Advances in Applied Mathematics 应用数学进展, 2022, 11(4), 2071-2079
Published Online April 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.114225

Hans汉斯

基于G-Mean加权随机森林算法的不平衡数据处理

邢　鸿，魏毅强，李晨龙*

太原理工大学数学学院，山西　晋中

收稿日期：2022年3月24日；录用日期：2022年4月18日；发布日期：2022年4月27日

摘　要

分类是机器学习领域的一个热点研究内容，不平衡数据导致在分类时产生了很大困难。企业破产预测可以被归结为一个二分类问题，根据企业的一些特征对企业的状态做出预测，可以帮助企业作出更好的决策以减少企业的损失，但在企业破产数据中，破产企业只占很少的一部分，数据存在严重的不平衡。针对企业破产预测中的不平衡数据，本文提出了一种基于G-mean的加权随机森林算法(BSWRF)。对随机森林采集出来的不平衡子训练集，运用自助采样法进行欠采样，将多数类采样到和少数类一致，形成平衡子训练集，将CART决策树作为基分类器，在每个子训练集上进行训练，同时在包外估计样本上测试，根据G-mean为每棵决策树赋予权重，加权投票得到最终的分类器，提高了分类性能。选择台湾企业破产数据集进行实验，在G-mean、Recall和AUC评价指标上，BSWRF的分类效果都优于随机森林和AdaBoost算法。

关键词

不平衡数据，随机森林，G-均值，欠采样，自助采样法

Unbalanced Data Processing Based on G-Mean Weighted Random Forest Algorithm

Hong Xing, Yiqiang Wei, Chenlong Li*

College of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi

Received: Mar. 24th, 2022; accepted: Apr. 18th, 2022; published: Apr. 27th, 2022

Abstract

Classification is a hot research content in the field of machine learning. Unbalanced data leads to

*通讯作者。

文章引用: 邢鸿, 魏毅强, 李晨龙. 基于 G-Mean 加权随机森林算法的不平衡数据处理[J]. 应用数学进展, 2022, 11(4): 2071-2079. DOI: 10.12677/aam.2022.114225

great difficulties in classification. According to the data of enterprise bankruptcy, it can be concluded that there is only a part of the prediction of enterprise bankruptcy loss, but there is only a small part of the prediction of enterprise bankruptcy loss according to the data of enterprise bankruptcy. For the unbalanced data in the enterprise bankruptcy forecast, this paper proposes a weighted random forest algorithm (BSWRF) based on G-mean. For the unbalanced sub-training set collected from random forest, the self-help sampling method is used for under sampling. Most classes are sampled to be consistent with a few classes to form a balanced sub-training set. The cart decision tree is used as the base classifier to train on each sub training set. At the same time, it is tested on the Out-of-Bags samples. Each decision tree is given weight according to G-mean, and the final classifier is obtained by weighted voting, which improves the classification performance. The bankruptcy data set of Taiwan companies is selected for the experiment. In G-mean and AUC evaluation indexes, the classification effect of BSWRF is better than random forest and AdaBoost algorithm.

Keywords

Unbalanced Data, Random Forest, G-Mean, Under-Sampling, Bootstrap Sampling

1. 引言

当今社会是一个信息数据快速发展的时代，对得到的数据进行分类是很有必要的，随着机器学习的兴起，已经产生了很多分类模型，这些模型可以分为单一模型、集成模型、混合模型等，这些模型在分类问题中有着很好的效果。但是现实生活中得到的数据很多时候都是不平衡数据集[1]，很多分类算法在处理不平衡数据集时，模型偏向提高整体精度，导致少数类的分类效果会非常差。比如：信用卡欺诈检测[2]中信誉不良的客户很少，大部分客户信誉都是好的；企业破产预测中[3]，破产的企业占的比例是很小的；癌症基因检测时[4]，检测出来病人有癌症基因的概率可能仅为百万分之一。这些领域，存在着数据不平衡的问题，对于不平衡问题，人们更加关心少数类的分类结果，因此，对不平衡数据集的研究是非常有意义的。

