Breast Cancer Risk Diagnosis based on Random Forest Classification

Li Li[1,2,3], Yuting Sun[4], Lei Xiao[5,*]

[1]Yunnan Association for Promotion of Trans-Asian Financial Cooperation and Development, Kunming 650092, China

[2]Pan-Asia Business School, Yunnan Normal University, Kunming 650092, China

[3]Champion Property & Casualty Insurance Co., Ltd., Kunming 650228, China

[4]School of Mathematics, Yunnan Normal University, Kunming 650092, China

[5]School of MARXISM Studies, Kunming University, Kunming 650214, China

基于随机森林分类器的乳腺肿瘤风险诊断

李丽[1,2,3]，孙玉婷[4]，肖磊[5,*]

[1]云南省泛亚金融合作发展促进会，昆明 650092，中国

[2]云南师范大学泛亚商学院，昆明 650092，中国

[3]诚泰财产保险股份有限公司，昆明 650228，中国

[4]云南师范大学数学学院，昆明 650092，中国

[5]昆明学院马克思主义学院，昆明 650043，中国

Abstract

In view of the good generalization performance of the random forest classifier, this paper uses the random forest classifier to analyze the risk of the 961 groups of breast tumor lesion tissue digital mammography image data. Empirical results show that the random forest classifier has better generalization performance than Decision Tree, Support Vector Machine and Recent Neighbor Method, and breast tumor severity of influential variable importance is as follows: Margin, Shape, Age and Density.

Key Words: Random Forest Classification; Breast Cancer; OOB Estimation

摘要

鉴于随机森林分类器良好的泛化性能，本文采用随机森林分类器对 961 组乳腺肿瘤病灶组织数字钼靶 X 线摄影图像数据进行了风险诊断分析。结果表明：采用随机森林分类器对乳腺肿瘤进行分类识别能够获得比决策树、支持向量机、最近邻方法更好的泛化性能，同时得到影响乳腺肿瘤严重程度的变量重要性依次为：病灶组织的边缘、 病灶组织的形状、患者年龄、病灶组织的密度。

关键词：随机森林分类器；乳腺肿瘤；OOB 估计

1. 引言

乳腺癌是女性最常见的恶性肿瘤之一，据资料统计，其发病率占全身各种恶性肿瘤的 7-10%。2014 年 ASCO 会议上美国癌症协会公布了 2014 年美国癌症死亡人数、发病率，女性人群中乳腺癌、肺癌和结肠癌最为常见，其中乳腺癌发病率最高（29%），死亡率占第二位（15%）。中国是乳腺癌发病率增长速度最快的国家之一，近年来乳腺癌发病率正以每年 3%的速度递增，已成为城市女性的第一杀手。乳腺癌的早期发现、早期诊断、早期治疗对降低死亡率有着重要意义，因而乳腺肿瘤的诊断

* **通讯作者：**肖磊，email:43063010@qq.com

也就变得尤为重要。

传统的乳腺肿瘤诊断方法是对病灶部位进行医学影像，专家根据经验对影像进行判断，最后对患者的肿瘤种类进行分类。而放射科医生的诊断过程是阅片、判断过程，在此过程中会受到医生个人经验及知识水平的限制和影响。特别是要发现一个病人的细微病灶要面对大量 X 光断层扫描图像，而由于阅片疲劳、个人的判读标准不一等原因，医生诊断时往往容易遗漏某些细微变化。于是人工智能诊断应运而生。由于乳腺病灶组织和正常组织的细胞核显微图像不同，因此根据两种图像不同，采用一种分类能力较强的算法可以进行乳腺肿瘤的诊断。随机森林分类器具有良好的泛化性能，所以本文选用随机森林算法对乳腺肿瘤数据进行分类，从而达到肿瘤诊断的目的。

