doi: 10. 3969 / j. issn. 1674 - 1242. 2024. 03. 002

基于双随机森林的发热待查智能诊断方法

杜建超[1]，丁俊瑶[1]，赵梦楠[1]，连建奇[2]，陈天艳[3]，WU Yuan[4]，周云[2]，石磊[3]
（1. 西安电子科技大学通信工程学院，陕西西安 710071；
2. 空军军医大学第二附属医院，陕西西安 710038；
3. 西安交通大学第一附属医院，陕西西安 710061；
4. Duke University Health System, Durham NC 27710）

【摘要】在机器学习预测模型中，不平衡数据集会降低少数类的预测准确性。针对发热待查数据集的不平衡特性，该文提出了一种基于 K-Means 聚类欠采样的双随机森林病因预测方法。首先通过 K-Means 聚类欠采样构建一个平衡数据集，并在此基础上创建一个基于 CART 投票机制的随机森林预测模型。然后对初始数据集用同样的方法创建一个随机森林预测模型。最后将两个随机森林预测模型联合，使用两者的 CART 一起投票预测。该文提出的方法增加了 CART 的数量，在保持原有数据集特性的同时，提高了少数类的投票权重。在发热待查数据集上的实验表明，该文所提方法不仅改善了少数类的预测性能，对其他类别的预测性能也有一定程度的提升。

【关键词】智能诊断；机器学习；发热待查；随机森林；不平衡数据集
【中图分类号】TP181　　【文献标志码】A　　文章编号：1674-1242（2024）03-0197-09

An Intelligent Diagnosis Method for FUO Based on Bi-random Forest

DU Jianchao[1], DING Junyao[1], ZHAO Mengnan[1], LIAN Jianqi[2], CHEN Tianyan[3], WU Yuan[4], ZHOU Yun[2], SHI Lei[3]
(1. School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi 710071, China;
2. The Second Affiliated Hospital of Air Force Medical University, Xi'an, Shaanxi 710038, China;
3. The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi 710061, China;
4. Duke University Health System, Durham NC 27710, U.S.A.)

【**Abstract**】In machine learning prediction models, imbalanced datasets reduce the accuracy of minority class predictions. A bi-random forest etiology prediction method based on K-Means clustering undersampling is proposed to address the imbalanced characteristics of the fever of unknown origin (FUO) dataset. Firstly, a balanced dataset is constructed through K-Means clustering undersampling, and a random forest prediction model based on the CART voting mechanism is created on this basis. Then, a random forest prediction model is also created using the same method for the initial dataset. Finally, two random forest prediction models are combined and their CART are used to vote together for prediction. The proposed method increases the number of CART, and enhances the voting weights of minority class while

收稿日期：2023-11-07。
基金项目：空军军医大学第二附属医院前沿交叉研究项目（2021QYJC-005）。
作者简介：杜建超（1977—），男，陕西省西安市人，副教授，博士生导师，从事智能图像处理、人工智能算法设计与应用研究。
通信作者：石磊，女，副主任医师，电话（Tel.）：13096937891；邮箱（E-mail）：dr.shilei@xjtu.edu.cn。
周云，女，副主任医师，电话（Tel.）：18009253837；邮箱（E-mail）：winzhouyun@126.com。

maintaining the characteristics of the original dataset. Experiments on FUO dataset show that the proposed method not only improves the prediction performance for minority class, but also improves the prediction performance for the other classes to a certain extent.

【Key words】 Intelligent Diagnosis; Machine Learning; Fever of Unknown Origin; Random Forest; Imbalanced Dataset

0　引言

发热待查（Fever of Unknown Origin，FUO）是指发热持续 3 周以上，经过至少 1 周系统全面的检查仍不能确诊的一组疾病。发热待查的病因诊断是当前医学界面临的一大难题 [1]。近年来，基于机器学习的医学疾病辅助诊断研究得到了快速发展[2-6]。从机器学习的角度来看，发热待查可以看作一个不平衡数据集上的多分类预测问题 [7]。由于发热待查的病因类型多达数十种，并且不同病因出现的频次相差数十倍，通常的机器学习算法在发热待查数据集上的表现欠佳。因此，找到一种针对发热待查不平衡数据集的有效分类方法，建立发热待查智能诊断模型，将能够帮助医生快速准确地确定病因。

传统的预测模型在处理不平衡数据集时，其预测结果会偏向于多数类，造成少数类的预测准确率很低，容易出现误判 [8-9]。在一些实际应用中，特别是在医学诊断领域，对少数类的准确预测是必要且十分重要的。例如，在发热待查的病因中，黑热病、感染性心内膜炎出现的概率很低，但是病情很紧急，如果出现误判，将带来严重后果。目前，解决不平衡数据集预测问题的方法主要从两个方面入手：一是通过采样方法调整数据集的平衡性，二是改进分类算法。文献 [10-12] 通过 SMOTE 过采样方法增加少数类样本，从而调整数据集的平衡性。然而，过采样方法虽然增加了少数类样本的数量，但是容易模糊正负样本的边界，从而使整体分类准确率下降 [13]。文献 [14-16] 通过欠采样方法删除部分多数类样来实现数据集的平衡，如随机欠采样、K 近邻欠采样等。然而，欠采样方法会不可避免地造成信息丢失问题，从而降低模型的整体分类性能。