为了在不平衡数据分类中取得较好的分类效果，国内外学者进行了相关研究，主要的方法有采样法和集成学习的方法。采样法主要有欠采样和过采样。欠采样方法除去样本中的部分多数类，使得多数类和少数类的数目接近时，进行学习。过采样方法增加样本中的部分少数类，当两类数目接近时，进行学习。Prusa 等人[5]运用随机欠采样来处理不平衡数据集，将随机欠采样应用于情绪分类的不平衡数据中，实验表明，随机欠采样技术显著提高了分类性能。Kang [6]等人将噪声滤波和欠采样技术结合在一起，该方法过滤出原始少数类中异常的实例，然后使用过滤之后的少数类来训练分类器。Chawla [7]提出一种过采样的方法，通过“合成”少数类来对少数类进行过采样，并进行实验表明将少数类过采样和多数类欠采样相结合的方法在部分数据集上的分类结果更好。Han [8]等人提出两种少数类过采样的方法，其方法只针对边界附近的少数样本进行过采样，实验结果表明，其方法比随机过采样方法效果更好。

集成学习通过构建多个基学习器，进行适当的组合形成性能好的强分类器。集成学习由于生成方式的不同，可以分为 Bagging 和 Boosting 两大类。Bagging 的代表算法是随机森林算法[9]，Boosting 的代表算法是 AdaBoost 算法[10]。Breiman [9]提出随机森林算法，并将其应用于分类问题。Wang 等人[11]将过

采样技术结合到 Bagging 模型中，并将其扩展到求解多分类问题上，提出一种集成模型分类方法 SMOTEBAGGING。任才溶[12]等人利用 K-means 算法对原始气象数据聚类，结合欠采样方法构建基于随机森林的预测模型，并将其应用于气象数据。Galar [13]等人综述了处理不平衡数据集方法的研究现状，并通过实验得出随机欠采样技术和集成方法结合的分类效果很好。在这些方法中主要存在两个方面的问题，在数据处理中运用采样技术，会导致部分有用信息被覆盖掉，在随机森林中，随机抽取的样本以及特征会导致不同的基分类器的分类性能不相同，每个基分类器赋予相同的权重会降低模型的效率，从而影响模型整体的分类性能。

本文针对企业破产预测中的不平衡数据进行研究，将破产预测归结为一个分类问题，根据企业的一些特征来判断企业的状态，本文提出基于 G-mean 的加权随机森林算法来对企业是否会破产做出分类。首先对随机森林中随机抽取出来的子训练集进行欠采样，使每一个子训练集变成平衡样本集，在每一个平衡样本集上用决策树训练一个基分类器，然后在包外样本上进行测试，得出包外样本上的 G-mean，将 G-mean 通过一个变换，作为每一个基分类器的权重，得到最终的分类模型，这种方法能够降低效果差的基分类器的决策能力，提高投票的公平性，从而提升模型性能。

2. 相关知识

2.1. 自助采样法

给定一个包含 n 个样本的数据集 D，每次从 D 中抽取一个样本放入 D'，然后将这个样本再放入 D 中，重复 n 次这样的操作，就得到了一个包含 n 个样本的数据集 D'，因为是有放回的，D 中的有一部分样本不会出现在 D' 中，可以通过下式计 D 中的样本不会出现在 D' 中的概率：

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n}=\frac{1}{\mathrm{e}}\approx 0.368 \tag{1}$$

上式表明，通过自助采样，数据集 D 中约有 36.8%的样本不会出现在数据集 D' 中，用 $\tilde{D}$ 表示这部分样本，在实际操作中，可以利用 $\tilde{D}$ 来验证模型，这种测试也成为“包外估计”。

2.2. 随机森林

随机森林是 Bagging 一种代表算法，随机森林以决策树为基分类器来构建模型，给定一个包含 n 个样本的数据集，通过自助采样法可以得到 N 个包含 n 个样本的子训练集，随机森林算法在每一个子训练集上训练出一个基决策树。并且随机森林算法在训练基分类器之前，加入了一个特征随机选择的步骤，假设决策树划分时，针对决策树的每个节点，有 f 个特征，先从这 f 个特征中随机选择 k 个特征作为一个子集，再从这个子集中选择一个最优特征来划分，参数 k 控制了随机性的引入成度。一般情况，k 的推荐值为 $k=\log_2 f$。对每一个基决策树的结果进行投票，选择数量最多的结果作为类别预测。最终的投票结果由下式表示：

$$H(x)=\arg\max_{y\in Y}\sum_{n=1}^{N} I\left(h_n(x)=y\right) \tag{2}$$

其中，$h_n(x)$ 是第 n 棵决策树的输出，$I(\cdot)$ 是一个示性函数，当决策树分类正确时，输出为 1，分类错误时，输出为 0。

2.3. AdaBoost

AdaBoost 算法是 Boosting 算法的代表算法，通过改变训练数据的概率分布，在不同的数据上训练出一系列基分类器。下面叙述 AdaBoost 算法。

输入：训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in X\subset R^n$，标记 $y_i\in Y=\{-1,+1\}$，基分类器。