在对乳腺肿瘤诊断的影像学检查中，主要有钼靶摄影和超声检查。自 20 世纪 70 年代开始，钼靶摄影一直作为诊断乳腺疾病首选的摄影学检查方法。德杰等[1]采用数字化钼靶 X 线对乳腺肿瘤进行诊断。超声检查在乳腺肿瘤诊断中的应用也很广泛, 赵红佳等[2]、周瑜[3]都分别用超声检查中的各种方法或是与其它方法相结合，对乳腺肿瘤进行了诊断。当然还有很多其它的诊断方法，如李佩佩[4]、孙光[5]分别用磁共振成像、微创技术来进行诊断，都取得了较好的成果。

国内外有很多学者对对乳腺肿瘤诊断的影像学数据进行了分析。Jun-Bao Li 等[6]、Eider Sanchez 等[7]、Gisele Helena Barboni Miranda 等[8]、Seral Şahan 等[9]分别采用了判别分析、决策 DNA 建模、计算机辅助系统诊断、K 近邻方法对乳腺肿瘤进行诊断。刘永春[10]等人曾采用过随机森林分类对乳腺肿瘤进行自动诊断，但他们研究的是美国威斯康辛州大学在 1992 年收集到的乳腺癌观测数据，与本文选用的数据在收集方法、数据容量、观测值等方面均有所不同，而且他们仅仅单一地对该数据用随机森林分类来进行处理，并未与其它分类方法进行比较，这些都是他们的文章与本文的区别所在。

在随机森林的理论研究与应用研究方面，吴喜之在《统计学：从数据到结论》[11]、《复杂数据统计方法》[12]中都对随机森林分类做了相关介绍，并说明了随机森林分类自身固有的特点和优良的分类效果。林成德[13]等人利用该方法进行企业信用评估；王志宏[14]等人采用随机森林分类方法对储层岩性进行识别；蒋诗泉[15]等人利用随机森林算法来选择葡萄酒质量评价指标。可见随机森林分类算法在各相关领域都取得了较好的应用。

本文剩余内容结构安排如下：第二部分介绍随机森林算法，包括定义、算法流程、性能指标等内容；第三部分建立模型对 961 组乳腺肿瘤病灶组织数字钼靶 X 线摄影图像数据进行分析，包括模型建立、算法实现、运行结果、对比等内容；最后是全文总结及改进方向。

2. 随机森林介绍

2.1. 随机森林的定义

随机森林算法是由Breiman提出的基于决策树分类器的融合算法，其基本思想是将很多的弱分类器集成为一个强分类器。大量的理论和实证研究都证明了随机森林算法具有较高的预测准确率,对异常值和噪声具有很好的容忍度,且不容易出现过拟合。

定义 1 随机森林是由多个决策树 $\{h(x,\theta_k),k=1,2,\cdots,n\}$ 组成的分类器,其中 $\{\theta_k\}$ 是相互独立且同分布的随机向量。最终由所有决策树综合决定输入向量 x 的最终类标签。

2.2. Bootstrap 法重采样

其核心思想是设集合 S 中含有 n 个不同的样本 $\{x_1,x_2,\cdots,x_n\}$，若每次有放回地从集合 S 中抽取一个样本，一共抽取 n 次,形成新的集合 S^* ,则集合 S^* 中不包含某个样本 x_i(i= 1, 2, n的概率为

$$p=\left(1-\frac{1}{n}\right)^n \tag{1}$$

当 $n\to\infty$ 时，则

$$\lim_{n\to\infty} p=\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n=e^{-1}\approx 0.368 \tag{2}$$

虽然新集合 S^* 的样本总数与原集合 S 的样本总数相等(都为 n)，但新集合 S^* 中可能包含了重复的样本(有放回抽取)，若除去重复的样本，新集合 S^* 中仅包含了原集合 S 中约 $1-e^{-1}=63.2\%$ 的样本。

2.3. 随机森林算法流程

随机森林是基于 Bootstrap 方法重采样，产生多个训练集。设样本的属性个数为 M，m 为大于零且小于 M 的整数。随机森林算法的流程如下:

1) 利用 Bootstrap 方法重采样，随机产生 T 个训练集 $S_1, S_2,\ldots,S_T$;

