随机森林中树的数量

刘 敏,郎荣玲,曹永斌
LIU Min, LANG Rongling, CAO Yongbin

北京航空航天大学 电子信息工程学院,北京 100191
College of Electrical and Information Engineering, Beihang University, Beijing 100191, China

LIU Min, LANG Rongling, CAO Yongbin. Number of trees in random forest. Computer Engineering and Applications, 2015, 51(5):126-131.

Abstract: Random Forest (RF) is a kind of ensemble classifier. This paper analyses the parameters influencing the performance of RF, and the result shows that the number of trees in random forest has significant effect on its performance. This paper carries on a research and summary on the method of determining the number of trees and evaluating the performance index of RF, with the classification accuracy used as the evaluation method, utilizing UCI data sets, an experimental analysis on the relationship between the number of decision trees in random forest and the data sets has been done. The experimental result shows that for the majority of data sets, when the number of trees is 100, the classification accuracy can meet the requirement. This paper compares RF with support vector machine having superior classification performance in the aspect of accuracy, and the result shows that the classification performance of random forest is similar to that of support vector machine.

Key words: random forest; support vector machine; classification accuracy

摘 要: 随机森林是一种集成分类器,对影响随机森林性能的参数进行了分析,结果表明随机森林中树的数量对随机森林的性能影响至关重要。对树的数量的确定方法以及随机森林性能指标的评价方法进行了研究与总结。以分类精度为评价方法,利用UCI数据集对随机森林中决策树的数量与数据集的关系进行了实验分析,实验结果表明对于多数数据集,当树的数量为100时,就可以使分类精度达到要求。将随机森林和分类性能优越的支持向量机在精度方面进行了对比,实验结果表明随机森林的分类性能可以与支持向量机相媲美。

关键词: 随机森林; 支持向量机; 分类精度
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1401-0264

1 概述

随机森林(Random Forest,RF)是由Leo Breiman将Bagging集成学习理论[1]与随机子空间方法[2]相结合,于2001年提出的一种机器学习算法[3]。RF利用bootstrap重抽样方法从原始样本中抽取同原始数据样本集个数相同的多个样本构成样本子集,利用每个样本子集构建决策树,然后融合多棵决策树的预测结果。

在构建RF时,有三个主要参数影响RF的性能和效率:

(1)森林中树的数量

设 *Ntree* 表示RF中树的数量。当 *Ntree* 较小时,RF的分类误差大、性能也比较差。另一方面,RF具有不过拟合性质,因此可以使 *Ntree* 尽量大,以保证集成分类器的多样性。但是构建RF的复杂度与 *Ntree* 成正比, *Ntree* 过大,使得RF构建时间花费过大。同时森林的规模达到一定程度时,将导致森林的可解释性减弱。因此在2002年数理统计研究所年会的Wald讲座上,Leo Breiman将RF称之为“黑箱”[4]。因此, *Ntree* 对RF的性能、可解释性和复杂性之间的平衡都具有重要意义。

(2)候选特征子集

在构建RF的过程中,为了保证随机性,在每棵决策

基金项目: 国家自然科学基金(No.61202078);国家高技术研究发展计划(863)(No.2011AA110101,No.2012AA121801)。

作者简介: 刘敏(1989—),女,在读硕士,研究领域为故障诊断;郎荣玲,女,讲师,研究领域为记载设备故障诊断与故障预报;曹永斌(1989—),男,在读硕士,研究领域为人工智能。E-mail:ronglinglang@163.com

收稿日期: 2014-01-17 **修回日期:** 2014-06-16 **文章编号:** 1002-8331(2015)05-0126-06
CNKI网络优先出版:2014-06-24, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1401-0264.html

