基于特征约简的随机森林改进算法研究

王 诚,高 蕊
(南京邮电大学 通信与信息工程学院,江苏 南京 210003)

摘 要:随机森林(random forest,RF)算法虽应用广泛且分类准确度很高,但在面对特征维度高且不平衡的数据时,算法分类性能被严重削弱。高维数据通常包含大量的无关和冗余的特征,针对这个问题,结合权重排序和递归特征筛选的思想提出了一种改进的随机森林算法 RW_RF(ReliefF & wrapper random forest)。首先引用 ReliefF 算法对数据集的所有特征按正负类分类能力赋予不同的权值,再递归地删除冗余的低权值特征,得到分类性能最佳的特征子集来构造随机森林;同时改进 ReliefF 的抽样方式,以减轻不平衡数据对分类模型的影响。实验结果显示,在特征数目很多的数据集中,改进算法的各评价指标均高于原算法,证明提出的 RW_RF 算法有效精简了特征子集,减轻了冗余特征对模型分类精度的影响,同时也证明了改进算法对处理不平衡数据起到了一定的效果。

关键词:随机森林;权重排序;特征约简;抽样方式;RW_RF 算法
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2020)03-0040-06
doi:10.3969/j.issn.1673-629X.2020.03.008

An Improved Random Forest Algorithm Based on Feature Reduction

WANG Cheng,GAO Rui
(School of Telecommunications & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:Although the Random Forest (RF) algorithm is widely used and highly accurate in the classification,its performance is severely weakened when faced with high and unbalanced features. High-dimensional data usually contains a large number of irrelevant and redundant features,so we propose an improved random forest algorithm RW_RF (ReliefF & wrapper random forest) based on the idea of weight sorting and recursive feature screening. Firstly different weights are assigned by ReliefF algorithm to all features according to the positive and negative classification ability,and then the redundant low-weight features are deleted recursively to obtain the feature subset with the best classification performance for the random forest construction. At the same time,the ReliefF sampling method is improved to mitigate the impact of unbalanced data on the classification model. The experiment shows that the evaluation indexes are improved as a whole,which proves that the proposed RW_RF algorithm effectively reduces the feature subset and the influence of redundant features on the classification accuracy of the model. It also proves that the improved algorithm is effective on processing unbalanced data.

Key words:random forest;weight sorting;feature reduction;sampling method;RW_RF algorithm

0 引 言

随着计算机网络的飞速发展,电子数据库的规模呈爆炸式增长,为帮助计算机更好地处理数据,得出可行的方法论,各分类回归算法不断焕发出新的生机。其中,随机森林算法是以决策树为基础的分类回归模型,它将多个单分类器集成,共同参与决策,因此分类精度要高于一般的单分类器。算法应用领域涵盖信用贷款[1]、生物医学[2]、图像[3]、销售[4]等。虽然其在大部分场景中能达到很好的效果,但在处理某些特殊数据,如不平衡且特征维度高的医疗数据时,过多的冗余特征使得模型极易过拟合;且模型为了提升整体分类精度,习惯将少数类归为多数类处理,得到虚假的分类精度。因此,算法不得不做出针对性的改进。

多年来学者们对原算法进行了很多改进,如通过聚类方式[5]、贪婪方法[6]挑选出一批具有代表性的高精度低相似性决策树,提高了部分数据集的分类精度,

收稿日期:2019-04-21 **修回日期**:2019-08-22 **网络出版时间**:2019-12-05
基金项目:中国博士后科学基金(SBH18028)
作者简介:王 诚(1970-),男,副教授,硕导,研究方向为互联网大数据挖掘及并行计算;高 蕊(1994-),女,硕士研究生,研究方向为互联网大数据挖掘及并行计算。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191205.1133.044.html