2) 利用每个训练集，生成对应的决策树 $C_1, C_2,\ldots,C_T$ 在每个非叶子节点(内部节点)上选择属性前，从 M 个属性中随机抽取 m 个属性作为当前节点的分裂属性集，并以这 m 个属性中最好的分裂方式对该节点进行分裂，在整个森林的生长过程中，m 的值维持不变;

3) 每棵树都完整成长，而不进行剪枝。对于测试集样本 X ，利用每个决策树进行测试，得到对应的类别 $C_1(X), C_2(X),\ldots,C_T(X)$;

4) 采用投票的方法，将 T 个决策树中输出最多的类别作为测试集样本 X 所属类别。

即得如下图 1 所示的随机森林算法示意图。

2.4. 随机森林的性能指标

随机森林分类性能受内外两方面因素影响。从外部因素来看，主要来自训练样本的情况，包括训练样本的正负类样本分布，即训练样本的平衡性问题；训练样本的规模，即样本的大小、样本的变量个数及变量的类型。从内部因素来看，主要包括单棵树的分类强度和树之间的相关度。这里我们将重点介绍衡量随机森林性能指标中的分类精度、泛化误差与 OOB 估计。

2.4.1. 分类精度

随机森林算法主要用于分类和预测，那么分类效果就是其主要的考核项目。考虑如下二分类数据的混淆矩阵。

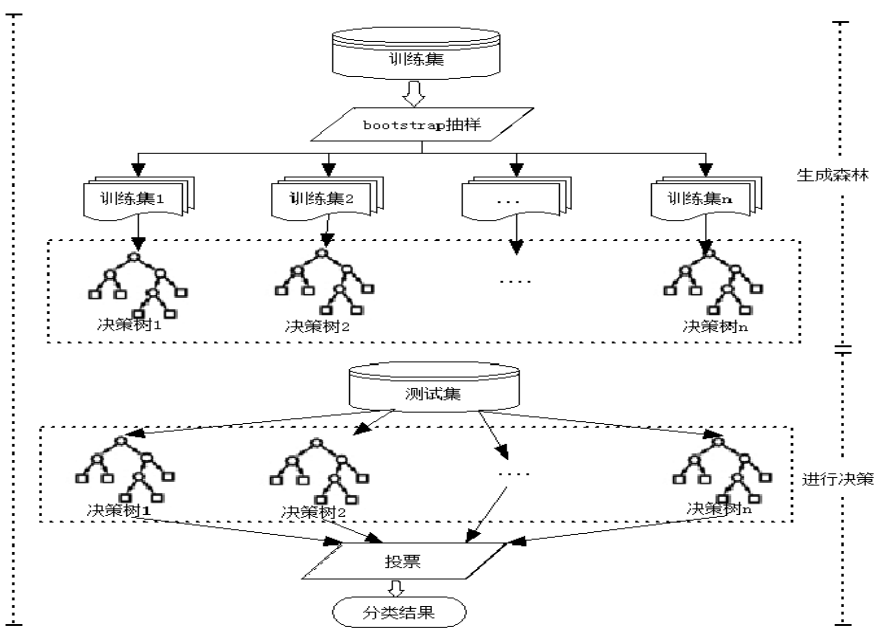

图 1. 随机森林算法示意图

表 1. 二分类数据的混淆矩阵

	Classified positive	Classified negative
Positive	TP	FN
Negative	FP	TN

假设数据集中有两个分类，分别叫正类（positive）和负类（negative），对应表中的列坐标，TP 和 TN 分别代表正确分类的正类和负类的样本数量；FP 和 FN 分别是错误分类的正类和负类的样本数量。

定义 2: 分类精度

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

这个指标是用来衡量随机森林对测试集的总体分类精度，一般而言总体的分类精度越高，算法的分类效果越好。

2.4.2. 泛化误差与 OOB 估计

1）泛化能力（Generalization Ability）

所谓泛化能力是指机器学习算法对新鲜样本的适应能力。机器学习的目的是学到隐含在数据背后的规律，对具有同一规律的学习集以外的数据，经过训练的分类器也能给出合适的输出，该能力称为泛化能力。

