

引用格式:李新鹏,高欣,何杨,等.不平衡数据集下基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型[J].微电子学与计算机,2020,37(3):14-19. [LI X P, GAO X, HE Y, et al. Prediction model of disk failure based on adaptive weighted bagging-GBDT algorithm under imbalanced dataset[J]. Microelectronics & Computer, 2020, 37(3): 14-19.]

不平衡数据集下基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型

李新鹏^{1,2}, 高 欣², 何 杨², 阎 博³, 孙汉旭²,
李军良⁴, 徐建航⁴, 刘震宇⁵, 庞 博⁵

(1 国家电网有限公司, 北京 100031; 2 北京邮电大学 自动化学院, 北京 100876;
3 国网冀北电力有限公司, 北京 100054; 4 南瑞集团(国网电力科学研究院)有限公司, 北京 100192;
5 国网冀北电力有限公司 承德供电公司, 河北 承德 067000)

摘 要: 针对磁盘数据集中正负样本数目严重不平衡导致基于机器学习的分类算法易出现故障预测准确率低的问题, 本文提出一种基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型。首先, 提出基于聚类的分层欠采样方法对健康磁盘样本进行多次抽样, 解决随机欠采样方法易丢弃潜在有用样本的问题; 其次, 将每次采样后样本与全部故障磁盘样本组合得到多个样本子集, 通过训练这些子集建立多个预测精度较高的 GBDT 子分类模型; 最后, 根据待测点邻域样本类别自适应确定各子模型权重, 据此通过加权硬投票集成最终的磁盘故障预测模型。在 8 组 KEEL 不平衡数据集上实验结果表明, 与现有典型不平衡学习算法相比, 少数类的召回率平均提升了 9.46%; 同时在磁盘公开数据集和某调度系统磁盘数据上对比验证了该方法在故障预测率上的先进性。

关键词: 磁盘故障预测; 不平衡数据集; 分层欠采样; Bagging-GBDT; 自适应加权

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2020)03-0014-06

Prediction model of disk failure based on adaptive weighted bagging-GBDT algorithm under imbalanced dataset

LI Xin-peng^{1,2}, GAO Xin², HE Yang², YAN Bo³, SUN Han-xu²,
LI Jun-liang⁴, XU Jian-hang⁴, LIU Zhen-yu⁵, PANG Bo⁵

(1 State Grid Corporation of China, Beijing 100031, China;
2 College of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China;
3 State Grid Jibei Electric Power Company Limited, Beijing 100054, China;
4 Nari Group (State Grid Electric Power Research Institute) Corporation, Beijing 100192, China;
5 State Grid Jibei Electric Power Company Limited Chengde Power Supply Company, Chengde 067000, China)

Abstract: Aiming at the problem that the classification algorithm based on machine learning is prone to low accuracy of fault prediction due to the serious imbalance between the number of positive and negative samples in the disk dataset, this paper proposes a disk fault prediction model based on adaptive weighted Bagging-GBDT algorithm. Firstly, a hierarchical under-sampling method based on clustering algorithm is proposed to sample healthy disk samples several times to solve the problem that the random undersampling method is easy to discard potentially useful samples. Secondly, each sample after sampling is combined with all the failed disk samples to obtain several

收稿日期: 2019-06-20; 修回日期: 2019-07-10

基金项目: 国家电网公司总部科技项目一调度控制系统测试验证与一体化运维关键技术研究与应用(DZ71-18-010)

subsets. By training these subsets, a number of GBDT sub-classification models with higher prediction accuracy are established. Finally, the weights of each sub-model are adaptively determined through the neighborhood sample label of the test sample, and the final disk failure prediction model is integrated by weighted hard voting. The experimental results on 8 sets of KEEL imbalanced datasets show that the recall of the negative is increased by an average of 9.46% compared with the existing typical imbalanced learning algorithm. At the same time, the advancement of the method in the fault prediction rate is verified on disk public datasets and the disk data of a scheduling system.

Key words: prediction of disk failure; imbalanced dataset; hierarchical under-sampling; Bagging-GBDT; adaptive weighted

1 引言

作为智能电网的重要组成部分,智能电网调度控制系统(D5000 系统)的可靠、稳定和高效运行在保障电网安全和国民经济社会生活方面具有重要意义^[1]. 磁盘是 D5000 系统存储数据的关键组件之一,尽管有如 RAID 等一系列的防御机制,系统的可用性和可靠性仍然经常受到严重的冲击. 因此,在构造故障预测分类器^[2]时,故障磁盘需被更多地识别出. 如何利用这些不平衡的数据^[3]获得更好的预测效果,这是需要考虑和解决的问题.