树的分裂节点，需要从原始特征集合中随机地抽取一个特征子集，子集中包含 *Mtry* 个候选特征（*Mtry* 一经选定，则在建树的过程中保持不变），再从候选特征中选择最优特征作为该节点的分裂特征。决策树通过这种多样化的分裂标准使得RF具有随机性。但是对于RF而言，过于随机化，即在每个节点选择过少的相关特征，所建的树的分类精度就会降低；相对的，大量选择相关特征，降低了分裂节点处的随机性。 *Mtry* 越小，随机化越强，分类精度就会相应降低， *Mtry* 越大，RF的多样性就会降低。因此 *Mtry* 可以用来协调RF中分类性能和多样性之间的平衡。目前的研究中 *Mtry* 多取为 $Mtry=\sqrt{M}$ 或 $Mtry=\mathrm{lb}\,M+1$ 。

（3）叶节点的样本数

设Nodesize表示叶节点的最小样本数，RF的性能对Nodesize不敏感[5-6]。

综上可以看出， *Ntree* 是随机森林的一个非常重要的指标，因此本文主要研究 *Ntree* 的选取。对于 *Ntree* 的选取，目前最常用的方法是固定 *Ntree*，就是指在取定 *Ntree* 情况下，构建森林。但是研究结果表明 *Ntree* 的取值与具体数据集的特点密切相关[7]。因此最近很多研究者致力于根据数据的特点设置 *Ntree* 值的研究，RF的修剪和动态确定是目前两种常用的根据数据特点确定 *Ntree* 的方法。

大量研究结果表明，很多情况下森林中1/2的树有时甚至1/6的树就能得到好的分类或预测结果[8]，因此可以对森林进行修剪，从过度生产的森林中选出优异的子集。RF修剪修剪过程中主要解决两个问题：选择的标准和选择的策略。

目前随机修剪的标准主要有：

（1）分类性能

选择的目的是为了获得性能优异的子集，因此很多研究者从分类性能出发，挑选森林中分类性能优异的决策树。评价RF性能的指标有精度[9-10]、ROC、AUC[7]等。

（2）多样性

集成分类器为了获得更好的性能，需要集成分类器中的元分类器（如RF中的决策树）具备多样性[8]。尽管目前还没有对多样性的统一定义，但是它是集成分类器泛化性能的重要特征之一[11]。

在确定了随机修剪准则的基础上，需要确定选择的策略，目前主要有加权[12-13]、搜索[11,14]、聚类[15-16]、排序[17]和最优化[18-19]等选择方法。

动态确定 *Ntree* 方法（也称为动态剪枝），在构建森林时，每一步，构建一棵新的决策树加入到森林中，如果这棵树使得森林的分类性能或多样性提升，则这棵树被保留[20-22]。动态构建森林避免了过度生产，节约了运行时间和存储空间，但因其研究相对较少，针对其停止准则、选择标准等问题还需要做进一步研究[8]。

由上面的分析可以看出，研究 *Ntree* 如何影响RF的性能是确定 *Ntree* 的基础。因此本文主要研究 *Ntree* 对RF性能的影响，下面首先介绍RF的评估方法。

2　随机森林的评估方法

RF是由一组决策树分类器 $\{h(\boldsymbol{x},\theta_k),k=1,2,\cdots,L\}$ 组成的集成分类器，其中 $\{\theta_k\}$ 是独立同分布的随机向量，L 是RF中决策树的个数，每棵树对输入数据向量 $\boldsymbol{x}$ 归属于哪一类进行未加权投票。这一章主要介绍对RF性能评估的主要方法。

2.1　精度

一般情况下，可以从三个方面评价分类器：精度，计算复杂度，模型描述的简洁度，其中分类器的精度是分类器性能的重要评价指标。精度通常是用分类器对于数据的分类正确率或是与之等价的错误率来衡量，它可以估计一个给定的分类器对数据正确分类的能力。下面介绍几种常用的精度估计的方法。

2.1.1　袋外数据估计方法

袋外数据（Out Of Bag，OOB）估计方法可以在不加入测试样本的情况下，评估森林以及森林中每棵树的分类精度。Breiman通过实验方法得到使用OOB数据得到的误差估计是无偏估计[3]。