因此，一些研究通过算法改进提高分类性能。此类方法的改进方式主要为代价敏感方法 [17] 和集成学习方法 [18]。前者的思想是根据不同的分类错误给予相应的惩罚力度，后者是以某种方式将多个基分类器组合起来共同决策，以获得比单个基分类器更好的分类效果。在集成学习中，随机森林 [19] 中的决策树机制使其容易通过改变决策树的数量或权重调整类间分类的结果，从而成为此类算法的主流。文献 [20] 提出了一种加权随机森林算法，赋予每棵树一定的权重，从而调整投票结果，提高分类算法的适应性。文献 [21] 提出了一种最近邻欠采样与随机森林联合的分类方法，通过最近邻算法对多数类进行欠采样，生成一个类间均衡的随机森林，从而在一定程度上增加了代表少数类的决策树数量。虽然最近邻算法简单，但其本质上是一种分类算法，在欠采样时难以选取出边界明确的中心样本，因此得到的均衡数据集样本不能很好地反映原始样本集的分布特点。K-Means 算法 [22] 是一种性能较好的聚类方法。通过 K-Means 算法对多数类样本进行若干次聚类，然后保留聚类中心附近的样本，删除与聚类中心距离较远的样本，可以实现性能优良的欠采样。

本文将 K-Means 聚类欠采样方法融合到随机森林的预测模型中，提出一种针对不平衡数据集的发热待查病因预测方法。利用 K-Means 聚类欠采样方法进行数据聚类和欠采样，形成一个新的均衡数据集，进而生成均衡的基分类器和随机森林。然后将该随机森林与通过初始数据集构建的随机森林联合形成双随机森林，并使用两者的决策树共同投票输出最终的预测结果。本文提出的预测方法在保持原始数据样本特征的同时，提高了少数类的投票权重，能够更好地预测出少数类样本的病因类别。

1　不平衡数据集

发热待查的病因有很多，分为感染性和非感染性两大类。感染性发热包括细菌、真菌、病毒、寄生虫等类别，非感染性发热包括血液病、肿瘤性疾

病、内分泌性发热、中枢性发热等类别。本文基于临床数据构建的数据集包含 560 条样本，分为 10 类感染性病因、6 类非感染性病因。每条样本均由人口学指标、伴随症状、体格检查、医生诊断结果等 186 项数据组成，这些数据项称为样本的特征。

在分类问题中，当数据集中各类样本的数量存在明显差异时，称为数据集不平衡。不平衡程度由下式给出。

$$\mathrm{IR}=\frac{\left|N^{\max}\right|}{\left|N^{\min}\right|} \tag{1}$$

式中，IR 为不平衡率；$\left|N^{\max}\right|$为数据集中样本数量最多的类别含有的样本数量；$\left|N^{\min}\right|$为数据集中样本数量最少的类别含有的样本数量。

图 1 展示了本文数据集中每种类别的样本数量。感染类和非感染类的不平衡比率 $\mathrm{IR}_1=393/167\approx2.35$。在具体的细分类别中，样本数量最多的“上呼吸道感染”类别和样本数量最少的“支原体感染”类别的不平衡比率 $\mathrm{IR}_2=97/11\approx8.82$，不平衡性十分严重。

2　预测模型

2.1　K-Means 聚类欠采样

K-Means 聚类欠采样过程以最少数类的样本数量为基准，对其他类别进行抽样。图 2 展示了 K-Means 聚类欠采样的效果。可以看出，K-Means 聚类欠采样在减少多数类样本的同时，有效保留了原始数据的分布特征，不仅实现了类间平衡，而且样本边界清晰，数据分布特征明显，有利于后续的分类和预测。

在发热待查数据集 $\boldsymbol{X}$ 中，设总的类别数量为 k，其中最少数类样本集记为 $\boldsymbol{X}_1$，其他 $k-1$ 个类别均称为多数类，其样本集记为 $\boldsymbol{\Delta}=\{\boldsymbol{X}_2,\boldsymbol{X}_3,\cdots,\boldsymbol{X}_k\}$。对每个多数类 $\boldsymbol{X}_j\in\boldsymbol{\Delta}$（$2\leqslant j\leqslant k$），以 $\boldsymbol{X}_1$ 的样本数量 $N_0=\left|\boldsymbol{X}_1\right|$ 为基准进行聚类欠采样。

