改进的并行随机森林算法及其包外估计*

钱雪忠，秦 静，宋 威
（江南大学 物联网技术应用教育部工程研究中心，江苏 无锡 214122）

摘 要：传统的包外估计记录全局数据与树之间的对应关系来测算泛化误差。然而基于 MapReduce 机制的并行随机森林算法(MR_RF)是建立在多个互不可见的分块数据上。对此分析 MR_RF 与 RF 的区别，设计了一个新的适用于 MR_RF 的包外泛化误差估计方法。主要将测算限定在数据块内，最终森林的泛化误差估计取块结果的平均。实验结果表明，新的包外估计方法与交叉验证在默认分块上的结果近似，却随着分块的增加出现偏差，对此分析了可能的原因，并给出选择集成方案思想，且分块大小与分类准确率成反比，与分类速率成正比。

关键词：MapReduce；随机森林；包外估计；泛化误差；交叉验证
中图分类号：TP301 **文献标志码**：A **文章编号**：1001-3695(2018)06-1651-04
doi:10.3969/j.issn.1001-3695.2018.06.011

Improved parallel random forest and its out_of_bag estimator

Qian Xuezhong, Qin Jing, Song Wei
(*Engineering Research Center of Internet of Things Technology Applications for Ministry of Education, Jiangnan University, Wuxi Jiangsu* 214122, *China*)

Abstract: Traditional out_of_bag (OOB) estimator needs to record the relationship of global data and the trees so as to estimate generation error. However parallel random forest based on MapReduce algorithm (MR_RF) is built on blocks that independently with each other. This paper analyzed the difference between the MR_RF and random forest, and designed a new oob estimator that was applicable to estimate MR_RF's generalization error. Its key idea was putting the OOB calculator just into that particular block and using the average result of all blocks as the final OOB estimator result. Experiments show that in the case of the default partition, the new method is as effective as cross validation. However it shows deviation as the blocks increase. This paper analyzed the reason and gave the idea of selective ensemble scheme. Meanwhile, the block size is proportional to the classification rate but inversely proportional to the classification accuracy. When dealing with large data classification problems, it is necessary to adjust the block size to take the compromise between accuracy and rate.

Key words: MapReduce; random forest(RF); out_of_bag estimator; generalization error; cross validation

0 引言

近年来，随着计算机和互联网技术的快速发展，网络上的数据膨胀速度异常迅猛。数据的不断产生和存储的需要使得大数据应用越来越成为人们关注的焦点。如何分析和处理如此庞大的数据给传统的统计科学和数据挖掘领域带来了巨大的挑战[1]。随机森林(RF)算法[2]是基于统计学习理论并应用于数据挖掘领域的算法之一。作为 bagging 集成的扩展变体，从数据扰动和属性扰动两方面增加基分类器的强度和分类器之间的多样性，正是由于和而不同的集成思想使得最终分类器的泛化性能有很大提升，能有效处理回归和分类问题，在网络安全入侵检测[3]、文本分类[4]、人脸识别[5]等领域都被广泛应用。包外估计是 bagging 集成的自然产物[6]。由于基分类器是构建在训练样本 T 的自助抽样集 Tb 上的，只有约 63.2% 原样本集出现在 Tb 中，而剩余的 36.8% 的数据 $T\backslash Tb$ 作为包外数据(OOB)，可以用于基分类器的验证集。经验证，包外估计是对集成分类器泛化误差的无偏估计[7~9]。在随机森林算法中数据集属性的重要性、分类器集强度和分类器间相关性计算都依赖于袋外数据，其重要性可见一般[10]。