对于不平衡数据的分类^[3]现在常用技术主要分为数据层面^[4-5]和算法层面^[6-10]. 有效地组合这两种方法可提高对不平衡数据的分类效果^[10].

针对调度服务器磁盘故障预测中正负样本数目不平衡的问题,本文提出一种基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型. 通过对健康磁盘样本进行多次基于聚类的分层欠采样以减少多数类样本有用信息的损失,并将采样后的样本与全部少数类样本组合得到多个训练样本子集,最后根据自适应加权投票方法进行多个 GBDT 分类器的集成,实现对磁盘故障的准确预测. 为验证模型的有效性,本文选取 KEEL 不平衡数据集、磁盘公开数据集及某调度系统磁盘数据进行实验分析.

2 梯度提升决策树(GBDT)算法

Friedman 在 2001 年提出梯度提升算法^[11](GB),是 Boosting 方法中的代表算法之一,因其出色的特征自动组合能力和高效的运算能力广受欢迎. GB 算法中最典型的基学习器是回归树,被称为梯度提升树算法(GBDT). 文献^[12]通过对 71 个数据集进行模型训练和测试也证实该算法相比于支持向量机、k-近邻等算法无论在计算效率还是预测性能上均具有明显的优势.

对于分类任务考虑的指标是预测精度,因此以预测精度较高的 GBDT 作为子分类模型,其核心思想是利用梯度提升算法优化损失函数以得到最优的分类回归树. 现给定磁盘训练样本集 $D_{\text{tra}} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ 为磁盘的输入特征向量, p 为输入特征总数, $y_i \in \{0, 1\}$ 为磁盘状态类别, N 为训练样本总数,具体步骤如下:

(1) 输入磁盘训练样本,损失函数 $L(y, h(x))$;

(2) 初始化回归树

$$h_0(x) = \arg \min_c \sum_{i=1}^N L(y_i, c) \quad (1)$$

(3) 对迭代次数 $m = 1:M$:

① 对所有磁盘训练样本,计算负梯度

$$r_{mi} = - \left[\frac{\partial L(y_i, h(x_i))}{\partial h(x_i)} \right]_{h(x_i)=h_{m-1}(x_i)} \quad (2)$$

② 利用 (x_i, r_{mi}) 拟合第 m 棵回归树,得到该树的叶结点区域 R_{mj} , $j = 1, 2, \dots, J$;

③ 对 $j = 1, 2, \dots, J$, 计算最优拟合值

$$c_{mj} = \arg \min_c \sum_{x_i \in R_{mj}} L(y_i, h_{m-1}(x_i) + c) \quad (3)$$

④ 更新模型

$$h_m(x) = h_{m-1}(x) + \sum_{j=1}^J c_{mj} I(x \in R_{mj}) \quad (4)$$

(4) 得到最终磁盘故障预测子模型

$$h_M(x) = \sum_{m=1}^M \sum_{j=1}^J c_{mj} I(x \in R_{mj}) \quad (5)$$

3 基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型

3.1 自适应加权 Bagging-GBDT 算法

Bagging^[13]是集成学习方法著名的代表之一,它通过随机采样法从训练样本集中随机有放回地选取多个样本子集,然后使用这些样本子集训练多个子分类器,再通过 Voting 方法组合所有的子分类器得到最终的分类模型,该方法具有较强的

泛化能力. 但磁盘数据集中存在严重的不平衡现象, 若直接采用 Bagging 算法会影响子分类器对故障磁盘的预测效果, 故对 Bagging 算法中采样环节进行改进, 使用全部故障磁盘训练样本, 仅对健康样本训练数据进行多次欠采样以保留更多的负样本信息.