对多数类 $\boldsymbol{X}_j$ 的欠采样过程如下。

（1）初始化。从 $\boldsymbol{X}_j$ 中随机选取 N_0 个样本

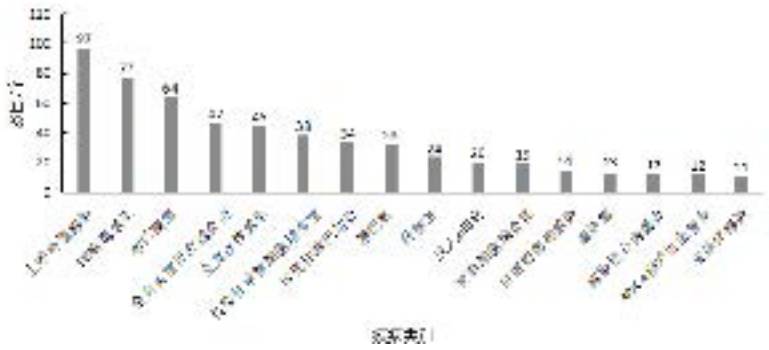
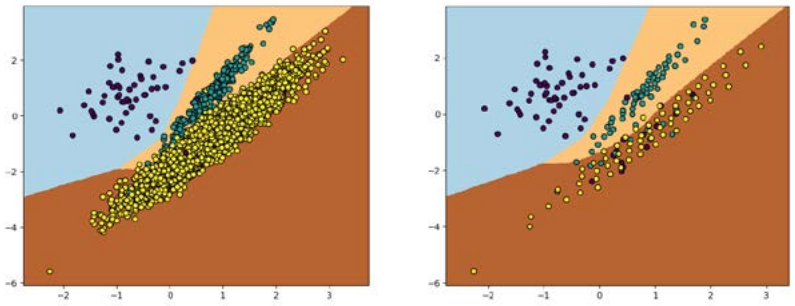


图 1　发热待查数据集分布

Fig.1　Distribution of FUO dataset



（a）原始数据分布　（b）K-Means 聚类欠采样后的数据分布

图 2　K-Means 聚类欠采样效果

Fig.2　K-Means clustering under sampling effect

$\hat{\boldsymbol{X}}_j = \{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_{N_0}\}$作为初始聚簇中心。

（2）计算距离。对$\boldsymbol{X}_j$中的每个样本$\boldsymbol{x}_i \in \boldsymbol{X}_j$，计算$\boldsymbol{x}_i$到每个聚簇中心$\hat{\boldsymbol{x}}_k (k = 1,2,\cdots,N_0)$的欧氏距离$d_{i,k}$。

（3）样本分簇。找到与$\boldsymbol{x}_i$距离最近的聚簇中心$\hat{\boldsymbol{x}}_k$，将$\boldsymbol{x}_i$并入以$\hat{\boldsymbol{x}}_k$为中心的聚簇$\psi(\hat{\boldsymbol{x}}_k)$。

（4）更新聚簇中心。当$\boldsymbol{X}_j$中的所有样本均完成分簇以后，将形成N_0个聚簇，根据以下公式重新计算每个聚簇$\psi(\hat{\boldsymbol{x}}_k)$的聚簇中心。

$$\tilde{\boldsymbol{x}}_k = \frac{1}{\left|\psi(\hat{\boldsymbol{x}}_k)\right|}\sum \boldsymbol{x}_k \tag{2}$$

式中，$\left|\psi(\hat{\boldsymbol{x}}_k)\right|$表示聚簇$\psi(\hat{\boldsymbol{x}}_k)$所含样本的数量，$\boldsymbol{x}_k \in \psi(\hat{\boldsymbol{x}}_k)$，$\sum \cdot$样证为向量加法。计算完成后，令$\hat{\boldsymbol{x}}_k = \tilde{\boldsymbol{x}}_k$，产生新的聚簇中心。

（5）聚类结束。若更新的聚簇中心$\hat{\boldsymbol{x}}_k (k = 1,2,\cdots,N_0)$均与更新前的聚簇中心相同，则聚类结束；否则，返回步骤（2）继续完成聚类。

（6）样本采样。找到每个聚簇$\psi(\hat{\boldsymbol{x}}_k)$中与聚簇中心$\hat{\boldsymbol{x}}_k$距离最近的样本$\boldsymbol{x}_k^*$，形成多数类$\boldsymbol{X}_j$的采样样本$\boldsymbol{X}_j^* = \{\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_{N_0}^*\}$，采样过程结束。

在采样过程中，步骤（2）中的欧氏距离是指两个样本之间的 2 范数，即

$$d_{i,k}(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_k) = \sqrt{\sum_n (x_i^n - \hat{x}_k^n)^2} \tag{3}$$

式中，x_i^n和$\hat{x}_k^n$分别为样本$\boldsymbol{x}_i$与$\hat{\boldsymbol{x}}_k$各个维度的数据。在步骤（6）的样本采样中，用距离聚簇中心最近的样本替代聚簇中心，因为聚簇中心可能是计算出来的插值样本点，而不是原始样本点。因此，为了与原始数据一致，在本算法中，选择与聚簇中心$\hat{\boldsymbol{x}}_k$距离最近的样本$\boldsymbol{x}_k^*$作为采样样本。

逐个完成每个多数类的聚类欠采样后，与少数类$\boldsymbol{X}_1$一起构成平衡数据集$\boldsymbol{X}^* = \{\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \boldsymbol{X}_3^*, \cdots, \boldsymbol{X}_k^*\}$。

2.2　双随机森林预测模型

随机森林由多棵相互独立的分类与回归树（Classification and Regression Tree，CART）组成，是一种基于决策树的集成学习方法[23-24]。随机森林的分类结果最终由所有的 CART 投票产生。因此，构建随机森林的关键是生成多棵 CART。

2.2.1　CART

每棵 CART 都是由原始样本集生成的一个副本训练得到的一棵二叉树。二叉树根据样本特征的取值产生分支，叶子节点为分类结果，由同一类别的部分样本组成。二叉树的数量越多，越有利于提高投票结果的准确性，但由于结果会收敛于一个稳定值，因此数量不需要太多。一般来说，二叉树的数量设置为 100 即可达到收敛要求。