但对于上文提到的特殊数据集效果甚微;其他改进如针对不平衡数据:改变性能评价标准[7]、重采样数据[8-9]、生成合成数据[10-11]等;又如针对特征选择的:粗糙集[12]、邻域互信息[13]、聚类[14]等,这些改进有一定成效,但很难融合以同时解决上述两种问题。其中 Marwa Hammami[15] 提出了 Filter 与 Wrapper 结合的高维数据特征构造的多目标混合滤波器包装进化算法,在消除冗余特征方面效果显著;李硕[16] 提出的基于改进的 ReliefF 算法结合支持向量机的非均衡特征选择方法有效解决了不平衡数据的问题。这两种改进一个针对特征排序,另一个针对特征约简,能很好互补。受此启发,文中提出一种基于特征约简的随机森林改进算法[17]:RW_RF。在随机森林的决策树构建过程中引入 Wrapper 递归特征消除与 ReliefF 算法结合的特征选择方法,尽可能挑选出拥有最佳分类性能的特征集,来减轻特征冗余和数据不平衡问题对模型的影响。

1　随机森林算法

随机森林算法(random forest,RF),本质是由多棵相互之间并无关联的决策树整合而成的多分类器,单条数据经过每一棵决策树投票,得票数最多的类别即为最终分类结果。

假设原始样本集 $D(X,Y)$,样本个数为 n ,要建立 k 棵树,随机森林的具体步骤大致如定义 1 所示。

定义 1:随机森林。

(1)抽取样本集:从原始训练集中随机有放回地抽取 n 个样本(子训练集)并重复 n 次,每一个样本被抽中的概率均为 $1/n$ 。被剩下的样本组成袋外数据集(OOB),作为最终的测试集。

(2)抽取特征:从总数为 M 的特征集合中随意抽取 m 个组成特征子集,其中 $m < M$ 。

(3)特征选择:计算节点数据集中每个特征对该数据集的基尼指数,选择基尼指数最小的特征及其对应的切分点作为最优特征与最优切分点,从节点生成两个子节点,将剩余训练数据分配到两个子节点中。

(4)生成 CART 决策树:在每个子节点的样本子集中重复执行步骤(3),递归地进行节点分割,直到生成所有叶节点。

(5)随机森林:重复执行步骤(2)~(4),得到 k 棵不同的决策树。

(6)测试数据:每一棵决策树都对测试集中的每一条数据进行分类,统计 k 个分类结果,票数最多的类别,即为该样本的最终类别。

2　算法改进

随机森林算法在处理不平衡且特征数非常多的数据时有几点弊端:第一,算法的分类思想是少数服从多数,因此在面对类别样本数相差悬殊的数据集时,容易将少数类归为多数类,造成很高的假分类精度;第二,过多的冗余特征会扰乱模型的学习能力,导致模型过拟合,限制了模型的普适性。因此,找出冗余度最小,且最能代表正负类数据之间的差异的特征子集,再生成随机森林,是文中算法改进的思路。基于此,提出了改进的随机森林算法 RW_RF。首先在构建 CART 决策树的特征选择步骤中,使用改进的 ReliefF 算法初步筛选掉一批不相关特征,并将留下的特征根据权重排序;接着运用 Wrapper 的递归特征选择思想,依次删除低相关特征和冗余特征,得到最佳分类特征子集;最后在随机森林中构建整个分类模型。

2.1　改进的 ReliefF 算法

ReliefF,是由 Relief 算法发展而来的一个经典的特征权重赋值算法,它将特征与正负类之间的相关性作为依据,给每个特征赋予相应的权重。

ReliefF 算法的思路为:首先从测试集中任意抽取一个样本 R_n ,接着随机抽取数量相同的 k 个 R_n 的同类与不同类样本(Same spe/Diff spe),分别计算特征 A 在 Same spe 和 Diff spe 样本间的距离,如果两类距离的均值相差悬殊,说明该特征对此类样本有较大的区分能力,继而增加该特征的权重;反之,若距离相同,说明没有区分能力,则降低该特征的权重。重复 m 次后得到的均值,作为该特征的权重。权重计算如下:

$$W(A) = W(A) - \frac{\sum_{j=1}^{k} \mathrm{diff}(A,R,H_J)}{mk} + \sum_{C \notin \mathrm{class}(R)} \frac{\left[\frac{p(C)}{1-p(\mathrm{class(R)})} \sum_{j=1}^{k} \mathrm{diff}(A,R,M_j(C))\right]}{mk} \tag{1}$$