考虑到随机欠采样方法具有一定盲目性, 会造成部分有用多数类样本的丢失^[5], 而基于聚类的欠采样算法盲目性相对较小, 因此提出一种基于 k-means 聚类的分层欠采样方法对多数类样本进行欠采样以实现正负样本数目均衡. 在聚类过程, 每个簇包含的样本个数不一定相同, 每个簇对应采样个数随着簇内样本总数的增加而增加. 步骤如下:

(1) 从训练数据集 $D_{\text{tra}} = \{(x, y)\}$ 中挑选所有多数类训练样本 D_{traMaj} , 设定采样总数 $\text{sample} = \text{count}_{\text{min}}$, 其中, $\text{count}_{\text{min}}$ 为少数类训练样本数目;

(2) 根据欧氏距离通过 k-means 聚类算法将 D_{traMaj} 聚成 k 个簇 D_{Maji} , 其中, $i = 1:k$;

(3) 计算每个簇对应样本总数 $\text{count}_{\text{Maji}}$, 应满足 $\text{count}_{\text{maj}} = \sum_{i=1}^k \text{count}_{\text{Maji}}$, 其中, $\text{count}_{\text{maj}}$ 为 D_{traMaj} 所包含的样本数目;

(4) 对每个簇所包含样本进行分层采样, 满足采样数目等于 $\text{sample}_{\text{maji}} = \frac{\text{count}_{\text{maji}}}{\text{count}_{\text{maj}}} * \text{count}_{\text{min}}$;

(5) 组合所有簇的采样样本, 得到采样后的多数类样本 $D'_{\text{traMaj}} = \sum_{i=1}^k D'_{\text{Maji}}$, 其中, D'_{Maji} 为第 k 个簇采样后的样本集.

由于 Boosting 方法相比 Bagging 方法具有更低的偏差, 所以根据基于 Boosting 思想的 GBDT 算法训练子分类模型; 在测试阶段, 通过待测样本邻域样本类别自适应加权投票集成多个 GBDT 子模型得到最终的分类模型. 步骤如下:

(1) 多数类样本基于 k-means 聚类算法进行 $D_{\text{tra}} = \{(x, y)\}$ 次有放回地分层欠采样, 且满足每次采样总数等于全部少数类训练样本数目, 得到 T 份多数类样本子集, 具体流程如图 1 所示;

(2) 分别将这些样本子集与全部少数类训练样本组合得到 T 份训练集;

(3) 使用 GBDT 算法对这些子集进行训练, 得到多个 GBDT 子分类模型;

(4) 根据待测样本在训练集中最近的 n 个样本类别计算权值 $D_{\text{tra}} = \{(x, y)\}$, 采用加权投票法进

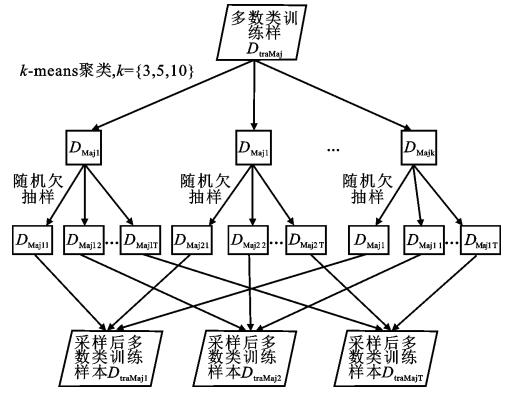


图 1 基于 k-means 的分层欠采样

行子分类模型集成得到最终的分类模型.

3.2 磁盘故障预测模型

本文提出的基于 Bagging-GBDT 算法的磁盘故障预测流程图, 如图 2 所示. 当磁盘测试样本送入故障预测模型时, 首先根据欧式距离统计其在训练样本集内 n 个邻居点中的故障类样本的数目 count , 计算故障类权重系数 $w_+ = \text{count}/n + 1$, 健康类权重系数默认设置为 1; 然后分别通过各个子预测模型对该样本进行测试得到预测类别; 最后, 根据权重系数对各种子模型的预测结果进行组合, 得到测试样本最终的预测类别, 具体细节如伪代码 1 所示.