训练 CART 的样本集通过对原始样本集进行有放回抽样的方式[25]获得。随机抽取一个样本，将其添加到一个新创建的副本集合中，再随机抽取下一个样本添加到副本集合中。若样本集中的样本数量为 N，则连续抽取 N 次，构成一个数据集副本。在抽样过程中，每次都从 N 个原始样本集中抽取一个样本，已经抽过的样本还可以再次被抽。由于每棵 CART 对应一个数据集副本，因此构建的副本数量与需要生成的 CART 数量相同。这种抽样方式扩展了样本集的容量，提高了随机森林的预测准确性和泛化能力。每棵 CART 都是一个弱分类器，它们共同组成一个随机森林。在预测时，由所有 CART 对待预测样本进行分类投票。票数最多的分类结果为最终的预测结果。

CART 的结构是二叉树形，由节点和边组成。节点包括根节点、内部节点和叶子节点。根节点和内部节点表示某一特征的分裂条件，叶子节点表示预测类别。CART 训练完成后，数据集中的每个样本都能根据其特征取值被划分到唯一的叶子节点，作为后续预测的基分类器。CART 的结构如图 3 所示。

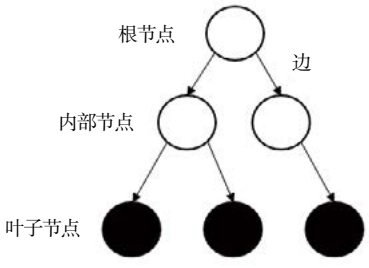


图 3　CART 的结构
Fig.3　The structure of CART

CART 是弱分类器，在训练时不采用全部特征，仅基于部分特征进行训练[26]。从样本的 M 个特征中随机选取 m 个特征，满足$0 \leqslant m \leqslant \log_2(M+1)$，可

以看出 $m << M$，以 m 个特征所对应的基尼系数为依据来训练和构建一棵 CART。

基尼系数代表以某一特征构建的分类模型的不确定程度，计算公式如下。

$$\mathrm{Gini}(\boldsymbol{O},\boldsymbol{a}_j)=\sum_{i=1}^{k}p_i(1-p_i)=1-\sum_{i=1}^{k}p_i^2 \tag{4}$$

式中，$\boldsymbol{O}$ 为一个抽样样本集，$\boldsymbol{a}_j$ 为某一特征项，k 为类别数量，$p_i\left(1\leqslant i\leqslant k\right)$为以 $\boldsymbol{a}_j$ 的某个取值为依据时第 i 个类别出现的概率。基于某个特征项计算得到的基尼系数越小，基于该特征项的分类确定程度越高，模型的分类性能越好。将具有最小基尼系数的特征项作为 CART 的根节点，然后在 CART 的两条分支上分别根据其他特征项的基尼系数依次递归便可生成一棵分类 CART。

设抽样得到的一个训练集$\boldsymbol{D}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，选取的特征集合$\boldsymbol{A}=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m\}$，其中 n 为样本数，对应 m 个特征的样本$\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{im}]^{\mathrm{T}}\in\boldsymbol{R}^m\left(1\leqslant i\leqslant N\right)$，特征$\boldsymbol{a}_j$ 在所有样本上的取值构成$\boldsymbol{a}_j=[x_{1j},x_{2j},\cdots,x_{nj}]^{\mathrm{T}}\in\boldsymbol{R}^n\left(1\leqslant j\leqslant m,n\leqslant N\right)$。

CART 的训练过程如下。

（1）遍历特征 $\boldsymbol{a}_j\in\boldsymbol{A}$，计算每个特征的分割基尼系数。

① 根据 $\boldsymbol{a}_j$ 的取值求分割点集合 $\boldsymbol{S}$。若 $\boldsymbol{a}_j$ 具有离散型特征，则$\boldsymbol{S}=\left\{s_l\mid s_l=x_{lj}\right\}(l=1,2,\cdots,n)$；若 $\boldsymbol{a}_j$ 具有连续型特征，则

$$\boldsymbol{S}=s_l\mid s_l=\frac{x_{lj}+x_{(l+1)j}}{2}\ \ (l=1,2,\cdots,n-1)$$

② 划分子集。根据分割点 s_l，将当前节点的样本集 $\boldsymbol{D}$ 划分为 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$。若 $\boldsymbol{a}_j$ 具有离散型特征，则划分方式如下。

$$\begin{aligned}\boldsymbol{D}_1&=\left\{\boldsymbol{x}_i\mid x_{ij}=s_l\right\}\\ \boldsymbol{D}_2&=\left\{\boldsymbol{x}_i\mid x_{ij}\neq s_l\right\}\end{aligned} \tag{5}$$

若 $\boldsymbol{a}_j$ 具有连续型特征，则划分方式如下。

$$\begin{aligned}\boldsymbol{D}_1&=\left\{\boldsymbol{x}_i\mid x_{ij}\leqslant s_l\right\}\\ \boldsymbol{D}_2&=\left\{\boldsymbol{x}_i\mid x_{ij}>s_l\right\}\end{aligned} \tag{6}$$