其中, $W(A)$ 为特征 A 的权重,$p(C)$ 为原数据集中类别为 C 的样本所占比例, $M_j(C)$ 为类 $C \notin \mathrm{class}(R)$ 中第 j 个最近邻样本。$\mathrm{diff}(A,R_1,R_2)$ 表示样本 R_1 和 R_2 在特征 A 上的差,如下:

$$\mathrm{diff}(A,R_1,R_2) = \begin{cases} \dfrac{|R_1[A] - R_2[A]|}{\max(A) - \min(A)}, \text{If A is continues} \\ 0, \text{If A is continues and } R_1[A] = R_2[A] \\ 1, \text{If A is continues and } R_1[A] \neq R_2[A] \end{cases} \tag{2}$$

由式(1)知,ReliefF 将 x_i 与异类 C 中距离 x_i 最近的 k 个样本在特征 A 上的差异取平均,再乘以 C 类样本占所有与 x_i 异类样本的比例,对所有与 x_i 异类的样本执行此操作,得到特征 A 在异类样本间的差异均值。$\boldsymbol{W} = \{w_1, w_2, \cdots, w_n\}$ 是最终得到的特征权重向量,按权重从大到小对特征进行排序。

考虑到数据不平衡的问题,对以上的 ReliefF 算法稍作改进。为弥补非平衡数据对分类性能的影响,通过修改抽样参数使它相对类均衡。具体做法是,计算权值时,将原本需要设定的 k 值固定为当前样本集中少数类个数,保证计算权重时当前特征对应的正负样本数量均衡,理论上避免了分类结果偏向多数类的情况。具体步骤如定义 2 所示。

定义 2:改进的 ReliefF 特征排序法。

输入:训练集 D,抽样次数 m,k = D 中少数类样本个数;

输出:各个特征的特征权重 W。

1. 置 0 所有特征权重 $W=\{0,0,\cdots,0\}$,T 为空集;

2. for i =1 to m do;

3. 从 D 中随机选择一个样本 R;

4. 从 R 的同类样本集中找到 R 的 k 个最近邻 H_j (j =1,2,…,k),从每一个不同类样本集中找到 k 个最近邻 $M_j(C)$;

5. for A =1 to N (all features) do;

6. 将所有特征值归一化映射到[0,1]范围内;

7. $$W(A)=W(A)-\frac{\sum_{j=1}^{k}\mathrm{diff}(A,R,H_j)}{mk}+\frac{\sum_{C\notin \mathrm{class}(R)}\left[\frac{p(C)}{1-p(\mathrm{class(R)})}\sum_{j=1}^{k}\mathrm{diff}(A,R,M_j(C))\right]}{mk};$$

8. 删除权值<0 的特征。

此改进目的是,先删除对分类效果有害的特征,再将剩余特征相对正类和负类的区分能力排序,以便接下来更方便地去除冗余和不相关的特征。

2.2 递归特征消除法

使用改进 ReliefF 算法快速筛选出分类性能最佳的特征,但没有达到消除冗余特征的要求,还需要进一步优化。文中借助 Wrapper 的递归特征选择思想来剔除冗余特征,找到最佳特征子集。具体方法为,将特征按计算好的权重排序,每次从特征集合中去掉 L 个权值最小的特征生成 CART 分类决策树,并计算其 AUC 值(具体见 3.1 节)。逐次迭代,直到找到 AUC 值最高的一组特征子集。这个过程采用 k 折交叉验证法来分割数据集,计算每次迭代的 AUC 值,选择值最大的一次迭代作为删除冗余特征的依据,具体过程如定义 3 所示。

定义 3:递归特征消除法。

输入:对应特征的权值 $W=\{w_1,w_2,\cdots,w_n\}$,数据集 D;