定义 2.1.1（OOB数据）在建立包含 L 棵树的RF的过程中，共需要 L 个训练样本子集，设这 L 个训练样本子集集合为 $\{S^i|i=1,2,\cdots,L\}$，每个 $S^i(i=1,2,\cdots,L)$ 通过Bootstrap抽样方法对样本集 $S=\{(x_i,y_i)|i=1,2,\cdots,N\}$ 进行抽样得到。 S 中每个样本 $(x_i,y_i),i=1,2,\cdots,N$ 未被抽取的概率为 $(1-1/N)^N,N=|S|$ 。 $\lim\limits_{N\to\infty}(1-1/N)^N\approx 0.368$，这表明对于任意 $S^i\subset S(i=1,2,\cdots,L)$，$S$ 中存在约37%的样本不会出现在 S^i 中，这一部分的数据称之为样本集 S^i 的袋外数据。

定义 2.1.2（OOB决策树集合）针对样本 S 中的任意一个样本 (x_i,y_i)，存在一个或者多个样本子集不包含样本 (x_i,y_i)，即存在样本子集 S^C 使得 $(x_i,y_i)\notin S^C$ 。设 $S'=\{S^C|C\in\{1,2,\cdots,L\}$，且 $(x_i,y_i)\notin S^C\}$，则 S' 中的样本子集均不包含样本 (x_i,y_i)。称由 S' 中的集合 $S^C,C\in\{1,2,\cdots,L\}$ 构建的决策树为样本 (x_i,y_i) 的OOB决策树集合。

定义 2.1.3（OOB误差）利用OOB决策树集合对样本 (x_i,y_i) 进行分类，记分类结果为 $f(x_i,y_i)$，则 $f(x_i,y_i)$ 表示OOB决策树集合中的所有树对数据 (x_i,y_i) 进行分类，将多个分类结果进行少数服从多数的投票得到的分类结果。

针对样本集 S ，OOB误差估计定义为：

$$\text{OOB error}=1-\frac{\sum_{i=1}^{N}I(f(x_i,y_i)=y_i)}{N}\qquad(1)$$

其中 I 是示性函数，$I(f(x_i,y_i)=y_i)=1$ 表示样本 (x_i,y_i) 的OOB决策树集合对数据 (x_i,y_i) 进行分类，结果正确。

2.1.2 K 交叉验证估计方法

K-次交叉验证是一种对分类机的精度进行验证的常用方法。该方法将样本集 S 分割成 K 个样本子集，一个子集被保留作为测试数据，其他 $K-1$ 个集合用来训练分类机。重复 K 次，平均 K 次的误差结果，最终得到误差估计。K 交叉验证，可以避免数据集的选取对结果的影响。

但是正确率的度量标准在当前的实际应用中发现了很多限制和不足。在数据不平衡的情况下，即数据的类别分布相差很大时，正确率并不能准确表达分类器的性能。如正确率为0.99的分类算法可能比随机猜测所有的信用卡使用都是“正常使用”的正确率要低，很高的正确率此时并不能够充分说明分类器性能的好坏。另一方面当分类错误代价不相等时，正确率只能保证出现错误的数量最小，但是并不能保证是总体代价最小。如在一些与人身生命安全相关的重要决策中，这种错误代价的关系往往无法具体衡量。从实用角度讲，分类学习算法应该尽量地减少代价高的错误出现，而不是着重于减少错误数量。因此有研究者提出利用受试者工作特征（Receiver Operating Characteristic，ROC）来衡量算法的精度。

2.1.3 ROC

ROC分析20世纪50年代起源于统计决策理论，Spackman最早将ROC分析技术引用到机器学习领域中，用来说明分类器命中率和误报率之间的关系。下面以二分类问题为例，介绍一下ROC的定义。

最简单的二分类问题是正例（Positive）和负例（Negative）的分类情况。给定一个分类器和一个实例，存在四种输出情况，如表1所示。

表1　二分类器输出

预测类别	真实类别	
	正例（Positive）	负例（Negative）
正例（Positive）	True Positive	False Positive
负例（Negative）	False Negative	True Negative