③ 对属于特征 $\boldsymbol{a}_j$ 的所有分割点 $s_l\in\boldsymbol{S}$，分别独立计算分割基尼系数。

$$\mathrm{Gini}(\boldsymbol{D},\boldsymbol{a}_j)\mid_{s_l}=\frac{\left|\boldsymbol{D}_1\right|}{\left|\boldsymbol{D}\right|}\mathrm{Gini}(\boldsymbol{D}_1,s_l)+\frac{\left|\boldsymbol{D}_2\right|}{\left|\boldsymbol{D}\right|}\mathrm{Gini}(\boldsymbol{D}_2,s_l) \tag{7}$$

式中，$\left|\boldsymbol{D}\right|$、$\left|\boldsymbol{D}_1\right|$、$\left|\boldsymbol{D}_2\right|$为对应集合的样本个数，$\mathrm{Gini}(\bullet,\bullet)$ 的计算公式如式（4）所示。

④ 选择最优分割点。对于特征 $\boldsymbol{a}_j$，计算完成后将得到 n 个（$\boldsymbol{a}_j$ 具有离散型特征）或 $n-1$ 个（$\boldsymbol{a}_j$ 具有连续型特征）分割基尼系数，取 $\mathrm{Gini}(\boldsymbol{D},\boldsymbol{a}_j)\mid_{s_l}$ 最小的分割点 $s_{\min}$ 为特征 $\boldsymbol{a}_j$ 的最优分割点，该点对应的分割后的样本集为 $\hat{\boldsymbol{D}}_1$、$\hat{\boldsymbol{D}}_2$。

（2）生成节点。对所有特征 $\boldsymbol{a}_j\in\boldsymbol{A}$ 计算完成后，将得到 m 个基尼系数：$\mathrm{Gini}(\boldsymbol{D},\boldsymbol{a}_j)\left(1\leqslant j\leqslant m\right)$。取基尼系数最小的特征作为新生成的节点，并将对应的子集 $\hat{\boldsymbol{D}}_1$、$\hat{\boldsymbol{D}}_2$ 分别分配到两条向边。

（3）递归建树。对样本集 $\hat{\boldsymbol{D}}_1$、$\hat{\boldsymbol{D}}_2$，删除已确认的特征 $\boldsymbol{a}_j$，更新特征集合 $\boldsymbol{A}$，待划分特征数 $m=m-1$，并返回步骤（1），递归构建每条边的分支子树。

（4）结束条件。若所有特征均已用来建树，或者递归深度达到预设的层数，则结束建树，返回创建的 CART。

对所有抽样的训练集都创建完 CART 后，这些 CART 构成了一个可进行分类预测的随机森林。设一个新的待预测发热待查样本为 $\boldsymbol{x}$，由构建的随机森林的每棵 CART 对样本 $\boldsymbol{x}$ 进行分类，最后分类数量最多的类别为 $\boldsymbol{x}$ 的预测结果。

2.2.2　双随机森林结构

对于不平衡数据集，少数类样本存在代表性数据缺乏的问题，并且在 CART 的构建过程中对少数类样本的观察会越来越少从而失去训练特征，导致模型容易将少数类误判成多数类[27]。因此，为了提升少数类的预测性能，本文构建了双随机森林预测模型，用原始样本集构造随机森林 RF_1，并用欠采样得到的平衡数据集构造随机森林 RF_2，将两个随机森林联合起来形成预测模型 BRF，由全部的 CART 投票输出预测结果。BRF 拥有更多的 CART，且兼容了原始样本的分布特性和少数类的投票权重。该预测模型结构如图 4 所示。

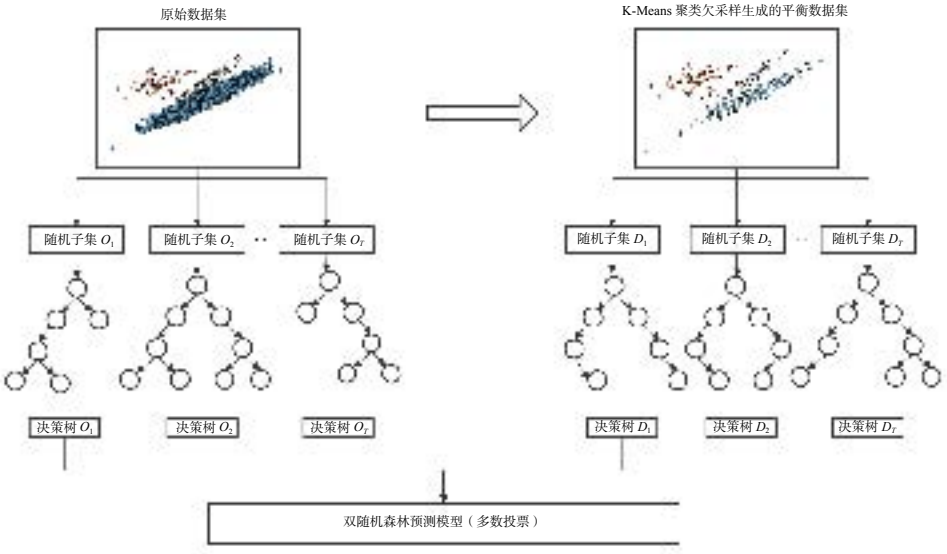


图 4　基于 K-Means 聚类欠采样的双随机森林预测模型结构
Fig.4　Structure of bi-random forest prediction model based on K-Means clustering undersampling