输出:最佳分类特征子集 FGSort。

1. 初始化:读入原始数据集 D, 设置 FGSort=Null;

2. 采用分层采样技术将数据集 D 划分为 6 等份,表示为:$D=D_1\cup D_2\cup\cdots\cup D_6$;

3. 设置 6 次迭代中每次训练得到的分类器的分类准确率向量 TLAuc[1∶6]=0;

4. for(i 从 1 到[N/L]) // i 代表循环变量, N 代表数据集中所有特征个数, L 为每次删除的特征数量;

5. 在数据集(D_1-D_5)上训练决策树分类器,对应特征子集记为 FGSort_i;

6. 计算当前迭代的准确率 TLAuc_i;

7. 剔除权重最低的 L 个特征;

8. end for;

9. 输出 6 次分类准确率最高的特征子集 FGSort_i。

此改进目的是将排好序的特征按末尾淘汰制训练决策树,选出分类性能最佳的子集。

2.3 改进特征选择法与随机森林算法结合的 RW_RF 算法

RW_RF 算法相比于原始随机森林算法有两点改进:第一,随机森林随机选择特征的步骤替换为上述改进的 ReliefF 算法,在初步排除一批不相关特征的同时,对剩下特征的分类能力进行排序;第二,建立决策树时,采用递归特征选择思想依次删除低权值特征,得到分类性能最好的特征子集,最后构建随机森林分类模型。改进算法部分流程如图 1 所示。

3 实验及结果分析

3.1 评价指标

传统二分类数据的评价准则有几个重要指标,其中 TP 表示正确预测的正类,FN 表示错误预测的正类,FP 表示错误预测的负类,TN 表示正确预测的负类。样本总数 N=TP+FP+FN+TN。

(1)分类精度(Accuracy)。

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}=\frac{\mathrm{TP}+\mathrm{TN}}{N}$$

(2)灵敏度/召回率/查全率(Sensitivity)。

$$\mathrm{Sensitivity/Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

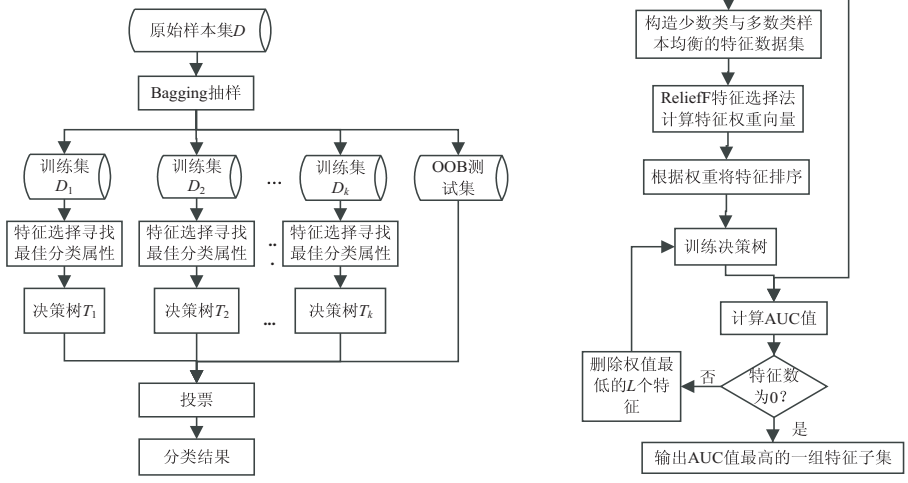
(3)特异度(Specificity)。

$$\mathrm{Specificity}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

(4)ROC 曲线/AUC。

ROC 曲线的横坐标为负样本错分的概率($\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$),纵坐标为正样本分对的概率($\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$),曲线从原点发出到终点(1,1)结束,并且不会随着正负类数量的变化而变化,因此很适合用来评价不平衡数据的分类结果。然而 ROC 曲线很难直