当正例被分类器预测分为正例时，为True Positive；被预测为负例时，为False Negative。反之，亦然。给定一个分类器和一个数据集合，定义：

$$\text{FP Rate}(FPR)=\frac{\text{错分负例个数}}{\text{负例总数}} \tag{2}$$

$$\text{TP Rate}(TPR)=\frac{\text{正确分类正例个数}}{\text{正例总数}} \tag{3}$$

ROC图是以 FPR 为 X 轴，以 TPR 为 Y 轴的二维曲线图，其中横轴和纵轴的长度相等，为单位1，如图1。

每个分类器在数据集上只产生一个（FP，TP）对，对应于ROC图中的一个点。ROC分析技术不仅是一种通用图形化分析性能的方法，而且ROC曲线的独特属性使它适合应用于类别分布不平衡或者分类错误代价不相等的领域，这使它在类别分布未知的领域和代价敏感学习中变得越来越重要。

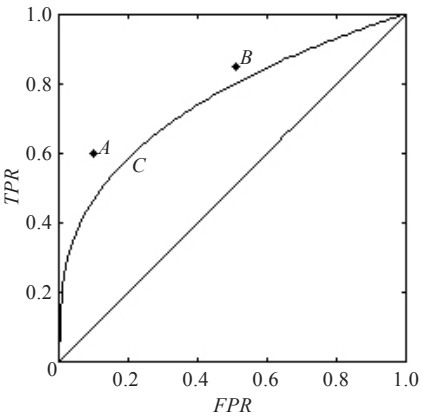

图1　ROC曲线

2.1.4 AUC

ROC曲线是对分类性能的二维图形描述，当对多个分类器进行性能比较时，无法对多条曲线进行直观的比较。因此为了能够直接比较多个分类器在任何比例分布和任何错误代价比的情况，可以将ROC曲线描述的分类器性能转换为一个数值来表示分类器的性能。一个通用的方法是计算ROC曲线下的面积AUC。根据ROC曲线中对优秀分类器的要求利用数学形态学对图像进行分析[22]，可以得到AUC值大的分类器性能优越，AUC值小的分类器性能次之。

2.2 多样性

RF作为一种集成分类器，为了获得更好的性能，RF中的决策树需要具备多样性[8]，即树之间需要存在不同（即互补）。它是RF提升泛化性能的最重要特征之一[11]。目前对RF多样性还没有统一的定义[23]，研究者们常用树之间的相关性和相似性来衡量RF的多样性。

2.2.1 相关性

设 $\{h(x,\theta_k),k=1,2,\cdots,L\}$ 为RF，$\{\theta_k\}$ 为独立同分布的随机向量；森林中每棵决策树针对数据 (x,y) 的分类结果为 $h(x,\theta_k),k=1,2,\cdots,L$；训练数据集的决策类别集为 $S^y=\{y_i|i=1,2,\cdots,N\}$。定义：

$$\hat{j}(x,y)=\arg\max_{j\neq y}\{g(j)=\sum_{m=1}^{L}I(h(x,\theta_m)=j)\} \tag{4}$$

其中 $j\in S^y$。上式代表了 $\{h(x,\theta_k),k=1,2,\cdots,L\}$ 对 (x,y) 进行分类，分类结果 j 除去正确类别 y，投票得票数 $g(j)$ 最多的类别。

定义原始余量函数（Raw Margin Function）为：

$$rmg(\theta,x,y)=I(h(x,\theta)=y)-I(h(x,\theta)=\hat{j}(x,y)) \tag{5}$$

由式（5）可得，输入数据 (x,y)，$rmg(\theta,x,y)$ 的可能取值如下：

$rmg(\theta,x,y)$	条件	单棵树对 (x,y) 分类结果
1	$h(x,\theta)=y$	正确
−1	$h(x,\theta)=\hat{j}(x,y)$	错误
0	$h(x,\theta)\neq y$且$h(x,\theta)\neq\hat{j}(x,y)$	错误