1) 对训练数据集初始化权重分布：$W_1=(w_{11},w_{12},\cdots,w_{1N}),w_{1i}=\frac{1}{N},i=1,2,\cdots,N$；

2) 对 $m=1,2,\cdots,M$，使用 W_m 对训练数据集进行学习，得到基分类器 $G_m(x)$；

3) 计算每一个 $G_m(x)$ 在训练集上的分类误差率如下：

$$e_m(x)=\sum_{i=1}^{N}P(G_m(x_i)\neq y_i)=\sum_{i=1}^{N}w_{mi}I(G_m(x_i)\neq y_i) \tag{3}$$

4) 计算 $G_m(x)$ 系数 α_m

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m} \tag{4}$$

5) 更新训练数据集上面的权重分布：

$$W_{m+1}=(w_{m+1,1},\cdots,w_{m+1,i},\cdots,w_{m+1,N}) \tag{5}$$

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}=\exp(-\alpha_m y_i G_m(x)),i=1,2,\cdots,N \tag{6}$$

其中 Z_m 是规范化因子，可以使 W_{m+1} 变成一个概率分布。Z_m 表达式如下：

$$Z_m=\sum_{i=1}^{N}w_{mi}\exp(-\alpha_m y_i G_m(x_i)) \tag{7}$$

6) 将基分类器进行线性组合：

$$f(x)=\sum_{m=1}^{M}\alpha_m G_m(x) \tag{8}$$

输出：最终分类器 $G(x)$：

$$G(x)=sign(f(x))=sign\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right) \tag{9}$$

3. 基于 G-Mean 加权随机森林算法

基于 G-mean 加权随机森林算法，首先在数据采样层面，针对训练集中采样出来的 N 个子训练集，此时的子训练集是不平衡数据集，先进行欠采样，将每个子训练集中的多数类采样到和少数类一致，得到 N 个平衡子训练集，在每个子训练集上训练一个决策树，并且在包外估计样本上测试每个决策树的性能，计算出包外估计样本上的 G-mean，将 G-mean 做一个变换，作为每个决策树的权重，用来加权投票得到最终的分类器。BSWRF 的功能框架如图 1 所示，BSWRF 算法的具体步骤如下：

1) 给定一个原始的不平衡数据集 D，将数据集划分为训练集和测试集。

2) 在训练集上进行 Bootstrap sampling 得到 N 个子训练集，此时的子训练集都是不平衡数据集，未被采集到的样本会形成包外估计样本，即每个子训练集会对应一个 OOB (Out-of-Bag)测试集。

3) 在 N 个子训练集上进行欠采样来平衡数据集，对多数类样本使用随机不放回采样，将多数类采样到和少数类数目相同。得到 N 个平衡的子训练集。

4) 在 N 个平衡子训练集上进行训练，得到 N 个决策树分类器 $\{h_1(x),h_2(x),\cdots,h_N(x)\}$。

5) 将每个决策树在其对应的 OOB 测试集上进行性能测试，得到每个 OOB 测试集上的 G-mean，G-mean 的计算如下：

$$G_n = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{10}$$

6) G-mean 的大小由 TP 和 TN 来共同决定，只有多数类和少数类同时分类准确率高，G-mean 才会较大，根据 G-mean 来为每个决策树赋予不同的权重，为能够同时分出多数类和少数类的决策树分类器赋予较大的权重，权重的大小计算如下：

$$w_n = \lg \frac{G_n}{1-G_n} \tag{11}$$

7) 通过加权投票，得到最终的分类器，表示如下：

$$\begin{aligned} H(x) &= \arg\max_{y \in Y} \left\{ \sum_{n=1}^{N} I\left(h_n(x) = y\right) * w_n \right\} \\ &= \arg\max_{y \in Y} \left\{ \sum_{n=1}^{N} I\left(h_n(x) = y\right) * \lg \frac{G_n}{1-G_n} \right\} \end{aligned} \tag{12}$$