观地评价模型好坏,因此将曲线下方的面积(AUC)作为量化指标,面积越大表示模型分类能力越强。



(a)随机森林流程　　(b)特征选择流程

图 1　改进的 RW_RF 算法

3.2　实验数据集

实验分别选择美国加州大学 UCI 公开数据集中共 5 个用于分类问题的数据集。其中包含特征数相对较少且类平衡的数据,如糖尿病引起的视网膜病变数据集、垃圾邮件区分数据集;还包括特征数相对较多且类不平衡的数据集,如癫痫诊断数据集、麝香判定数据集;最后是斯堪尼亚卡车故障数据集,此数据集极不平衡,正负类比例接近 40∶1。这 5 个具有代表性的数据集,可以全面地展现改进的 RW_RF 算法在特征选择和处理不平衡数据方面的优势。具体参数如表 1 所示。

表 1　选用的 UCI 数据集具体参数

数据库	样本数	特征数	多数类/%	少数类/%
糖尿病视网膜病变(DB)	1 151	20	58.3	41.7
垃圾邮件区分(SB)	4 601	57	60.6	39.4
癫痫(ES)	11 500	179	82.5	17.5
麝香区分(MUSK)	6 598	168	84.6	15.4
斯堪尼亚卡车(APS)	16 000	171	97.6	2.4

3.3　实验过程

实验所用的 RW_RF 算法采用 Java 编程实现,主要用到 Weka 包来封装。硬件执行环境配置为:Intel(R) Core(TM) i7-7700HQ CPU @2.80 GHz 处理器、16 GB 内存、64 位 Windows 10 企业版操作系统。随机森林决策树个数设置为 50,取样次数 m 设置为当前数据集少数类个数 k;构建 CART 决策树时基尼指数设为 0.01。

此外,文中通过 3 种算法的对比来验证提出的 RW_RF 算法的分类效果。第一种算法是未经任何改进的原始随机森林算法;第二种是在原始随机森林算法中加入上文所提的改进 ReliefF 算法,命名为 R_RF 算法;第三种即提出的 RW_RF 算法。

分别将 5 个数据集在上述 3 种算法中进行分类,比较各自的分类精度(Accuracy)、灵敏度(Sensitivity)、特异度(Specificity)和 AUC(area under the curve)指标以及相关的参数。

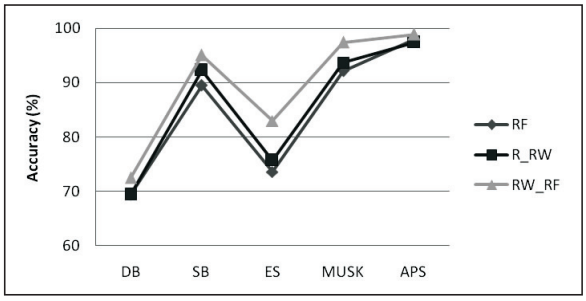
3.4　实验结果分析

实验结果如表 2 所示。

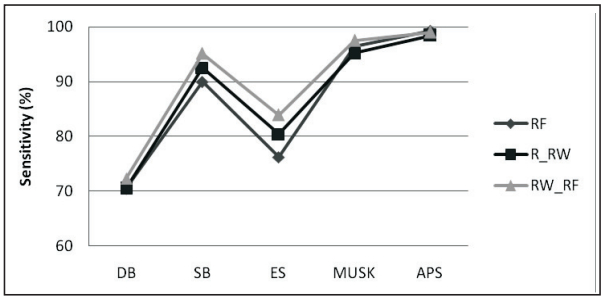
表 2　各数据集在 3 种算法中的性能指标对比

算法	Accuracy/%			Sensitivity/%			Specificity/%			AUC/%		
	RF	R_RW	RW_RF	RF	R_RW	RW_RF	RF	R_RW	RW_RF	RF	R_RW	RW_RF
DB	69.4	69.4	72.5	70.5	70.5	72.3	72.9	72.9	73.1	89.3	89.3	90.4
SB	89.5	92.3	95.1	90.0	92.5	95.1	88.6	91.8	94.3	96.6	97.0	98.7
ES	73.6	75.7	83.0	76.2	80.4	83.9	61.7	66.6	78.9	91.4	92.6	93.3
MUSK	92.1	93.7	97.4	96.4	95.2	97.5	69.1	85.7	87.6	97.3	99.0	99.4
APS	97.9	97.4	98.9	99.2	98.4	99.0	43.2	60.0	69.3	95.2	96.7	98.5