处理大数据的技术重点是实现快速、可伸缩的并行化分析处理。现在主流的方法是遵循 MapReduce 流程，借助 Hadoop 分布式平台，先将大数据分成多个小的子集，然后对子集处理的结果进行合并[11]。然而传统的袋外估计需要记录数据与该数据构建树之间的对应关系，这是面向全局数据的，所以在面对需要将数据分块操作的 MapReduce 模式下，显然不适用。业界对基于 MapReduce 模式的并行随机森林(MR_RF)的研究有以下几方面。就性能受分块大小即 m 影响这方面，Kleiner 等人[12]提出了新的 BLB(bag of little BootStrap)，并证明其与 BootStrap 都有一样的统计性能；就不同大数据处理分析方法方面，Genuer 等人[13]结合 R 平台的并行随机森林展开研究，对比二次抽样、BLB 及分治三种算法处理大数据的准确率和效率问题；就 MR_RF 用于不平衡大数据的处理方面，Río 等人[14]对比分析过采样、欠采样和代价敏感方法对不平衡数据处理的有效性。而 MR_RF 的袋外估计泛化误差研究得比较少。基

收稿日期：2017-02-15；**修回日期**：2017-03-20 **基金项目**：国家自然科学基金资助项目(61673193)；中央高校基础研究资助项目(JUSRP51510,JUSRP51635B)

作者简介：钱雪忠(1967-)，男，江苏无锡人，副教授，主要研究方向为数据库技术、数据挖掘、网络安全等(emily139617@126.com)；秦静(1992-)，女，江苏无锡人，硕士研究生，主要研究方向为数据挖掘；宋威(1981-)，男，湖北恩施人，副教授，博士，主要研究方向为数据挖掘、人工智能和模式识别、信息检索.

于此,本文在借鉴前人研究的基础上大胆尝试,首先对比分析 MR_RF 与 RF 的区别;其次根据分而治之的 MapReduce 思想,设计新的包外估计方法;最后利用该方法与交叉验证方法对比验证其有效性,与此同时,改变分块的大小,观测分类效率的变化。

1 相关知识

MR_RF 是由 Mahout 开源项目实现的基于 MapReduce 机制的并行化随机森林算法。其主要内容有:

MapReduce 机制:能够将工作流和数据分配到分布式集群节点之上,然后各个节点并行计算。作为一种线性可伸缩的编程模型,将任务分为 map(映射)和 reduce(约简)两个处理阶段,每个阶段都以一系列的〈key,value〉键值对作为输入输出。

MR_RF 步骤:算法分为模型构建和数据分类预测两个阶段。

a)森林构建。基于分治思想,将数据分区(一般按 64 M/分区)传递给不同的数据节点进行计算,在 map 函数中,针对分区数据的 bagging 子集用 RF 算法构建基分类器,map 任务并行运行,多个决策树汇聚成森林。

b)分类预测。首先将测试集分成几个互相独立的子集分发到不同的数据节点,在 map 任务中,遍历森林里的每一棵树,采用简单投票,取预测结果的众数作为最终的分类结果。将结果汇聚,计算分类准确率。

利用 OOB 对 RF 泛化性能估计:给定数据集 $D[N,S]=\{(x_i,y_i),i=1,2,\cdots,N\}$,其中 $x_i\in R^S$ 为预测变量,$y_i\in\{1,2,\cdots,J\}$ 是类标签。对每个观测(x_i,y_i),令 $V_i=\{m:(x_i,y_i)\notin D^{(m)}\}\subset\{1,2,\cdots,M\}$ 表示不包含它的 BootStrap 样本集编号,利用基分类器集合 $C_{m\in V_i}$对 x_i 进行预测,则预测的类标签见式(1)。泛化误差见式(2)。

$$y_i^*=\arg\max_{1\leqslant b\leqslant J}N_b \tag{1}$$

其中:

$$N_b=\mathrm{card}\{m:C_{m\in V_i}(x_i)=b\}$$

$$\varepsilon^{OOB}=\frac{1}{N}\sum_{i=1}^{N}I(y_i\neq y_i^*) \tag{2}$$

其中:$I(\ )$是取值为 1 或 0 的示性函数;card(\ )是计数函数。许多实验结果表明[7~9],上述袋外估计泛化误差与采用交叉验证法所得的结果之间几乎没有差别,某些情况下甚至更好。