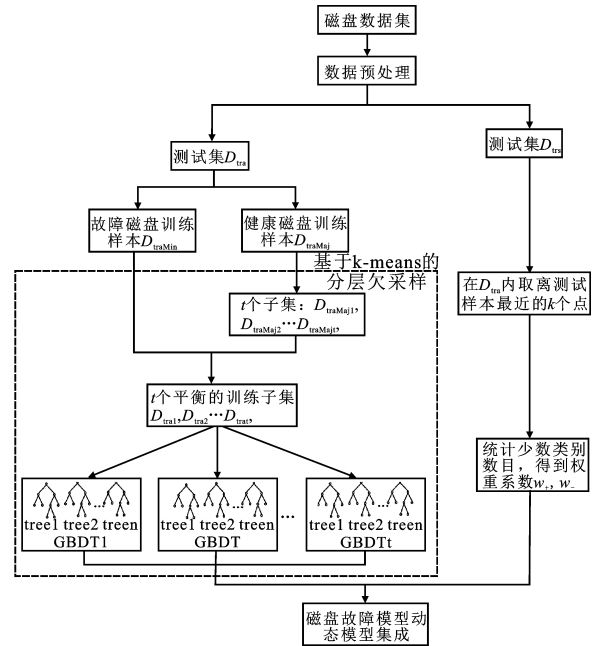


图 2 磁盘故障预测流程图

伪代码 1: 模型的构建与故障预测

输入: 训练数据集 $D_{\text{tra}} = \{(x, y)\}$, D_{traMaj} 为健康磁盘训练样本, D_{traMin} 为故障训练样本, 磁盘测试

数据集 D_{tes} ,子分类器个数 T ;

输出:磁盘测试样本预测值 y_{ic} .

Step1. 利用 k-means 聚类算法将 D_{traMaj} 聚成 k 个簇 D_{Maji} ,其中, $i = 1:k$;

Step2. 模型训练 for $t = 1:T$:

(1) 对 k 个簇 D_{Maji} 分层欠采样得到 $D_{traMajt}$, 且满足 $D_{traMajt} = D_{traMin}$.

(2) $D_{trat} = D_{traMajt} \cup D_{traMin}$, 通过 GBDT 算法训练 D_{trat} 得到子分类器 h_t :

$$h_t = h_{tM}(x) = \sum_{m=1}^M \sum_{j=1}^J c_{mj} I$$

end for

Step3. 自适应集成 for each x_{ic} in D_{tes} :

(1)for each x_i in the n nearest neighbors of the x_{ic} , 若 $y_i = 1$, 则:

count = count + 1

end for

(2) 令 $w_+ = \text{count}/n + 1, w_- = 1$;

(3) 输出 y_{ic} :

$$y_{ic} = \operatorname{argmax}_y \sum_{t=1}^T I(h_t(x_{ic}) = y) \times w$$

end for

4 实验分析

4.1 实验设置

现有的分类方法通常采用测试精度作为衡量分类器性能的指标. 对于不平衡数据集,以测试精度作为衡量标准并不合适^[3]. 在故障诊断中,故障样本数目远远少于正常样本数目,即使获得较高的精确率,但也无法准确反映对故障样本的识别率,因此测试精度不能很好地评价分类器在不平衡数据集上的分类性能. 所以实验采用少数类的 Recall₊^[8] 指标作为评价标准,Recall₊ 为少数类查全率,也是故障磁盘的预测率,定义分别如下:

$$\text{Recall}_+ = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

式中,TP 表示被正确预测为少数类的样本数目;FN 指的是被错分为多数类的少数类样本数目.

4.2 KEEL 数据集实验

为了验证本文提出方法的有效性,从机器学习权威 KEEL 数据库(<http://www.keel.es/>)中选取 8 组具有代表性的不平衡数据集进行对比实验分析. 对于具有两类以上类别的数据集,选择其中某一类别作为少数类,其他类别作为多数类,数据集的相

关信息如表 1 所示.

表 1 数据集描述

数据集名称	样本总数	特征数	少数类所属类别	不平衡率
Pima	768	8	tested_positive	1.87
Haberman	306	3	positive	2.78
Vehicle	846	18	van	3.25
New-thyroid	215	5	2	5.14
Cleveland	303	13	3	7.66
Ecoli	336	7	imU	8.19
Balance-scale	625	4	B	11.76
Yeast	1 484	8	ME2	28.10

实验采用 10 折交叉验证,总模型个数 $t = 30$, GBDT 子分类模型参数设置如下:loss='deviance'; learning_rate=0.1; n_estimators=100. 测试样本邻居点个数 $n = 5$, 聚类簇数 $k = 5$. 为充分验证本文方法的有效性,除了将其与基于 Bagging 思想的典型算法随机森林(RF)^[13]、基于 Boosting 思想的 GBDT^[11]作对比,还将其与结合 SMOTE 过采样的 GBDT-SMOTE、结合基于聚类的欠采样方法 GBDT-KM^[5]、基于 Bagging 思想处理不平衡数据分类问题的 ModifiedBagging^[8]算法进行对比,实验结果表 2 所示.