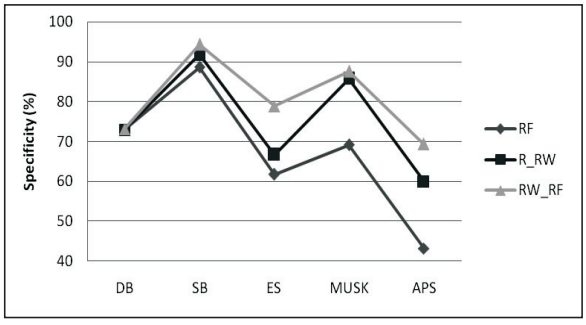
将 5 个数据集在 RF、R_RF 和 RW_RF 算法中的各性能指标结果绘制成折线图,如图 2 所示。



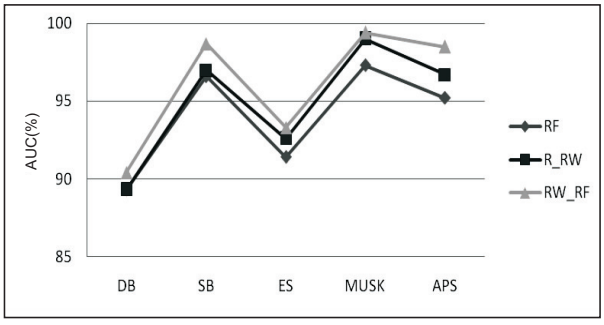
(a)准确率



(b)召回率



(c)特异度



(d)AUC

图 2 改进前后各指标对比结果

图 2 展示了 5 个数据集在原始 RF、改进 R_RF 和最终 RW_RF 算法下分类精度、灵敏度、特异度的对比结果。其中 DB 和 SB 数据集本身的特征数和样本数较少,由结果可知改进的 R_RF 算法模型没有带来太大的性能提升。然而,在特征数较多的数据集 ES、MUSK 和 APS 中,两种改进算法均达到了很好的分类效果。结合表 3 可知,在改进的 R_RF 模型中训练后,3 个数据集的特征数由原来的 179、168、171,分别约简到 165、143、139,初步删除了大量无关特征值后,4 个分类指标结果都有大幅提升,参见图 2(d),在数据集 ES、MUSK 中整体分类性能提升最为明显,说明加入改进的 ReliefF 算法有效删除了不相关的特征,提高了模型分类性能。在 RW_RF 算法中,特征被进一步约简至 138、127、114,各指标结果相较 R_RF 又有或多或少的提升,说明 Wrapper 递归特征消除法能在 ReliefF 的基础上进一步约简冗余特征,尽可能得到对分类最有帮助的特征集合。

表 3 改进前后特征数对比

数据集/算法	RF 特征数	R_RF 特征数	RW_RF 特征数
DB	20	20	16
SB	57	52	46
ES	179	165	138
MUSK	168	143	127
APS	171	139	114

在处理数据不平衡问题中,改进的算法也体现了优越性。APS 数据集不平衡问题最严重,由图 2(a)可知,在原始 RF 中 Accuracy 分类精度非常高,但是由图 2(b)、(c)可知,其正类分类准确率接近 100%,负类分类准确率都不足 50%。但是经过改进算法模型训练后,负类分类正确率有明显提升,在 R_RF 中特异度达到了 60%,在 RW_RF 中更是达到了 69%,说明所提出的 ReliefF 抽样改进方式确实能减轻随机森林算法在处理不平衡数据集中的短板。参见图 2(d),RW_RF 的折线均在 R_RF 和 RF 之上,说明提出的 RW_RF 算法具有最佳的分类性能。