2 改进的并行随机森林及其袋外估计

2.1 MR_RF 与 RF 的区别

RF 将多棵树的构建任务分而治之,而 MR_RF 将 RF 借助 Hadoop 分布式平台,使构建任务和数据同时分而治之。MR_RF 与直接在全局数据上处理的 RF 区别具体如下:a)分块原则仅仅是根据给定块大小将数据截断成块,而不考虑块数据的代表性和内部类别分布,文献[15]指出现实生活中的大数据很少有随机散布的,相反是根据某个属性团簇在一起,这样分块的数据一致性高而构建出的基分类器多样性低;b)不同于在整个数据集上的自助抽样 BootStrap,现在是在分块上有放回地自助抽样,类似 m out of n BootStrap,该方法最终的分类准确率与分块大小 m 的选取有很大的关系[12];c)由于分块之间数据是互不可见的,树构建时袋内数据和袋外数据的关系也只局限在块内,为此设计了 MR_RF 袋外泛化误差估计方法。

2.2 MR_RF 的袋外泛化误差估计

模仿 RF 的袋外泛化误差估计计算方法,本文把计算范围缩小到块上。OOB 泛化性能估计:

对块 τ_k 上每个观测样本(x_i,y_i),令 $V_i=\{m_k:(x_i,y_i)\notin D_{\tau_k}^{m_k}\}\subset\{1,2,\cdots,m_k\}$ 表示不包含它的 BootStrap 样本集编号,利用基分类器集合 $C_{m_k\in V_i}$对 x_i 进行预测,令 $N_b=\mathrm{card}\{m_k:C_{m_k\in V_i}(x_i)=b\}$,则预测的类标签见式(3),块泛化误差见式(4),MR_RF 的泛化误差见式(5)。

$$y_i^*=\arg\max_{1\leqslant b\leqslant J}N_b \tag{3}$$

$$\varepsilon^{OOB\_\tau k}=(1/N_{\tau_k})\sum_{i=1}^{N}I(y_i\neq y_i^*) \tag{4}$$

$$\varepsilon^{OOB}=\frac{1}{N}\sum_{k=1}^{K}N_{\tau_k}\varepsilon^{OOB\_\tau k} \tag{5}$$

下节将通过实验验证 MR_RF 的包外泛化误差估计的有效性和分块大小对分类准确率的影响。

2.3 改进的 MR_RF

改进的 MR_RF 利用袋外数据直接给出分类模型的泛化误差估计,省去了预测阶段。分类阶段由 initial、map 和 final 三个步骤组成。

a)Initial 阶段。首先将数据分区到不同的数据节点上,这个由 Hadoop 平台自动完成。

b)Map 阶段。输入键值对〈key,value〉,key 为编码一个实例的二进制流,value 是实例的具体数据。在对各个分区数据进行自助的数据集构建代价敏感 CART 树。特别地用三个统计变量记录过程中的计算数据。第一个统计变量记录袋内袋外关系,另一个全局变量统计分块上数据与树构建的关系,最后用一个统计变量计算出 MR_RF 的泛化误差值,计算所得树的强度和分块泛化误差都存入 HDFS 文件中。输出键值对〈key′,value′〉,key′是树 ID 和分区 ID 组合字符串,value′是一个基分类器,包括基分类器本身和它的强度。

c)Final 阶段。汇总各个分区树,形成森林,放入分布式缓存。图 1 是改进的 MR_RF 构建树流程。

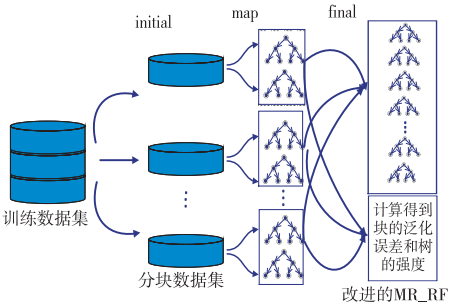

图1 改进的MR_RF构建树流程

3 实验

3.1 实验环境

本文搭建的 Hadoop 平台为服务器虚拟的 4 台集群节点,由 100 M 宽带互连。配置为 CPUIntel Xeon® E5-2630、内存 4 GB、缓存 15 MB、Hadoop 版本为 Hadoop 2.0.0 CDH 4.5、OS 版本为 Centos 7.0、Mahout 版本为 Mahout 0.7 CDH 4.5。