双随机森林预测模型包含 $2T$ 棵 CART，将测试样本 $\boldsymbol{x}$ 作为输入，每棵 CART 都将得出一个预测结果。

$$f_t(\boldsymbol{x})=\begin{cases}1, & c=1\\ 2, & c=2\\ \vdots & \\ k, & c=k\end{cases} \tag{8}$$

式中，$f_t(\cdot)$为第 $t\left(1\leqslant t\leqslant 2T\right)$棵 CART 的判别函数；$c$ 表示样本类别，$1,2,\cdots,k$为类别标签。得到所有 CART 的预测结果后，采取如下投票机制，得票最多的类别为样本的最终类别。

$$f_{\mathrm{BRF}}(\boldsymbol{x})=\underset{i=1,2,\cdots,k}{\arg\max}\left(\sum_{t=1}^{T}\{I(f_t^1(\boldsymbol{x})=i)\}+\sum_{t=1}^{T}\{I(f_t^2(\boldsymbol{x})=i)\}\right) \tag{9}$$

式中，$f_t^1(\cdot)$表示 RF_1 中第 t 棵 CART 的输出；$f_t^2(\cdot)$表示 RF_2 中第 t 棵 CART 的输出；$I(\cdot)$是一个判断函数，当 CART 的输出满足要求时$I(\cdot)=1$，不满足要求时$I(\cdot)=0$。

3　实验结果与分析

3.1　实验条件

实验平台采用 Pycharm2019、Anaconda 版本为 4.12.0、Python 版本为 3.7 的仿真环境，通过十折交叉验证的方法，对发热待查数据集的 16 类共 560 条样本数据进行模型评估。

分别从以下两个方面进行实验。

（1）性能对比。将本文方法与标准随机森林、BRAF[21]、AdaBoost[28] 等主流分类方法进行比较。

（2）消融实验。将在原始不平衡数据集上直接使用标准随机森林的分类结果、在 K-Means 聚类欠采样后的平衡数据集上使用标准随机森林的分类结果与本文方法的分类结果进行比较。

3.2　评价指标

主要采用 3 个指标评价模型的性能 [29]：准确率（Accuracy）、召回率（Recall）和 F_1，其中准确率反映整体分类正确率，其值越高，说明分类方法的分类正确率越高；召回率在二分类时代表正样本的召回率，本文在处理发热待查二分类问题时，以非感染类为正样本，感染类为负样本；F_1 是基于查准率和查全率的加权调和平均指标，反映了分类方法的整体分类情况，其值越高，说明分类方法在处理不平衡数据集时整体分类性能越好。

在二分类中，混淆矩阵如表 1 所示。

表 1　混淆矩阵
Tab.1　Confusion matrix

	预测为正类	预测为负类
实际正类	实际为正，预测为正（TP）	实际为正，预测为负（FN）
实际负类	实际为负，预测为正（FP）	实际为负，预测为负（TN）

各评价指标的计算公式如下。

$$\text{Accuarcy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{10}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$\text{F}_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{12}$$

式中，$\text{Precision}=\text{TP}/(\text{TP}+\text{FP})$。

对于多分类方法，F_1 对应的性能评价指标计算公式为：

$$\text{F}_{1-\text{macro}}=\frac{1}{k}\sum_{i=1}^{k}\text{F}_1^i \tag{13}$$

式中，k 为类别数量；F_1^i 为类别 i 的 F_1 值。

3.3　实验结果

（1）本文方法与标准随机森林、BRAF、AdaBoost 等分类方法在感染类和非感染类二分类上的性能对比如表 2 所示。最佳分类结果用星号（*）标记。

表 2　二分类性能对比

Tab.2　Comparison of the performance of the two classifications

指标	本文方法	标准随机森林	BRAF	AdaBoost
Accuracy	**0.8912 ***	0.8686	0.8839	0.8590
Recall	**0.8455 ***	0.6714	0.8210	0.7508
F_1	**0.8203***	0.7477	0.7973	0.7548

可以看出，本文方法在处理发热待查不平衡数据集的二分类问题时，Accuracy、Recall 和 F_1 这 3 项指标都达到了 80% 以上，性能优于标准随机森林、BRAF 和 AdaBoost。本文方法的 Recall 明显提高，表明其降低了非感染少数类的误判率。

（2）本文方法与标准随机森林、BRAF、AdaBoost 等方法在 16 个细分类别上的多分类性能对比如表 3 所示。表中列出了 16 个细分类别的整体 Accuracy、$\text{F}_{1\text{-macro}}$、每个分类的召回率 Recall_i 和平均召回率 Recall。其中，Recall_i 对应的标签序号 $i=1,2,\cdots,16$ 表示样本数量按照从大到小的顺序排列。最佳分类结果用星号（*）标记。