综上所述,RW_RF 算法不论在消除冗余特征还是减轻不平衡数据对模型的影响方面,都带来了有效的提升。相比于初始随机森林算法,RW_RF 算法更适用于解决特征维度高且不平衡的数据分类问题。

4 结束语

围绕多特征及不平衡数据的特殊性对随机森林算法做出了一些改进。将 ReliefF 算法和 Wrapper 递归特征选择法融合来代替随机森林算法中的特征选择过程,得到 RW_RF 算法,并选择 5 组有代表性的 UCI 数据集进行分类测试。结果表示,RW_RF 算法有更好的分类性能,证明了该改进算法对解决数据的特征冗余和数据不平衡问题有积极意义。

由于使用了递归构造决策树的方法,使得算法时间复杂度大大增加。为了进一步优化模型性能,接下来考虑实现算法并行化,如将模型在 Spark 并行计算框架中运行,以此来提高整体运算效率。

参考文献:

[1] ZHAO Y,MA X. Study on credit evaluation of electricity users based on random forest[C]//2017 Chinese automation congress (CAC). Jinan:IEEE,2017:4729-4732.

[2] ZHAO B,CAO Z,WANG S. Lung vessel segmentation based on random forests[J]. Electronics Letters,2017,53(4):220-222.

[3] ZHANG Y,CAO G,LI X,et al. Cascaded random forest for hyperspectral image classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2018,11(4):1082-1094.

[4] TORIZUKA K,OI H,SAITOH F,et al. Benefit segmentation of online customer reviews using random forest[C]//2018 IEEE international conference on industrial engineering and engineering management (IEEM). Bangkok: IEEE, 2018: 487-491.

[5] 王日升,谢红薇,安建成. 基于分类精度和相关性的随机森林算法改进[J]. 科学技术与工程,2017,17(20):67-72.

[6] MASHAHEKHI M, GRAS R. Rule extraction from decision trees ensembles: new algorithms based on heuristic search and sparse group lasso methods[J]. International Journal of Information Technology and Decision Making,2016,16(6):1707-1727.

[7] 尹　华,胡玉平. 一种代价敏感随机森林算法[J]. 武汉大学学报:工学版,2014,47(5):707-711.

[8] GARCIA S, HERRERA F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy[J]. Evolutionary Computation,2014,17(3):275-306.

[9] OH S,LEE M S,ZHANG B. Ensemble learning with autive example selection for imbalanced biomedical data classification [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2011,8(2):316-325.

[10] HAN H,WANG W Y,MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// International conference on intelligent computing. [s. l.]:[s. n.],2005:878-887.

[11] MALDONADO S,LOPEZ J. Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification[J]. Applied Soft Computing,2018,67:94-105.

[12] 吴辰文,王　伟,李长生,等. 一种结合随机森林和邻域粗糙集的特征选择方法[J]. 小型微型计算机系统,2017,38(6):1358-1362.

[13] LIU J H,LIN Y J,LIN M L,et al. Feature selection based on quality of information[J]. Neurocomputing, 2017, 225: 11-22.

[14] SHAH F P,PATEL V. A review on feature selection and feature extraction for text classification[C]//2016 international conference on wireless communications, signal processing and networking (WiSPNET). Chennai:IEEE,2016:2264-2268.

[15] HAMMAMI M,BECHIKH S,HUNG C. A multi-objective hybrid filter-wrapper evolutionary approach for feature construction on high-dimensional data[C]//2018 IEEE congress on evolutionary computation (CEC). Rio de Janeiro: IEEE, 2018:1-8.

[16] 李　硕. 非均衡医学数据的特征选择与分类[D]. 杭州:浙江大学,2018.

[17] 刘　凯,郑山红,蒋　权,等. 基于随机森林的自适应特征选择算法[J]. 计算机技术与发展,2018, 28(9):101-104, 111.