对于独立同分布的向量 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}'$，$rmg(\boldsymbol{\theta},x,y)$ 和 $rmg(\boldsymbol{\theta}',x,y)$ 的相关系数定义为：

$$\rho(\boldsymbol{\theta},\boldsymbol{\theta}')=\frac{\mathrm{cov}_{X,Y}(rmg(\boldsymbol{\theta},x,y),rmg(\boldsymbol{\theta}',x,y))}{sd(\boldsymbol{\theta})sd(\boldsymbol{\theta}')} \tag{6}$$

其中：

$$sd(\boldsymbol{\theta})=\sqrt{D_{X,Y}(rmg(x,\boldsymbol{\theta}))}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\{rmg(x_i,\boldsymbol{\theta})-\frac{1}{N}\sum_{i=1}^{N}rmg(x_i,\boldsymbol{\theta})\}^2}$$

表示了 $rmg(\boldsymbol{\theta},x,y)$ 对于数据集 (X,Y) 的标准差。$\rho(\boldsymbol{\theta},\boldsymbol{\theta}')$ 可用于衡量森林中任意两棵树 $h(x,\boldsymbol{\theta})$ 和 $h(x,\boldsymbol{\theta}')$ 对数据集 (X,Y) 的分类结果的相关程度。$\rho(\boldsymbol{\theta},\boldsymbol{\theta}')$ 越大，两棵树的相关程度越大。

2.2.2　相似性

相似性的概念也是用来描述森林中树之间的关系。

定义：

$$\rho_{\theta_i}=\frac{1}{K-1}\sum_{\theta_i\neq\theta_j}\rho(\theta_i,\theta_j) \tag{7}$$

ρ_{θ_i} 表示森林中的树 θ_i 与森林中其他树之间的平均相似性大小，衡量了森林中每棵树对于森林的互补作用。ρ_{θ_i} 值越大，树与森林其他树的关联程度越大，则树 θ_i 对森林的互补作用越小。

3　实验分析

实验分为两个部分，首先以精度为依据分析RF中 *Ntree* 与数据特点的关系，第二个实验将RF与支持向量机（Support Vector Machine，SVM）在精度方面进行比较。实验中采用的数据均来自UCI数据库[24]，实验数据信息如表2。

数据密度定义为：

$$D=\log_a\frac{N}{c} \tag{8}$$

其中 N 是数据中的样本个数，c 是数据样本的类别个数，a 是数据样本的条件属性个数。

在参数 $Mtry=\sqrt{M}$ 和 $Nodesize=1$ 下，数据的test error和OOB error与 *Ntree* 的关系如图2和表3所示。

从图2和表3实验结果可以看出：

（1）在森林中树的个数不足时，RF的分类精度test error和OOB error随着树的增长而迅速下降。

（2）此现象满足RF作为集成分类器优于单个决策树分类器的特征。

（3）当树的个数足够多时，test error和OOB error趋于稳定，在一定的值上下小幅度波动。由表3可以看出当 $Ntree>100$ 时，即使 *Ntree* 的值大幅度增加时，RF对于数据的分类错误误差均保持在1%以内。实验也验证了RF的不过拟合能力，即随着树的数量的增加，达到足够多时，RF的泛化误差收敛于一个极限值。

（4）根据图2和表3可以看出，当 $Ntree=100$ 时，森林的分类正确率的性能基本可接近最优值。

（5）在本实验中采取的均是数据密度 $D>1$ 即高密度数据进行test error和OOB error分析。从本实验可

表2　实验数据介绍

样本	样本类型	样本数 n	条件属性 a	类别个数 c	数据密度 D
Sonar_lisan	离散	208	60	2	1.13
Ionosphere	连续	351	34	2	1.46
Glass_lisan	离散	214	9	7	1.56
Vehicle_lisan	离散	846	18	4	1.85
Ecoli	连续	336	7	8	1.92
Breast	离散	699	9	2	2.66
Iris	连续	150	4	3	2.82
car	离散	1 728	6	4	3.39
Haberman	连续	306	3	2	4.58