8) 将最终得到的分类器 $H(x)$ 用于测试集上去测试分类效果。

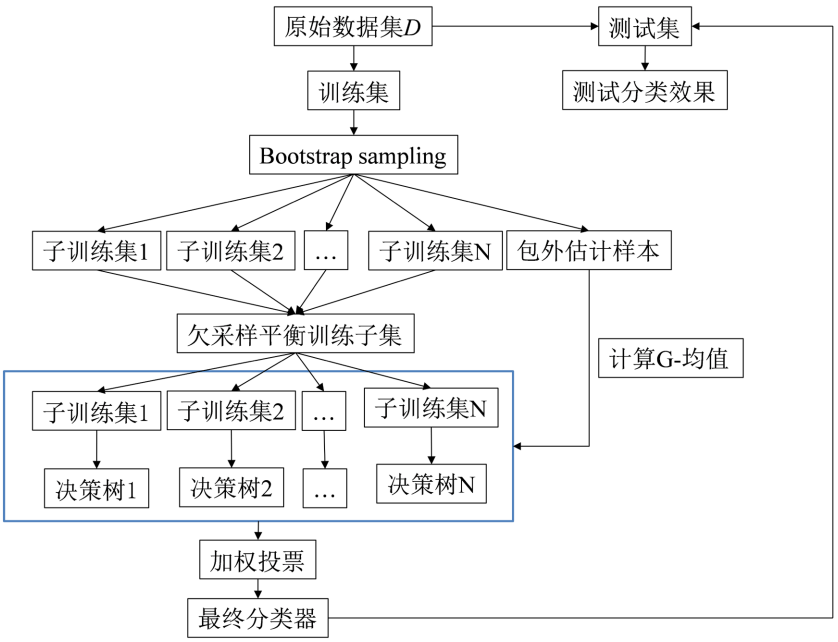

Figure 1. Functional block diagram
图 1. 功能框架

4. 实验结果与分析

4.1. 数据集

数据集选择台湾企业破产预测的数据集，该数据集是一个不平衡数据集，数据集中多数类样本和少数类样本的不平衡率比例达到了 29.995，数据集的信息如表 1 所示。

Table 1. Sample data set
表 1. 样本数据集

数据集	数据集大小	多数类样本	少数类样本	特征数量	不平衡率
Taiwan	6819	6599	220	95	29.995

4.2. 评价指标

在不平衡数据集的分类中，使用分类准确率这个评价指标没办法有效地评估模型的效果，因为它考虑的是总体的分类情况，并且认为少数类样本和多数类样本的重要程度是一样的，这明显是不符合实际的，在实际中，对于不平衡数据集的分类，人们更关心少数类的分类准确率。因此本文采用 AUC 值、召回率(Recall)和 G-mean 来评价模型分类效果。用混淆矩阵来更好的描述这些指标，混淆矩阵如表 2 所示：

Table 2. Confusion matrix
表 2. 混淆矩阵

类别	实际正类	实际负类
实际正类	*TP*	*FN*
实际负类	*FP*	*TN*

其中，*P* 表示少数类样本，*N* 表示多数类样本，*TP* 表示将少数类预测为少数类，*FN* 表示少数类预测为多数类，*FP* 表示将多数类预测为少数类，*TN* 表示将多数类预测为多数类。基于混淆矩阵，各评价指标如下：

$$\text{TPR}=\frac{TP}{TP+FN},\ \text{FPR}=\frac{FP}{FP+TN} \tag{13}$$

$$\text{G-mean}=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{14}$$

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

G-mean 用来表示多数类和少数类精度的几何平均，可以同时反映出少数类和多数类的分类情况。ROC (Receiver Operating Characteristic Curve)曲线是由 FPR 和 TPR 为横纵坐标画的曲线，AUC 值是 ROC 曲线下方的面积，在样本不平衡的情况下，AUC 依然可以对分类器做出合理的解释，ROC 曲线越靠近左上角越好，对应的 AUC 面积越大越好。Recall 值反应少数类中有多少被预测正确，Recall 值越大越好。

4.3. 实验结果与分析

首先将数据集中的多数类样本和少数类样本随机打乱之后，取 70%作为训练集，30%作为测试集，在训练集上训练分类模型，然后在测试集上验证模型性能。为了验证 BSWRF 算法的有效性，选取了两种经典的集成分类算法随机森林算法和 AdaBoost 算法进行对比。在构建模型时，统一使用 50 棵决策树作为基分类器，计算模型在测试集上的 G-mean 值、Recall 值和 AUC 值，来比较三种模型的分类效果。