3.2 实验数据集

本文选取的数据集为 KDD Cup 1999,将 KDD Cup 1999 这个多分类数据集按照不同类别形成不平衡二元分类大数据集,

从上至下数据量依次增大。数据集的具体信息如表 1 所示。其中:nIns 表示实例的数量;nAttr 表示属性的数量;class(n : p)表示负类与正类的数量;IR 是负类与正类数量的比值,其表示全局不平衡指数。

表 1　不平衡数据集

数据集	nIns	nAttr	class(n : p)	IR
normal_VS_U2R	972833	41	972781 : 52	18707.327
normal_VS_R2L	973907	41	972781 : 1126	863.926
normal_VS_PRB	1013883	41	972781 : 41102	23.667
DOS_VS_U2R	3883422	41	3883370 : 52	74680.192
DOS_VS_R2L	3884496	41	3883370 : 1126	3448.819
DOS_VS_PRB	3924472	41	3883370 : 41102	94.481
DOS_VS_normal	4856151	41	3883370 : 972781	3.992

3.3　不平衡数据评价指标

传统的分类学习方法中,一般采用分类精度作为评价指标。然而该指标对于不平衡数据集不适用,因为正类比例如果不足 1% 的时候即使正类被分为负类,精度依旧可以达到 99%,所以本文采用 G-mean 标准为评价指标,它是由混淆矩阵而来,具体如表 2 所示。其中:*TP* 表示正类样本判为正类的数目;*TN* 表示负类样本判为负类的数目;*FN* 与 *FP* 分别表示实际为正类和负类而判断错误的样本。

表 2　混淆矩阵

分类	预测正类	预测负类
实际正类	*TP*	*FN*
实际负类	*FP*	*TN*

G-mean 值表示的是正类召回率和负类召回率的几何平均值。(+)表示越大越好。其定义如下:

$$\text{G-mean}(+) = \sqrt{\text{sen} \times \text{spe}} \tag{6}$$

其中:sen 为 $\frac{TP}{TP+FN}$;spe 为 $\frac{TN}{FP+TN}$。

G-mean 只有在负类和正类的分类召回率同时都高的情况下,G-mean 的值才最大。那么利用袋外数据估计 MR_RF 的 G-mean 值为

$$\varepsilon^{\text{OOB\_GM}} = \frac{1}{N}\sum_{k=1}^{K} N_{\tau_k} \varepsilon^{\text{OOB\_GM}_{\tau_k}} \tag{7}$$

其中:$\varepsilon^{\text{OOB\_GM}_{\tau_k}} = \sqrt{\text{sen}_{\tau_k} \times \text{spe}_{\tau_k}}$;$\text{sen}_{\tau_k} = \frac{1}{N_{\tau_k,p}} \sum_{i=1}^{N_{t_k,p}} I(y_i = y_i^*)$;$\text{spe}_{\tau_k} = \frac{1}{N_{\tau_k,n}} \sum_{i=1}^{N_{t_k,n}} I(y_i = y_i^*)$;$N_{\tau_k,p}$ 表示分块 τ_k 的正例数;$N_{\tau_k,n}$ 表示分块 τ_k 的负例数。

3.4　实验设计

a)在默认块大小(64 MB)下,本文比较 MR_RF 在交叉验证和袋外泛化误差估计两方法之间的差值,以验证新的泛化误差估计的有效性;b)测试 MR_RF 在不同分块大小下精度的变化,并观测新的泛化误差的有效性;c)测试不同分块大小对 MR_RF 分类效果的影响;d)测试分块大小对新泛化误差对预测时间的影响。