表 2 Recall₊ 值(%)对比

数据集名称	RF	GBDT	GBDT-SMOTE	Modified Bagging	Bagging-GBDT
Pima	58.08	61.36	68.64	66.95	71.92
Haberman	87.12	84.81	86.91	86.91	87.50
Vehicle	35.00	35.00	40.00	58.86	60.00
New-thyroid	61.77	59.38	73.91	76.95	79.87
Cleveland	33.19	27.36	39.58	57.92	56.94
Ecoli	93.97	91.97	96.00	97.50	99.50
Balance-scale	92.50	89.17	91.67	95.21	96.67
Yeast	49.17	55.00	69.17	86.67	89.16
平均值	63.85	63.01	70.74	78.38	80.20

从图 3 中可以看出,对大部分数据集,Bagging-GBDT 算法在 Recall₊ 指标上优于其他五种对比算法,体现了此方法的稳定性. 其中,在 Pima 数据集上提升效果最明显,与表现第二 GBDT-SMOTE 算

法相比,在 Recall₊ 上提升了 3.28%,在 New-thyroid 数据集提升效果次之,与表现第二 Modified-Bagging 算法相比,在 Recall₊ 上提升了 2.92%,这说明 Bagging-GBDT 算法能有效应用于不平衡数据的分类。

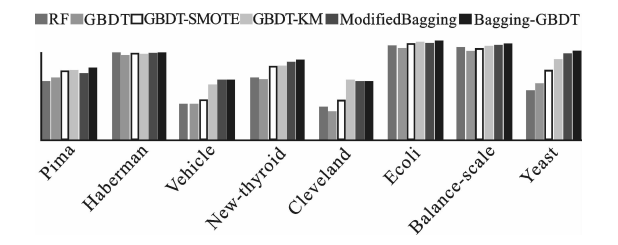


图 3 六种方法在 8 组公开数据集的 Recall₊ 对比

4.3 磁盘数据集实验

4.3.1 磁盘公开数据集实验

从 Backblaze 磁盘公开数据集(<https://www.backblaze.com/hard-drive-test-data.html>) 中选取来自日立公司某型号磁盘 2016 年上半年的数据集,该数据集中包括 4 460 块健康磁盘和 43 块故障磁盘,不平衡比例达到 103:1. 每个磁盘均以天为单位进行取样,选取该年最后一周的数据记录. 某些 smart 属性取值在健康磁盘与故障磁盘上相同,对故障预测并无意义,故只选取部分相关属性作为模型输入特征属性。

将整个数据集划分为训练集与测试集,其中早期 70%数据作为训练集,后 30%数据作为测试集. 采用本文提出的 Bagging-GBDT 算法进行训练与测试,同样与 RF、GBDT、GBDT-SMOTE、GBDT-KM、ModifiedBagging 算法进行对比,实验结果如表 3 所示。

表 3 磁盘公开数据集故障预测率对比

对比方法	故障预测率/%
RF	64.21
GBDT	66.32
GBDT-SMOTE	70.52
GBDT-KM	87.36
ModifiedBagging	85.26
Bagging-GBDT	88.42

从表 3 实验结果可以看出,相比于未结合采样方法的 RF、GBDT 两种分类算法, Bagging-GBDT 对故障样本的预测率分别提升 24.21%与 22.1%,表明进行样本数目平衡处理对解决不均衡数据分类问题的必要性;与结合过采样方法的 GBDT-SMOTE 算法相比, Bagging-GBDT 对故障样本的

预测率提升了 17.9%,这是由于过采样方法在合成大量故障样本的同时可能会带来噪声信息影响分类效果;GBDT-KM、ModifiedBagging 均是结合欠采样的典型不平衡学习方法,其中, Bagging-GBDT 算法相比于 ModifiedBagging 算法在故障预测值上提高了 3.16%,这是由于 Bagging-GBDT 算法并没有直接采用随机欠采样对健康磁盘样本进行删除,而是通过基于 k-means 聚类算法的分层欠采样删除部分健康磁盘样本,这样减少了健康磁盘样本有用的信息缺失,另外 Bagging-GBDT 算法相比于 GBDT-KM 算法也具有优势,也说明在所提方法迭代过程中增加模型多样性的同时,也一定程度上减少了多数类样本的信息缺失。

4.3.2 某调度系统磁盘数据实验

在国家电网公司某网省级 D5000 系统磁盘数据集上进行验证,选取了某型号磁盘 2017 年的记录数据作为原始数据集,共包括 200 个健康磁盘,2 个故障磁盘. 其中早期 70%数据作为训练集,后 30%数据作为测试集,用来检验模型的性能,这里同样只使用部分相关的 smart 属性作为输入特征。