从表 3 中可以看出本文提出的 BSWRF 与其他两种算法的对比效果，在对不平衡数据的处理上，性能有了进一步提升。BSWRF 算法与随机森林算法相对比，G-mean 提升了 83.8%，BSWRF 算法与 AdaBoost 算法相比，G-mean 提高了 61.7%。在 AUC 评价指标上，BSWRF 与另外两种算法的效果比较接近，但还

是优于另外两种算法，BSWRF 算法比随机森林算法相比提高了 4.42%，BSWRF 算法比 AdaBoost 算法提高了 1.1%。在 Recall 指标中，随机森林和 AdaBoost 算法的结果为 0.341 和 0.494，表明对少数类的预测准确率并不高，BSWRF 算法的 Recall 值达到了 0.897，表明 BSWRF 算法可以对少数类做出较好的预测，整体来说，BSWRF 相比于另外两种算法，在不平衡数据集上展示出了更好的分类性能。

Table 3. Performance comparison of three models
表 3. 三种模型性能对比

算法	G-mean	AUC	Recall
随机森林	0.475	0.906	0.341
AdaBoost	0.540	0.936	0.494
BSWRF	**0.873**	**0.946**	**0.897**

图 2 中绘制了 3 种算法的 ROC 曲线图，图中的蓝色曲线代表 AdaBoost 算法，绿色曲线代表随机森林算法，红色曲线代表本文所提出的 BSWRF 算法，从图 2 中，可以明显看出红色曲线更加靠近左上角，包围的右下角面积更大，绿色曲线所包围的右下角面积最小，从 ROC 曲线中，可以直观地看出，本文所提出的 BSWRF 算法在处理不平衡数据集上效果更好。

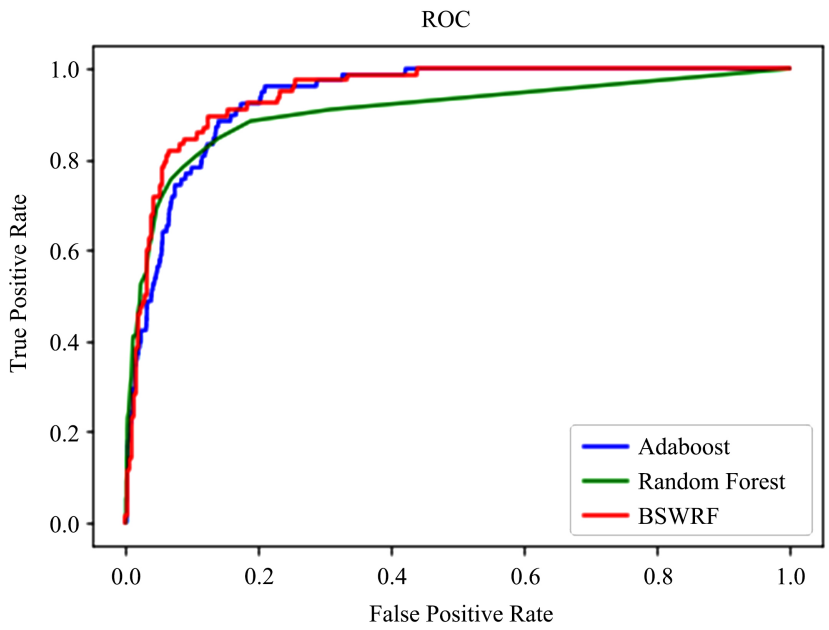

Figure 2. ROC curves of three models
图 2. 三种模型的 ROC 曲线图

本文还运用 ENN (Edited Nearest Neighbors)算法对数据进行清洗，剔除异常点之后，再运用 BSWRF 算法对数据进行分类。ENN 算法的原理是针对多数类中的每一个样本，如果其 K 个临近点中，有超过一半的点不属于多数类，则该样本就会被视为异常点，然后剔除掉。用 ENN 算法进行数据清洗之后的结果如表 4 所示：

Table 4. Sample data set after ENN algorithm cleaning
表 4. ENN 算法清洗之后的样本数据集

数据集	数据集大小	多数类样本	少数类样本	特征数量	不平衡率
Taiwan	6487	6267	220	95	28.486

表 4 展示了运用 ENN 算法进行数据清洗之后的数据集，在对多数类样本进行异常值剔除之后，多数类样本的数量减少，不平衡率由原始数据集种的 29.995 降低至 28.486。下面在这个数据集上运用 BSWRF 算法和另外两种对比算法进行分类。分类结果如表 5 所示：

Table 5. Performance comparison of three models
表 5. 三种模型性能对比

算法	G-mean	AUC	Recall
随机森林	0.595	0.935	0.456
AdaBoost	0.621	0.922	0.590
BSWRF	**0.893**	**0.949**	**0.915**