MR_RF 的参数集中在森林构建阶段。其中,树的最大深度默认为无限,属性数量为 sqrt(M),M 为属性总数,K 表示树的数量设置为 100。本文采用 5 次 5 折分层交叉验证,取其均值,袋外误差估计方法也是 5 次结果的平均值。

3.5　实验结果分析

1)MR_RF 袋外泛化误差估计的有效性验证

首先本文比较 MR_RF 在不同数据集上利用交叉验证方法估计泛化误差 e^{TSGM} 与包外泛化误差估计方法 e^{OOBGM} 的差别,用 $|e^{\text{TSGM}}-e^{\text{OOBGM}}|$ 表征两者之间的差距,$e^{\text{OOB}}/e^{\text{TS}}$ 越接近 1 说明两者越接近。表 3 是两种算法在不同数据集上的平均 G-mean 值。从表中可以看出以下结论:

a)多个数据集结果平均值可以看出两者相差千分之一,两者的比值也达到 0.999,所以说两种方法衡量数据集的泛化误差能力近似,而利用袋外误差方法可以在树构建的同时给出树性能的估计以及森林的泛化性能估计,而免去了多次 k 折交叉验证的计算时间,大数据量环境下的时间效率有很大提高。

b)随机森林在遇到极不平衡的数据集(如 Kddnormal_VS_U2R 和 kddDOS_VS_U2R)时分类性能与随机二分类性能相差无几,分类效果不佳。而对于相对平衡(如 kddDOS_VS_R2L)和几近平衡的数据集(如 kddDOS_VS_normal)时,并行随机森林的分类效率很高。

表 3　MR_RF 泛化性能(GM)

数据集	ave^{TS}	ave^{OOB}	$\text{ave}\lvert e^{\text{TS}}-e^{\text{OOB}}\rvert$	$e^{\text{OOB}}/e^{\text{TS}}$
normal_VS_U2R	0.530	0.536	0.006	1.011
normal_VS_R2L	0.937	0.933	0.004	0.996
normal_VS_PRB	0.995	0.995	0.000	1.000
DOS_VS_U2R	0.548	0.538	0.010	0.982
DOS_VS_R2L	0.991	0.992	0.001	1.001
DOS_VS_PRB	1.000	1.000	0.000	1.000
DOS_VS_normal	1.000	1.000	0.000	1.000
平均值	0.857	0.856	0.001	0.999

2)测试不同分块大小对 MR_RF 泛化误差衡量效果的影响

因为与 normal 组合的三个数据集是大小在同一级别的,所以取几近平衡的数据集 Kddnormal_VS_PRB、相对不平衡数据集 Kddnormal_VS_R2L 和极不平衡数据集 Kddnormal_VS_U2R 进行具体分析。根据不同的分块大小,比较其精度的变化,Kddnormal_VS_PRB 如图 2 所示。

从图 2 中可以看出 eoobGM 与 etsGM 两个值随着分片的增大,几乎重叠在一起,说明两者界定泛化误差的能力相近。

对于相对不平衡数据集 Kddnormal_VS_R2L 来说,从图 3 可以看出:eoobGM 与 etsGM 两个值随着分片的增大,开始出现了越来越大的偏差。随着分片的增加,etsGM 测算的泛化误差略微下降,从 0.937 下降到 0.922,而 eoobGM 却从 0.933 下降到 0.88。对 eoobGM 的值测算来自分块数据结果,此时分块大小改变对其值影响大。不仅是 GM 值,连 sen 值也是如此,oobsen 随着分块增多,急速递减,而 etssen 下降平缓,两者的差距也随着分块数增加越来越大,而分块对负类的召回率几乎没有影响,最终使得整体 GM 值随着分块的增大而减小。所以分块性能的综合没有总体性能好。

