在 F1-measure 这一个值上略低于 HFS 算法，但是综合其他两个指标依然可以说明本文算法在 CM1 和 MW1 数据集上具有很好的分类性能。可视化图如图 3 所示。

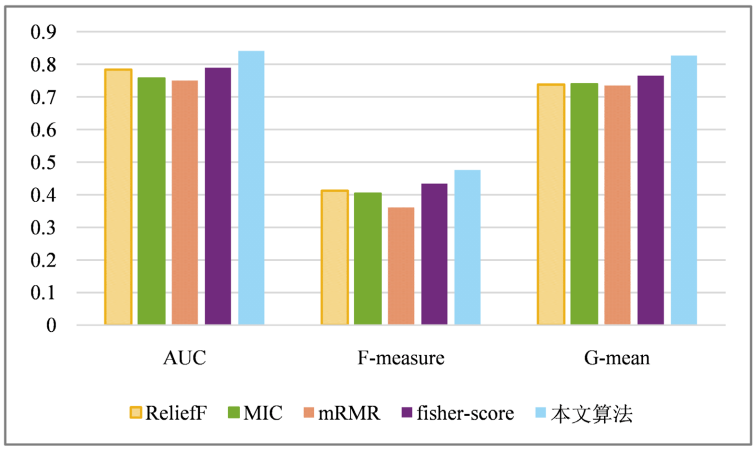

Figure 3. Compares with the integrated feature selection method
图 3. 与集成特征选择方法对比

5. 结论

针对不平衡数据的高维性和不平衡性，本文通过 SMOTE-Tomek 算法解决类不平衡性，其次通过将特征选择建模成一个多准则决策过程解决高维性，在该方法中，根据多种特征选择方法获得特征权重并构建决策矩阵后，将该数据作为决策数据应用到 VIKOR 方法中，得到特征重要性排序，并在此基础上引入包裹式思想，找到最优特征子集。选择 AUC、F-measure 和 G-mean 作为评价指标，在 NASA MDP 数据集上与单一特征选择方法和其他集成特征选择方法进行对比，实验表明本文所提方法可以明显提高少数样本的分类指标值，模型的鲁棒性更好。

基金项目

国家重点研发计划项目(2020YFB1710502)。

江苏省重点研发计划(社会发展)——城市轨道交通设施的监控、巡检和应灾综合系统科技示范。

参考文献

[1] Liu, Q., Lu, G.Y., Huang, J.R. and Bai, D.X. (2020) Development of Tunnel Intelligent Monitoring and Early Warning System Based on Micro-Service Architecture: The Case of AnPing Tunnel. *Geomatics, Natural Hazards and Risk*, **11**, 1404-1425. https://doi.org/10.1080/19475705.2020.1797906

[2] BenSaid, F. and Alimi, A.M. (2021) Online Feature Selection System for Big Data Classification Based on Multi-Objective Automated Negotiation. *Pattern Recognition*, **110**, Article ID: 107629. https://doi.org/10.1016/j.patcog.2020.107629

[3] Jia, W., Sun, M., Lian, J. and Hou, S.J. (2022) Feature Dimensionality Reduction: A Review. *Complex &Intelligent Systems*, **8**, 2663-2693. https://doi.org/10.1007/s40747-021-00637-x

[4] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[5] He, H.B., *et al*. (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008 *IEEE International Joint Conference on Neural Networks* (*IEEE World Congress on Computational Intelligence*), Hong Kong, 1-8 June 2008, 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

[6] Fernández-Navarro, F., Hervás-Martínez, C. and Antonio Gutiérrez, P. (2011) A Dynamic Oversampling Procedure Based on Sensitivity for Multi-Class Problems. *Pattern Recognition*, **44**, 1821-1833. https://doi.org/10.1016/j.patcog.2011.02.019

[7] Yu, H.L., Ni, J. and Zhao, J. (2013) ACOSampling: An Ant Colony Optimization-Based Undersampling Method for Classifying Imbalanced DNA Microarray Data. *Neurocomputing*, **101**, 309-318. https://doi.org/10.1016/j.neucom.2012.08.018

[8] Yen, S.J. and Le, Y.S. (2009) Cluster-Based Under-Sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications*, **36**, 5718-5727. https://doi.org/10.1016/j.eswa.2008.06.108

[9] Chawla, N.V. (1996) Data Mining for Imbalanced Datasets: An Overview. In: Maimon, O. and Rokach, L., Eds., *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, 123-140. https://doi.org/10.1007/0-387-25465-X_40

[10] Yong, Y. (2012) The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm. *Energy Procedia*, **17**, 164-170. https://doi.org/10.1016/j.egypro.2012.02.078

[11] 邹春安, 王嘉宝, 付光辉. MetaCost 与重采样结合的不平衡分类算法——RS-MetaCost[J]. 软件导刊, 2022, 21(3): 34-41.

[12] Li, J., Cheng, K., Wang, S., *et al*. (2017) Feature Selection: A Data Perspective. *ACM Computing Surveys*, **50**, 1-45. https://doi.org/10.1145/3136625

[13] Saeys, Y., Abeel, T. and Van de Peer, Y. (2008) Robust Feature Selection Using Ensemble Feature Selection Techniques. In: Daelemans, W., Goethals, B. and Morik, K., Eds., *ECML PKDD* 2008: *Machine Learning and Knowledge Discovery in Databases*, Springer, Berlin, 313-325. https://doi.org/10.1007/978-3-540-87481-2_21

[14] Chai, J. and Ngai, E.W.T. (2020) Decision-Making Techniques in Supplier Selection: Recent Accomplishments and What Lies Ahead. *Expert Systems with Applications*, **140**, Article ID: 112903. https://doi.org/10.1016/j.eswa.2019.112903

[15] Acuña-Soto, C.M., Liern, V. and Pérez-Gladish, B. (2019) A VIKOR-Based Approach for the Ranking of Mathemati-

cal Instructional Videos. *Management Decision*, **57**, 501-522. https://doi.org/10.1108/MD-03-2018-0242

[16] Kononenko, I. (1994) Estimating Attributes: Analysis and Extensions of RELIEF. In: Bergadano, F. and De Raedt, L., Eds., *Machine Learning*: *ECML*-94, Springer, Berlin, 171-182. https://doi.org/10.1007/3-540-57868-4_57

[17] Kanimozhi, U. and Manjula, D. (2017) An Intelligent Incremental Filtering Feature Selection and Clustering Algorithm for Effective Classification. *Intelligent Automation & Soft Computing*.

[18] 施启军, 潘峰, 龙福海, 李娜娜, 苟辉朋, 苏浩辂, 谢雨寒. 特征选择方法研究综述[J]. 微电子学与计算机, 2022, 39(3): 1-8.

[19] Drotár, P., Gazda, M. and Vokorokos, L. (2019) Ensemble Feature Selection Using Election Methods and Ranker Clustering. *Information Sciences*, **480**, 365-380. https://doi.org/10.1016/j.ins.2018.12.033

[20] Jacob, S. and Raju, G. (2017) Software Defect Prediction in Large Space Systems through Hybrid Feature Selection and Classification. *The International Arab Journal of Information Technology*, **14**, 208-214.