2）泛化误差（Generalization Error）

泛化误差是反应泛化能力的一个指标，泛化误差越小，该机器的学习性能越好，反之则性能越差。随机森林的泛化误差从理论上是可以计算出来的。然而在实际环境中，样本的期望输出和分布情况都是未知的，无法直接通过计算泛化误差来评估随机森林的泛化能力。目前，主要有两种方法用于估计分类器的泛化误差，一种是分析模型，而另一种就是交叉验证方法。分析模型很难对随机森林的有效参数个数做出良好估计，因此只能对一些简单的单一性线性分类问题比较有用。交叉验证的方法是利用带标注的验证集来估计泛化误差。在交叉验证中，一般采用 k-fold 交叉验证的方法，如果样本集合比较大，那将需要很大的时间开销。

3）OOB 估计

在对于随机森林算法，比较好的估计泛化误差的方式是 OOB 估计。如前所述，随机森林是利用 Bootstrap 方法重采样进行训练集的生成的，在生成这些数据集时，初始的训练集中有一些样本是不能被抽取的，这些样本的个数是初始数据集的$(1-1/N)^N$（其中 N 为初始训练集中样本的个数）。可以证明，当N 足够大时，$(1-1/N)^N$ 将收敛于 $1/e \approx 0.368$，这个数据说明，有将近37%的样本不会被抽出来，这些不能从初始数据集中抽取出来的样本组成的集合，称之为袋外数据，简记为 OOB。使用 OOB 数据来估计随机森林算法的泛化能力，称之为 OOB 估计。以每一棵决策树为单位, 利用未被该森林选中的所有的训练样本点的集合，统计该树的 OOB 误分率；将所有树的误分率取平均得到随机森林的 OOB 误分率，就可以得到一个 OOB 误差估计。

交叉验证虽然也可以用来估计随机森林的泛化误差，但是交叉验证进行估计时,需要进行数据集的划分和合并处理，使得算法的时间复杂度和空间复杂度会过高，从而使算法的运行效率降低。这一点在使用 OOB 估计时是不会产生的，因为 OOB 数据是在随机森林算法生成的过程中产生的，因此 OOB 估计就可以在每棵决策树生成时同时计算出 OOB 误差率，最终只需要使用很少的资源就可以得到随机森林算法的泛化误差估计。因此，和交叉验证方式相比,OOB 估计的效率是很高的，有大量的实验数据表明，两者的估计效果是差不多的。Breima 通过实验已经证明，OOB 估计是随机森林的泛化误差的一个无偏估计。所以在本文中随机森林算法的性能最终使用 OOB 估计进行评价,OOB 估计越小，算法的性能越好。

3. 基于随机森林分类器的乳腺肿瘤风险诊断

3.1. 模型建立及数据来源

基于随机森林分类器的乳腺肿瘤风险诊断过程的设计思路为: 将乳腺肿瘤病灶组织的数字钼靶 X 线摄影图像的量化特征作为模型的输入变量，良性乳腺肿瘤和恶性乳腺肿瘤作为模型的输出变量。用训练集数据进行随机森林分类器模型的创建，然后对测试集数据进

行仿真测试并对测试结果进行分析。最后将该方法与其它分类方法进行比较。

乳腺肿瘤数据来源于德国埃朗根纽伦堡（Erlangen-Nuremberg）大学放射学研究所建立的乳腺肿瘤病灶组织数字钼靶 X 线摄影图像数据库。数据网址为：http://archive.ics.uci.edu/ml/datasets/Mammographic。数据特征包含了四个量化特征：被诊断者年龄、肿瘤形状、边缘和密度，这些特征都与肿瘤的性质有着密切联系。该数据库共包括 961 个病例数据，其中良性（0）为 516 例，恶性（1）为 445 例。