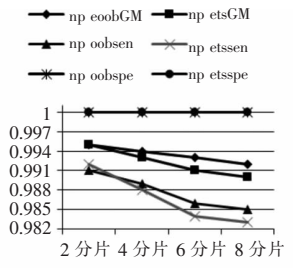

图 2　Kddnormal_VS_PRB

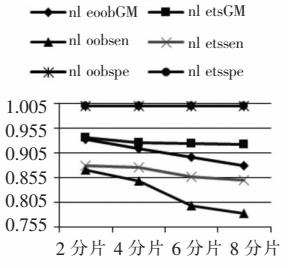

图 3　Kddnormal_VS_R2L

对于极不平衡数据集 Kddnormal_VS_U2R,从图 4 可以看出,eoobGM 随着分块数的增加,从 2 分片的 0.536 逐渐递减到 8 分块的 0.181,而 etsGM 在分块 2 时还有 0.53,而随着分块的增加锐减趋于 0,分块性能的综合比整体性能来得好。

3)测试不同分块大小对 MR_RF 分类效果的影响

从图 2 中可以看出:a)随着分片的增加,GM 值略微下降;b)虽然 sen 和 spe 都随着分片的增加而略微减小,但是 MR_RF 对正负类的召回率都很高,所以整体的 GM 值也很高,MR_RF 能轻松处理几近平衡问题。

从图 3 中这个相对不平衡数据集:所有指标值都随着分块数量的增加而呈现非递增的趋势。对正类的召回率随着分块上数据的减少而减少,而对负类召回率却没有影响,所以整体的分类效果也随着分块数的增加而递减。

而对于极不平衡数据,从图 4 可以看出,MR_RF 处理能力弱,分块数量对分类效率影响明显,随着分块有效数量的减少,少有正类被识别,分类器失去意义。

4)测试分块大小对新泛化误差预测时间影响

纵坐标值 a 和 b 分别表征 min 和 s。从图 5 可以看出,随着分块数的增加,改进的 MR_RF 算法处理时间缩短。需要注意的是,当分块为默认大小时,数据量较多,且多次自助抽样耗时又占用大量内存,所以任务容易内存溢出而重启,最终导致分块为 2 时,时间需要很长。

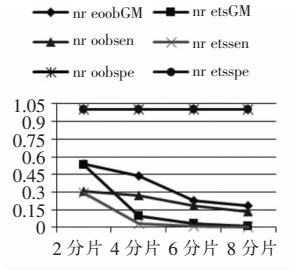

图 4 Kddnormal_VS_U2R

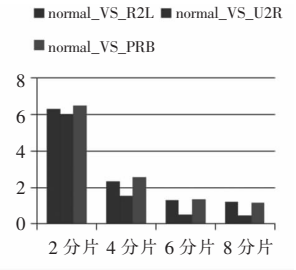

图 5 分块对新泛化误差预测时间的影响

3.6 分类器强度和选择性集成方法

首先分析出现上述袋外泛化误差随着分块增加出现偏差的原因。总的来说,对几近平衡数据集随着分块数量的增多,分块结果的均值 eoobGM 与总体 etsGM 衡量效果近似。对于相对不平衡数据分块性能的综合不如总体的性能,但对于极不平衡数据集,分块性能又优于总体的性能。对比 MR_RF 的预测过程和图 1 的预测过程。MR_RF 利用五折交叉验证测算出来的 etsGM 是取分块树集合的众数投票结果,而 MR_RF 利用袋外数据预测是在分块树这个小的集合上进行的众数投票,所以存在偏差是正常的。而为了缩小这种偏差,就需要保证分块的性能良好,最后集成最终的全局分类器也有良好的性能,为此提出了选择性集成的设想。