表 3　多分类性能对比

Tab.3　Comparison of the performance of the multi-classifications

指标	本文方法	标准随机森林	BRAF	AdaBoost
Accuracy	**0.7409***	0.6787	0.7217	0.7196
$\text{F}_{1\text{-macro}}$	**0.6578***	0.5062	0.6289	0.6369
Recall_1	0.8732	0.9320	0.8845	**0.8557***
Recall_2	0.5883	**0.7130***	0.5442	0.6494
Recall_3	**1.0000***	**1.0000***	**1.0000***	**1.0000***
Recall_4	0.5000	**0.7319***	0.6021	0.5745
Recall_5	0.5698	**0.5867***	0.6067	0.5333
Recall_6	**0.8895***	0.7553	0.8526	0.7632
Recall_7	**0.7941***	0.7441	0.78534	0.7059
Recall_8	**0.8515***	0.8273	0.8212	0.7879
Recall_9	**0.7958***	0.6708	0.7917	0.6667
Recall_{10}	**0.6050***	0.1300	0.4700	0.3500
Recall_{11}	0.6421	0.3158	0.5789	**0.7368***
Recall_{12}	0.32143	0.0000	0.27143	**0.7143***
Recall_{13}	**0.8077***	0.1077	0.7000	0.6923
Recall_{14}	**0.6833***	0.0833	0.5417	0.5833
Recall_{15}	0.8333	0.3250	0.8417	**0.9167***
Recall_{16}	**0.5780***	0.0091	0.1273	0.1818
Recall	**0.7083***	0.4957	0.6512	0.6695

可以看出，本文方法在 Accuracy、$F_{1\text{-macro}}$ 这两项指标上相较于标准随机森林、BRAF 和 AdaBoost 均有明显提升。在平均召回率指标上，本文方法在 9 个类别上表现最佳，特别是在后面的少数类上，召回率提升得非常明显，表明本文方法降低了少数类的误判率。虽然本文方法对于其中某些类别的召回率表现低于其他方法，但其平均召回率表现最好，达到了 0.7083。

3.4　消融实验

将 K-Means 聚类欠采样后的数据集生成随机森林预测模型（对照组 1），将原始不平衡数据集生成随机森林预测模型（对照组 2），与本文方法进行消融实验对比。

表 4 列出了本文方法与对照组的二分类性能对比结果。从表中可以看出，本文方法的 Accuracy、F_1 指标都优于对照组 1 和对照组 2。虽然本文方法的 Recall 低于对照组 1，但从综合指标 F_1 来看，对照组 1 在提高非感染少数类的预测准确性的同时，却降低了感染多数类的预测性能。

表 4　二分类性能对比

Tab.4　Comparison of the performance of the two classifications

指标	本文方法	对照组 1	对照组 2
Accuracy	**0.8912 ***	0.8398	0.8686
Recall	0.8455	**0.9225***	0.6714
F_1	**0.8203***	0.7696	0.7477

表 5 列出了本文方法与对照组的多分类性能对比。从表中可以看出，本文方法的 Accuracy、Recall、$F_{1\text{-macro}}$ 都优于对照组 1 和对照组 2，尤其是在 $F_{1\text{-macro}}$ 这一综合指标上，本文方法提高了 15% 以上。

表 5　多分类性能对比

Tab.5　Comparison of the performance of the multi-classifications

指标	本文方法	对照组 1	对照组 2
Accuracy	**0.7409***	0.4742	0.6787
Recall	**0.7083***	0.5687	0.4957
$F_{1\text{-macro}}$	**0.6578***	0.4461	0.5062

4　结论

发热待查难以确诊病因，是医学界的一大难题。利用机器学习方法对发热待查疾病进行辅助诊断是本文的主旨。由于发热待查的病因类型多，数据集严重不平衡，机器学习算法无法被有效地应用在发热待查的病因诊断上。针对这一多分类问题，本文提出了一种基于 K-Means 聚类欠采样的双随机森林模型，通过 K-Means 聚类采样构建平衡数据集，在此基础上训练一个有利于改善少数类预测性能的随机森林模型，与由初始数据集训练的随机森林联合构成双随机森林预测模型。该模型增加了 CART 的数量，并利用平衡数据集提高了少数类的投票权重。通过与其他主流方法的分类性能进行对比实验和消融实验，证明了本文方法在预测发热待查样本类别方面性能优良，不仅能有效改善少数类的预测准确性，对其他类别的预测性能也有较大提升，是一种较好的发热待查病因预测智慧诊疗方法。

参考文献

[1] 张文宏，李太生 . 发热待查诊治专家共识 [J]. **上海医学**，2018，41（7）：385-400.
ZHANG Wenhong, LI Taisheng. Treatment expert consensus of fever of unknown origin diagnosis[J]. **Shanghai Medical Journal**, 2018, 41(7):385-400.

[2] OGUNLEYE A A, WANG Q G. XGBoost model for chronic kidney disease diagnosis[J]. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 2020, 17(6): 2131-2140.

[3] ZHANG X, MARIA B, MARIA P G, *et al*. Disease classification risk through machine learning algorithms—lessons learned from COVID-19[J]. **Transplantation**, 2022, 106(9S):196.

[4] ZHANG Q, ZHAO H, HANG Y, *et al*. Research on Parkinson's disease diagnosis based on improved particle swarm optimization support vector machine algorithm[C]//2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining(MLCCIM). Xiamen:IEEE, 2022:365-369.

[5] ZHANG H, GUO L, WANG D, *et al*. Multi-source transfer learning via multi-kernel support vector machine plus for b-mode ultrasound-based computer-aided diagnosis of liver cancers[J]. **IEEE Journal of Biomedical and Health Informatics**, 2021, 25(10):3874-3885.

[6] SHI F, CHEN B, CAO Q, *et al*. Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest CT images[J]. **IEEE Transactions on Medical Imaging**, 2022, 41(4):771-781.