表3　*Ntree* 取值对RF分类精度影响　%

样本	*Ntree*=100		*Ntree*=300		*Ntree*=500		*Ntree*=800		*Ntree*=1 000	
	Test	OOB	Test	OOB	Test	OOB	Test	OOB	Test	OOB
Sonar_lisan	6.86	9.88	6.84	7.34	6.84	6.70	5.53	6.64	5.88	7.44
Glass_lisan	27.96	28.97	29.16	28.71	29.64	28.29	30.02	28.09	30.90	27.78
Vehicle_lisan	26.25	26.93	26.47	26.83	26.01	26.52	26.81	26.29	26.68	26.14
Iris	3.33	4.59	5.33	4.74	5.33	4.74	4.00	4.22	5.33	5.04
Breast	3.18	3.60	3.44	3.42	3.59	3.24	3.25	3.27	3.10	3.27
Haberman	28.11	28.61	28.17	28.69	29.83	28.69	29.72	28.88	28.89	29.19
Car	6.82	6.91	6.83	7.32	7.11	7.08	6.71	7.06	5.59	5.83

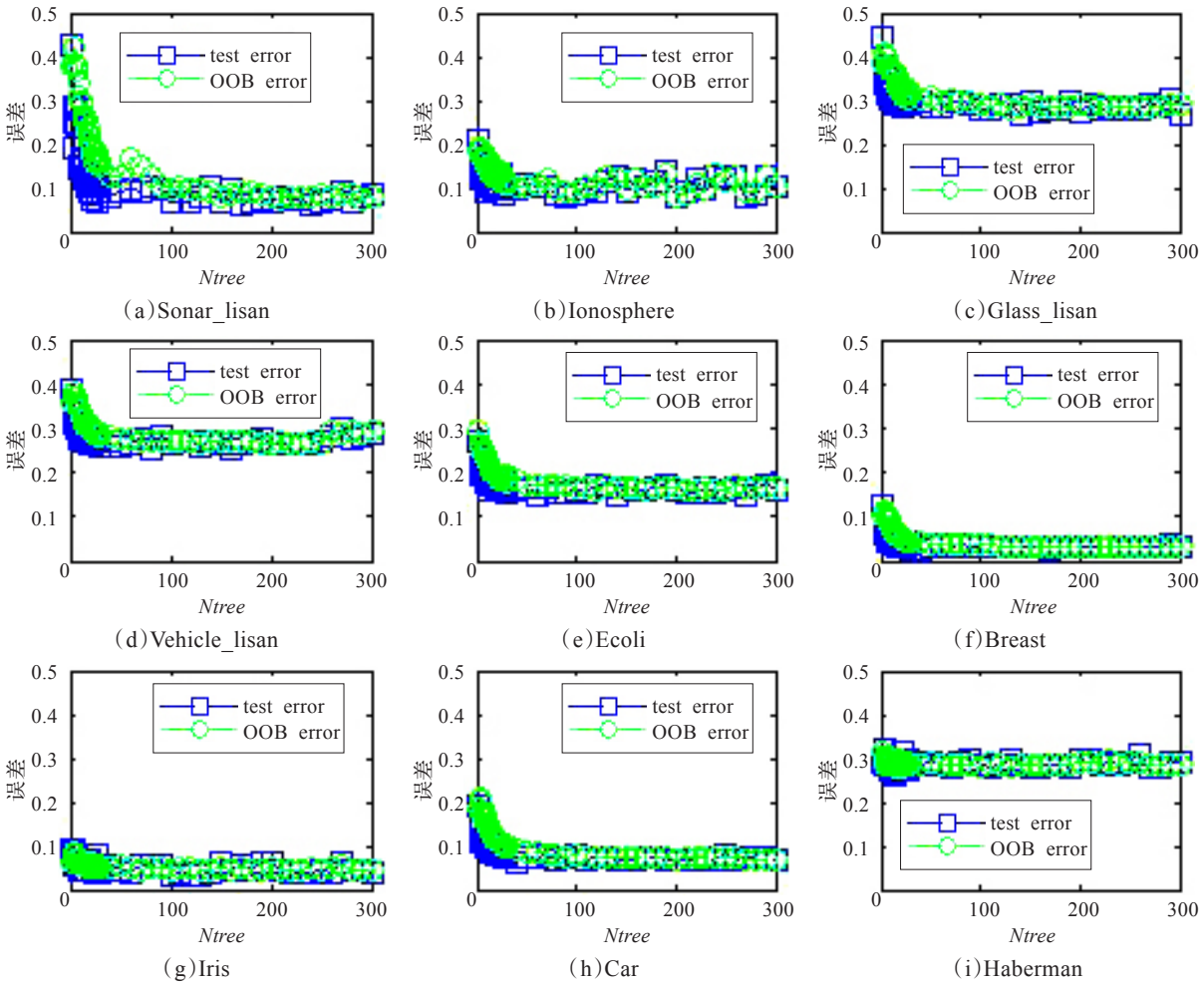

图2　*Ntree* **对** test error **和** OOB error **影响**

以看出,对于高密度数据,数据密度对于 *Ntree* 值影响并不显著。

上一章分析、介绍几种常用的精度估计的方法,支持向量机不存在袋外数据,因此袋外估计方法不适用。因此采用 3-折交叉验证,获取误差估计,对比 RF 和 SVM 的分类器性能。

通过上面的实验结果,采用 *Ntree*=100 获取 *Ntree* 参数下 RF 的接近最优性能。