3.2. 随机森林算法实现

本研究算法实现及图制作均在 R 软件中操作，运用软件包 package(random forest)进行操作。研究类别为两类，输入变量 4 个。

3.3. 运行结果与分析

当树的数量为 200～500 时，分类误差趋于稳定，整体的分类误差也趋于最小化，因此本文将随机森林树的数量定为默认值 500，然后用全部样本进行训练，验证方法采用 10 折交叉验证，画出变量重要性图，得到以下分类结果。

表 2. 二分类的混淆矩阵及结果

	0（良性）	1（恶性）
0（良性）	445	71
1（恶性）	30	415
误判率	0.1050989	

由表 2 中的混淆矩阵可以看出，样本总共有 101 个分错，其中恶性肿瘤有 71 个错分在良性肿瘤中，即良性肿瘤的错分率最高为 0.1376；而良性肿瘤有 30 个错分在恶性肿瘤里，即恶性肿瘤错分率最低为 0.0674；整体误判率为 0.1050989。

由表 3 可以看出，经过随机森林分类 10 折交叉验证，得到乳腺肿瘤数据做分类的训练集和测试集的平均误判率分别为 0.1011692 和 0.1893848。

表 3. 10 折交叉验证结果

训练集平均误判率	0.1011692
测试集平均误判率	0.1893848

表 4. OOB 估计结果

	0	1	误判率
0	409	107	0.2073643
1	70	375	0.1573034
误判率	0.1842		

从表 4 可以看出，在 OOB 估计中，良性肿瘤的误判率为 0.2073643，恶性肿瘤的误判率为 0.1573034，整体误判率为 0.1842，与 10 折交叉验证中测试集的误判率基本相符。在上文已经提到，OOB 估计是随机森林的泛化误差的一个无偏估计，所以本文最终选取 OOB 估计的误判率，即用随机森林分类对乳腺肿瘤数据进行分类时，正确率为 81.58%。

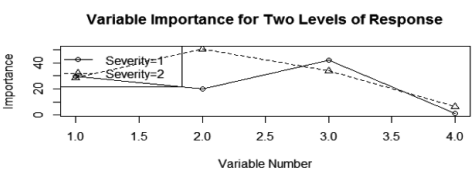

图 2. 变量重要性图

图 2 显示了各个变量对严重程度（Severity）两个水平的相对影响，由此可以看出，诊断为良性肿瘤的变量重要性依次为：病灶组织的边缘（Margin）>患者年龄（Age）> 病灶组织的形状（Shape）> 病灶组织的密度（Density）；诊断为恶性肿瘤的变量重要性依次为：病灶组织的形状（Shape）> 病灶组织的边缘（Margin）> 患者年龄（Age）> 病灶组织的密度（Density）。

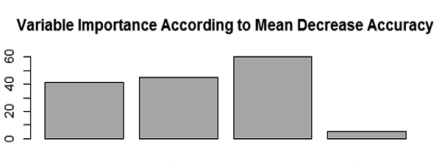

图 3. 平均降低的基尼系数变量重要性图

由图 3 中平均降低的基尼系数变量重要性可知，影响乳腺肿瘤严重程度（Severity）的变量重要性依次为：病灶组织的边缘

（Margin）> 病灶组织的形状（Shape）> 患者年龄（Age）> 病灶组织的密度（Density）。

3.4. 与其它三种分类方法的比较

为比较随机森林分类器与其他分类器的诊断效果，本文还对乳腺肿瘤数据运用决策树、支持向量机、最近邻方法[16]进行了分类，10 折交叉验证结果如表 5 所示。

表 5. 四种分类方法 10 折交叉验证结果

分类方法	训练集误判率	测试集误判率
决策树	0.1817556	0.195755
随机森林	0.1011692	0.1893848
支持向量机	0.1781735	0.1967535
最近邻方法	0.1305372	0.2289615