同样利用袋外数据来计算每个基分类器的分类准确率,也就是基分类器的强度。在块上抽样的时候本文就记录下袋外数据和袋内数据,对于特定块,有一个自助抽样集 $\boldsymbol{\Theta}^i_{\tau_k}$,其袋外数据用 $O(\boldsymbol{\Theta}^i_{\tau_k})$ 记录,遍历袋外每个数据,计算该分类器的强度,见式(8):

$$s(t)=s(h(\boldsymbol{\Theta}^i_{\tau_k}))=\frac{1}{N_{\tau_k}}I(h(x,\boldsymbol{\Theta}^i_{\tau_k})=y) \tag{8}$$

对于如何选择,借鉴 AdaBoost[16] 简单而鲁棒的方法,树的权重记为 $\alpha_t=\frac{1}{2}\ln(\frac{s_t}{(1-s_t)})$,设置一定阈值,选择符合条件的树进行集成。

4 结束语

本文首先对比 MapReduce 模式下的随机森林算法与顺序版的区别,MR_RF 受分块操作影响使得传统包外估计很难计算。新设计的袋外泛化误差估计方法将计算局限在块上,并采用块平均作为整体的 MR_RF 泛化误差估计。实验证明该方法在默认分块下是衡量泛化误差的有效方法,但随着分块的增加,新的泛化误差的有效性存在偏差。在分析原因的基础上提出了选择性集成方案。其次数据集分类准确率受不平衡程度和分块大小的影响,表现为随着分块数的增多和不平衡指数的增大而减小;而分类的时间效率又随着分块的增大而减小,所以 MR_RF 处理不平衡的数据分类时需要调节分块大小以获取效率和速率的折中。

参考文献:

[1] Gupta R, Gupta H, Mohania M. Cloud computing and big data analytics: what is new from databases perspective? [M]//Big Data Analytics. Berlin: Springer, 2012: 42-61.

[2] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[3] Hasan M A M, Nasser M, Pal B, *et al*. Support vector machine and random forest modeling for intrusion detection system (IDS)[J]. Journal of Intelligent Learning Systems and Applications, 2014, 6(1): 45-52.

[4] Klassen M, Paturi N. Web document classification by keywords using random forests[M]//Networked Digital Technologies. Berlin: Springer, 2010: 256-261.

[5] Thaweekote V, Songram P, Jareanpon C. Automatic nipple detection based on face detection and ideal proportion female using random forest[C]//Proc of IEEE International Conference on Computational Intelligence and Cybernetics. Piscataway, NJ: IEEE Press, 2013: 11-15.

[6] 张春霞, 郭高. Out-of-bag 样本的应用研究[J]. 软件, 2011, 32(3): 1-4.

[7] Wolpert D H, Macready W G. An efficient method to estimate bagging's generalization error[J]. Machine Learning, 1999, 35(1): 41-55.

[8] Bylander T. Estimating generalization error on two-class datasets using out-of-bag estimates[J]. Machine Learning, 2002, 48(1-3): 287-297.

[9] Hernándezlobato D, Martínez-Mñoz G, Suárez A. Out of BootStrap estimation of generalization error curves in bagging ensembles[C]// Proc of International Conference on Intelligent Data Engineering and Automated Learning Ideal. Berlin: Springer, 2007: 47-56.

[10] Martínez-Muñoz G, Suárez A. Out-of-bag estimation of the optimal sample size in bagging[J]. Pattern Recognition, 2010, 43(1): 143-152.

[11] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[12] Kleiner A, Talwalkar A, Sarkar P, *et al*. A scalable BootStrap for massive data[J]. Journal of the Royal Statistical Society, 2014, 76(4): 795-816.

[13] Genuer R, Poggi J M, Tuleau-Malot C, *et al*. Random forests for big data[J]. Big Data Research, 2017, 9(9): 28-46.

[14] Río S D, López V, Benítez J M, *et al*. On the use of MapReduce for imbalanced big data using random forest[J]. Information Sciences, 2014, 285(C): 112-137.

[15] Laptev N, Zeng Kai, Zaniolo C. Early accurate results for advanced analytics on MapReduce[J]. Proceedings of the VLDB Endowment, 2012, 5(10): 1028-1039.

[16] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.