从表 4 中可以看出,从与分类性能优越的 SVM 进行比较,RF 的分类精度还是比较高的,有的甚至高于 SVM。表 4 从交叉验证的误差比较了不同数据 RF 和 SVM 的分类性能,下面利用 ROC 曲线比较 RF 与 SVM 分类性能。

表4　RF **和** SVM **测试误差比较**

数据集	RF	SVM
Sonar_lisan	6.86	27.54
Ionosphere	7.69	4.31
Glass_lisan	27.69	11.27
Vehicle_lisan	26.25	3.93
Ecoli	14.60	2.70
Breast	3.18	3.03
Iris	3.33	2.08
Car	6.65	7.19
Haberman	28.11	29.70

图 3 采用 ROC 分析,针对 Breast、Car、Haberman 数据,生成了不同数据中某两类的 RF 和 SVM 的(FP,TP)点。其中 RF 中的 *Ntree* 参数采用实验建议值 100。通过对 ROC 的分析,可以通过绘制凸起的曲线外壳比较 RF 和 SVM 的分类性能。针对 car 和 haberman 数据,SVM 的 ROC 点在曲线凸壳下面,此时选取 RF 可以获得更佳性能。针对 Breast 数据,RF 和 SVM 均在 ROC 曲线凸壳上。

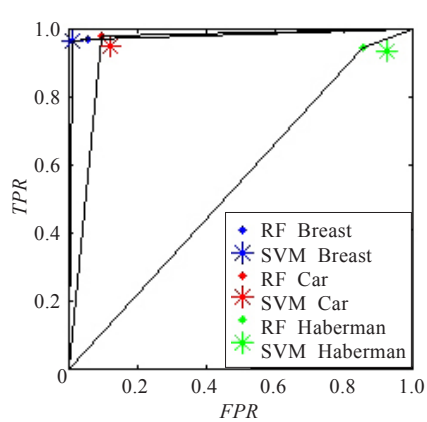

图3　**数据** ROC **曲线**

4　结束语

RF是一种性能优越的集成分类器,其可以克服单分类器对于数据不平衡、重要度不一致等问题。本文首先分析并总结了其性能影响要素以及评价方法。重点研究了RF中树的数量的选取,首先研究了数的数量确定方法,然后通过实验分析的方法,给出了森林分类性能接近最优时树的数量的建议值100。

支持向量机是一种性能优越的单分类器,为了验证RF在*Ntree*为100时的性能,文章从精度和ROC两个方面,比较了RF与支持向量机的性能。实验结果表明,RF的分类性能与支持向量机相当。

参考文献:

[1] Breiman L.Bagging predictors[J].Machine Learning,1996,24(2):123-140.