将本文提出的 Bagging-GBDT 算法与 RF、GBDT、GBDT-SMOTE、GBDT-KM、ModifiedBagging 算法进行对比,实验结果如表 4 所示。

表 4 实采数据故障预测率对比

对比方法	故障预测率/%
RF	85.71
GBDT	87.25
GBDT-SMOTE	92.58
GBDT-KM	95.78
ModifiedBagging	98.86
Bagging-GBDT	99.62

从表 4 实验结果可以看出, Bagging-GBDT 算法相比于其他五种对比算法分类效果最好,在故障预测率上达到了 99.62%,这说明该算法能有效地预测出磁盘的故障。

5 结束语

针对电网调度系统磁盘故障预测中正负样本不平衡问题,本文提出了一种基于自适应加权 Bagging-GBDT 算法. 为了验证该方法的有效性,以少数类的识别率作为评价指标,在 8 组典型的 KEEL 不平衡数据集、公开磁盘数据集及某调度系统磁盘数据集进行实验对比,实验结果表明 Bagging-GB-

DT 算法相比典型不平衡分类算法能够有效地提升对不平衡数据中少数类样本的识别效果。

在 Bagging-GBDT 算法的训练模型过程中,为了获得最优的预测效果,需要设置合适的子分类模型个数,本文主要是通过多次实验得到最优个数,如何根据实验数据自适应地确定子分类模型个数将是后续研究重点。

参考文献:

[1] 杨清波,李立新,李宇佳,等. 智能电网调度控制系统试验验证技术[J]. 电力系统自动化, 2015, 39(1): 194-199.

[2] LI J, REBECCA J STONES, WANG G, et al. Hard drive failure prediction using decision trees[J]. Reliability Engineering & System Safety, 2017(164):55-65.

[3] 李金孟,林亚平,祝团飞. 基于 Hubness 与类加权的 k 最近邻分类算法[J]. 计算机工程, 2018, 44(4): 248-252.

[4] CHAWLA N V, BOWYER K W, HALL L C, et al. SMOTE: synthetic minority oversampling technique[J]. Journal of Artificial Intelligence Research, 2002(16):321 - 357.

[5] LIN W C, CHIH-FONG TSAI, YA-HAN HU, et al. Clustering-based undersampling in class-imbalanced data [J]. Information Sciences, 2017;409-410.

[6] 周宇航,周志华. 代价敏感大间隔分布学习机[J]. 计算机研究与发展, 2016, 53(9):1964-1970.

[7] SUN Z B, SONG Q B, ZHU X Y, et al. A novel ensemble method for classifying imbalanced data [J]. Pattern Recognition, 2015, 48(5):1623-1637.

[8] SAMAD NEJATIAN, HAMID PARVIN, ESHAGH FARAJI. Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification[J]. Neurocomputing, 2018(276):55-66.

[9] MIKEL GALAR, ALBERTO FERNÁDEZ, EDURNE BARRENECHEA, et al. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling[J]. Pattern Recognition, 2013, 46(12):3460-3471.

[10] TANG B, HE H B. GIR-based ensemble sampling approaches for imbalanced learning[J]. Pattern Recognition, 2017(71):306-319.

[11] JEROME H FRIEDMAN. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.

[12] ZHANG C S, LIU C C, ZHANG X L, et al. An up-to-date comparison of state-of-the-art classification algorithms[J]. Expert Systems with Applications, 2017(82):128-150.

[13] 李慧,张南南,曹卓,等. 基于机器学习的恐怖分子预测方法研究[J/OL]. 计算机工程:1-7. [2019-05-29].

作者简介:

李新鹏 男, (1989-), 博士研究生, 高级工程师. 研究方向为机器学习、电力系统自动化.

高欣(通讯作者) 男, (1974-), 博士, 副教授. 研究方向为机器学习、数据挖掘; Email: xlhhh74@bupt.edu.cn.

何杨 女, (1996-), 硕士研究生. 研究方向为机器学习、数据挖掘.

阎博 男, (1985-), 博士, 高级工程师. 研究方向为机器学习、电力系统自动化.

孙汉旭 男, (1960-), 博士, 教授. 研究方向为智能系统的分析与设计研究.

李军良 男, (1981-), 硕士研究生, 高级工程师. 研究方向为电力系统自动化、软件工程研究.

徐建航 男, (1988-). 研究方向为电力系统自动化.

刘震宇 男, (1976-). 研究方向为电力系统自动化.

庞博 男, (1975-). 研究方向为电力系统自动化.