不难看出，与其它三种方法相比较，随机森林的训练集和测试集的误判率最小，因此运用随机森林分类器进行乳腺肿瘤的风险诊断要优于其它三种方法。

4. 结语

本文采用随机森林分类器对乳腺肿瘤风险进行诊断识别，并与决策树分类、支持向量机分类、最近邻方法分类的结果进行了比较。研究发现随机森林分类器进行乳腺肿瘤风险诊断具有更好的泛化性能，该分类方法的正确率为 81.58%。同时，本文得到影响乳腺肿瘤严重程度（Severity）的变量重要性依次为：病灶组织的边缘（Margin）> 病灶组织的形状（Shape）> 患者年龄（Age） >病灶组织的密度（Density）。这些结论都将对依据乳腺肿瘤病灶组织数字钼靶 X 线摄影图像诊断肿瘤性质提供很好的参考价值。本文只依据乳腺肿瘤病灶组织数字钼靶 X 线摄影图像的数据进行分析，并未对其它检测方法的数据进行分析，针对其它数据，运用随机森林进行分类诊断，其正确率我们无法确定。

Acknowledgements

This study was supported by 2016 year Yunnan Philosophy and Social Science Planning Project (YB2016016), Scientific Research Foundation of Yunnan Provincial Education Department and Doctor Research Foundation of Yunnan Normal University.

致谢

本研究得到了本文受 2016 年度云南省哲学社会科学规划项目（YB2016016）、云南省教育厅科学研究基金资助性项目和云南师范大学博士科研启动项目的资助。

参考文献

[1] 德杰，李彩英，祁永富等.数字钼靶 X 线摄影形态及边缘征象对乳腺良、恶性病变的诊断价值.*临床放射学杂志*，28(10)：1377-1380,2009.

[2] 赵红佳，董宝玮，许荣等.超声造影定量分析在乳腺肿瘤诊断中的应用价值.*中国乳腺病杂志*，2(6)：634-640,2008.

[3] 周瑜.超声弹性成像和"萤火虫"技术在乳腺肿瘤诊断中的应用研究.*大连：大连医科大学*，1-29,2012.

[4] 李佩佩.3.OT 磁共振弥散加权成像在乳腺肿瘤诊断中的应用.*昆明：昆明医科大学*，1-44,2010.

[5] 孙光.微创技术在乳腺肿瘤诊断中的应用价值研究.*吉林大学*，1-24,2007.

[6] Junbao L, Y Peng, D Liu. Quasiconformal kernel common locality discriminant analysis with application to breast cancer diagnosis. *Information Sciences*, 223,2013.

[7] Eider Sanchez, W Peng, C Toro etc. Decisional DNA for modeling and reuse of experiential clinical assessments in breast cancer diagnosis and treatment. *Neurocomputing*, Vol.146, 2014.

[8] Gisele Helena B M, J C Felipe. Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization. *Computers in Biology and Medicine*, 2015.

[9] Seral S, K Polat, H Kodaz, S Gunes. A new hybrid method based on fuzzy-artificial immune system and k algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*, 37 (3): 415-423,2006.

[10] 刘永春，宋弘. 基于随机森林的乳腺肿瘤诊断研究. *电视技术*，38(15): 253-255, 2014.

[11] 吴喜之.统计学：从数据到结论.*北京：中国统计出版社*，123-139,2013.

[12] 吴喜之.复杂数据统计方法--基于 R 的应用.*北京：中国人民大学出版社*，54-68,2012.

[13] 林成德，彭国兰.随机森林在企业信用评估指标体系确定中的应用.*厦门大学学报*，46(2)：199-203,2007.

[14] 王志宏，韩璐，戚磊. 随机森林分类方法在储层岩性识别中的应用[J]. *辽宁工程*

技术大学学报：自然科学版，34(9)：1083-1088,2015.

[15] 蒋诗泉，刘中侠，蒋诗平，周兴才.随机森林算法在红葡萄酒质量评价指标体系选择中的应用.*食品工业科技报*，35(7)：264-267,2014.

[16] Mu Zhang, Zongfang Zhou. A Credit rating model for enterprises based on projection pursuit and k-means clustering algorithm. *Journal of Risk Analysis and Crisis Response*, 2(2):131-138,2012.