[2] Ho T.The random subspace method for constructing decision forests[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(8):832-844.

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

[4] Zhang H,Wang M.Search for the smallest random forest[J].Statistics and ITS Interface,2009,2(3).

[5] Díaz-Uriarte R,De Andres S A.Gene selection and classification of microarray data using random forest[J].BMC Bioinformatics,2006,7(1).

[6] Svetnik V,Liaw A,Tong C,et al.Random forest:a classification and regression tool for compound classification and QSAR modeling[J].Journal of Chemical Information and Computer Sciences,2003,43(6):1947-1958.

[7] Oshiro T M,Perez P S,Baranauskas J A.How many trees in a random forest?[M]//Machine learning and data mining in pattern recognition.Berlin Heidelberg:Springer,2012:154-168.

[8] Kulkarni V Y,Sinha P K.Pruning of random forest classifiers:a survey and future directions[C]//2012 International Conference on Data Science & Engineering (ICDSE),2012:64-68.

[9] Dietterich T G.Approximate statistical tests for comparing supervised classification learning algorithms[J].Neural Computation,1998,10(7):1895-1923.

[10] Alpaydm E.Combined 5×2 cv F test for comparing supervised classification learning algorithms[J].Neural Computation,1999,11(8):1885-1892.

[11] Bernard S,Heutte L,Adam S.On the selection of decision trees in random forests[C]//International Joint Conference on Neural Networks,2009:302-307.

[12] Tsymbal A,Pechenizkiy M,Cunningham P.Dynamic integration with random forests[M]//Machine learning:ECML.Berlin Heidelberg:Springer,2006:801-808.

[13] Cunningham P.A taxonomy of similarity mechanisms for case-based reasoning[J].IEEE Transactions on Knowledge and Data Engineering,2009,21(11):1532-1543.

[14] Gatnar E.A diversity measure for tree-based classifier ensembles[M]//Data analysis and decision support.Berlin Heidelberg:Springer,2005:30-38.

[15] Giacinto G,Roli F,Fumera G.Design of effective multiple classifier systems by clustering of classifiers[C]//Proceedings 15th International Conference on Pattern Recognition,2000,2:160-163.

[16] Martínez-Muñoz G,Suárez A.Pruning in ordered bagging ensembles[C]//Proceedings of the 23rd International Conference on Machine Learning,2006:609-616.

[17] Latinne P,Debeir O,Decaestecker C.Limiting the number of trees in random forests[M]//Multiple classifier systems.Berlin Heidelberg:Springer,2001:178-187.

[18] Orrite C,Rodríguez M,Martínez F,et al.Classifier ensemble generation for the majority vote rule[M]//Progress in pattern recognition,image analysis and applications.Berlin Heidelberg:Springer,2008:340-347.

[19] Banfield R E,Hall L O,Bowyer K W,et al.A comparison of decision tree ensemble creation techniques[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(1):173-180.

[20] Tripoliti E E,Fotiadis D I,Manis G.Automated diagnosis of diseases based on classification:dynamic determination of the number of trees in random forests algorithm[J].IEEE Transactions on Information Technology in Biomedicine,2012,16(4):615-622.

[21] Tripoliti E E,Fotiadis D I,Manis G.Dynamic construction of random forests:evaluation using biomedical engineering problems[C]//2010 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB),2010:1-4.

[22] Haralick R M,Sternberg S R,Zhuang X.Image analysis using mathematical morphology[J].IEEE Trans on Pattern Anal Machine Intell,1987,9:532-550.

[23] Kuncheva L I,Whitaker C J.Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J].Machine Learning,2003,51(2):181-207.

[24] Frank A,Asuncion A.UCI machine learning repository[EB/OL].[2013-12-10].http://archive.ics.uci.edu/ml/datasets.html.