表 5 是通过 ENN 算法进行数据清洗之后，3 种模型的分类效果，从表 5 中可以看出，进行数据清洗之后，无论是从 G-mean，还是 AUC 评价指标上来说，BSWRF 算法的分类效果依然是最好的。并且将表 5 和表 3 进行对比，发现经过数据清洗之后，3 种模型上的分类效果都有所提升。3 种模型在 AUC 评价指标上都有小幅度的提升，在 G-mean 和 Recall 评价指标上，由于随机森林算法和 AdaBoost 算法本身分类效果较差，所以有较大提升，BSWRF 算法有小幅度提升。总之，在经过 ENN 算法之后，不平衡数据集上的整体分类效果有所提升，并且 BSWRF 算法依然展现出最好的分类效果。

5. 总结

本文提出了一种基于 G-mean 的加权随机森林算法，在随机森林采样出来的不平衡子集上运用欠采样方法，进行数据集平衡，得到平衡的训练子集，有效地解决了不平衡数据集中少数类样本信息难以学习的问题，并且在得到的平衡子训练集上进行学习，得到基决策树，然后在包外样本上对基决策树的性能进行评估，计算出 G-mean 值，并且根据 G-mean 来为每个基决策树赋予权重，得到基于 G-mean 的加权随机森林。并且在台湾破产企业数据集上进行了实验，与随机森林算法和 AdaBoost 算法进行对比，表明本文所提出的 BSWRF 算法在处理不平衡数据集问题上更有效。此外，本文算法还存在改进的空间，未来可以针对不平衡数据集的边界样本进行研究，提出针对边界数据的解决方法。

基金项目

国家自然科学基金(61901294)；山西省应用基础研究计划资助项目(201901D211105)。

参考文献

[1] 王乐, 韩萌, 李小娟等. 不平衡数据集分类方法综述[J]. 计算机工程与应用, 2021, 57(22): 42-52.

[2] Zakaryazad, A. and Duman, E. (2016) A Profit-Driven Artificial Neural Network (ANN) with Applications to Fraud Detection and Direction and Direct Marketing. *Neurocomputing*, **175**, 121-131. https://doi.org/10.1016/j.neucom.2015.10.042

[3] Kim, H.J., Jo, N.O. and Shin, K.S. (2016) Optimization of cluster-Based Evolutionary Undersampling for the Artificial Neural Networks in Corporate Bankruptcy Prediction. *Expert Systems with Applications*, **59**, 226-234. https://doi.org/10.1016/j.eswa.2016.04.027

[4] Salem, H., Attiya, G. and El-Fishawy, N. (2015) Gene Expression Profiles Based Human Cancer Diseases Classification. 2015 11*th International Computer Engineering Conference* (*ICENCO*), Cairo, 9-30 December 2015, 181-187. https://doi.org/10.1109/ICENCO.2015.7416345

[5] Prusa, J., Khoshgoftaar, T.M., Dittman, D.J., *et al.* (2015) Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. 2015 *IEEE International Conference on Information Reuse and Integration*, San Francisco, CA, 13-15 August 2015, 197-202. https://doi.org/10.1109/IRI.2015.39

[6] Kang, Q., Chen, X.S., Li, S.S., *et al.* (2016) A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification. *IEEE Transactions on Cybernetics*, **47**, 4263-4274. https://doi.org/10.1109/TCYB.2016.2606104

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

[8] Han, H., Wang, W.Y. and Mao, B.H. (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *International Conference on Intelligent Computing*, Springer, Berlin, Heidelberg, 878-887. https://doi.org/10.1007/11538059_91

[9] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[10] Freund, Y. and Schapire, R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119-139. https://doi.org/10.1006/jcss.1997.1504

[11] Wang, S. and Yao, X. (2009) Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models. 2009 *IEEE Symposium on computational Intelligence and Data Mining*, Nashville, TN, 30 March-2 April 2009, 324-331. https://doi.org/10.1109/CIDM.2009.4938667

[12] 任才溶, 谢刚. 基于随机森林和气象参数的 PM 2.5 浓度等级预测[J]. 计算机工程与应用, 2019, 55(2): 213-220.

[13] Galar, M., Fernandez, A., Barrenechea, E., *et al.* (2011) A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C* (*Applications and Reviews*), **42**, 463-484. https://doi.org/10.1109/TSMCC.2011.2161285