[7] SUN Y M, WONG A K C, KAMEL M S. Classification of imbalanced data: a review[J]. **International Journal of Pattern Recognition and Artificial Intelligence**, 2009, 23(4):687-719.

[8] 陆克中，陈超凡，蔡桓，等 . 面向概念漂移和类不平衡数据流的在线分类算法 [J]. **电子学报**，2022，50（3）：585-597.
LU Kezhong, CHEN Chaofan, CAI Huan, *et al*. Online classification algorithm for concept drift and class imbalance data stream[J]. **Acta Electronica Sinica**, 2022, 50(3):585-597.

[9] STEFANOWSKI J. Dealing with data difficulty factors while learning from imbalanced data[M]//S. Matwin, J. Mielniczuk. Studies in Computational Intelligence: vol 605. Berlin: Springer, 2016:333-363.

[10] CHAWLA N V, BOWYER K W, HALL L O, *et al*. SMOTE: synthetic minority over-sampling technique[J]. **The Journal of Artificial Intelligence Research**, 2002,16(1):321-357.

[11] CHANG C C, LI Y Z, WU H C, *et al*. Melanoma detection using XGB classifier combined with feature extraction and k-means SMOTE techniques[J]. **Diagnostics**, 2022,12(7).

[12] MAHMUD S M H, CHEN W, JAHAN H, *et al*. iDTi-CSsmoteB: identification of drug-target interaction based on drug chemical structure and protein sequence using XGBoost with over-sampling technique SMOTE[J]. **IEEE Access**, 2019(7):48699-48714.

[13] SONG L L, XU Y K, WANG M H, *et al*. PreCar_deep: a deep learning framework for prediction of protein carbonylation sites based on Borderline-SMOTE strategy[J]. **Chemometrics and Intelligent Laboratory Systems**, 2021, 218:104428.

[14] LIN C, TSAI C F, LIN W C. Towards hybrid over-and under-sampling combination methods for class imbalanced datasets: an experimental study[J]. **Artificial Intelligence Review**, 2022, 56(2): 845-863.

[15] 沈晔，李敏丹，夏顺仁 . 计算机辅助乳腺癌诊断中的非平衡学习技术 [J]. **浙江大学学报（工学版）**，2013，47（1）：1-7.
SHEN Ye, LI Mindan, XIA Shunren. Learning algorithm with non-balanced data for computer-aided diagnosis of breast cancer[J]. **Journal of Zhejiang University (Engineering Science)**, 2013, 47(1):1-7.

[16] 何云斌，冷欣，万静 . 不平衡数据加权边界点集成欠采样方法 [J]. **西安电子科技大学学报**，2021，48（4）：176-183，191.
HE Yunbin, LENG Xin, WAN Jing. Unbalanced data weighted boundary point integration undersampling method[J]. **Journal of Xidian University**, 2021, 48(4):176-183, 191.

[17] 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**, 2018, 29(8):3573-3587.

[18] 徐继伟，杨云 . 集成学习方法 : 研究综述 [J]. **云南大学学报（自然科学版）**，2018，40（6）：1082-1092.
XU Jiwei, YANG Yun. A survey of ensemble learning approaches[J]. **Journal of Yunnan University** (Natural Sciences Edition), 2018, 40(6):1082-1092.

[19] BREIMAN L. Random Forests[J]. **Machine Learning**, 2001, 45(1):5-32.

[20] 常玉清，孙雪婷，钟林生，等 . 基于改进随机森林算法的工业过程运行状态评价 [J]. **自动化学报**，2021，47（9）：2214-2225.
CHANG Yuqing, SUN Xueting, ZHONG Linsheng, *et al*. Industrial operation performance evaluation of industrial processes based on modified random forest[J]. **Acta Automatica Sinica**, 2021, 47(9):2214-2225.

[21] MOHAMMED B E D, ELEMAN T, TODD P. Biased random forest for dealing with the class imbalance problem[J]. **IEEE Transactions on Neural Networks and Learning Systems**, 2019, 30(7):2163-2172.

[22] JAIN A K. Data clustering: 50 years beyond k-means[J]. **Pattern Recognition Letters**, 2010, 31(8): 651-666.

[23] BREIMAN L, FRIEDMAN J H, OLSHEN R A, *et al*. Classification and regression trees[M]. Monterey: Wadsworth International Group,1984.

[24] POLIKAR R. Ensemble based systems in decision making[J]. **IEEE Circuits and Systems Magazine**, 2006, 6(3): 21-45.

[25] BREIMAN L. Bagging predictors[J]. **Machine Learning**, 1996, 24(2):123-140.

[26] HO T K. The random subspace method for constructing decision Forests[J]. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, 1998, 20(1):832-844.

[27] HE H, GARCIA E A. Learning from imbalanced data[J]. **IEEE Transactions on Knowledge and Data Engineering**, 2009, 21(9):1263-1284.

[28] HASTIE T, ROSSET S, ZHU J, *et al*. Multi-class AdaBoost[J]. **Statistics and Its Interface**, 2009, 2(3):349-360.

[29] SOKOLOVA M,LAPALME G. A systematic analysis of performance measures for classification tasks[J]. **Information Processing & Management**, 2009, 45(4):427-437.