

# 基于 Spark 和随机森林的乳腺癌风险预测分析

苗立志<sup>1,2,3</sup>, 刁继尧<sup>4</sup>, 娄冲<sup>4</sup>, 崔进东<sup>4</sup>

(1. 南京邮电大学 地理与生物信息学院, 江苏 南京 210023;

2. 南京邮电大学 江苏省智慧健康大数据分析与服务工程实验室, 江苏 南京 210023;

3. 南京邮电大学 泛在网络健康服务系统教育部工程研究中心, 江苏 南京 210003;

4. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

**摘要:**现代医疗正在朝着智能健康的方向发展。在此大背景下,为了提高乳腺癌风险的发现及预测效果,文中采用大数据分析技术并基于随机森林模型,应用多个弱分类器,将多个决策树获得的结果进行集成,得到疾病发病率;并采用管道学习方法来训练模型,基于该模型开展了致病因素分析以及结果预测。同时,通过皮尔逊相关系数和 Spearman 等级相关系数来进行相关度分析,研究权重较高的影响因子,提高乳腺癌风险的监测和早期预防。实验结果表明,在乳腺癌致病细胞细胞核的相关参数中,Perimeter、Texture 和 Concave points 影响因子对于乳腺癌的致病影响程度较大,更易导致疾病的发生。基于管道训练方法所建立的模型预测精度可达 99.04%,精度高、方法可靠。最终的实验研究结果对于乳腺癌风险的发现具有一定程度的参考意义。

**关键词:**Apache Spark;随机森林;疾病预测;机器学习;智能健康;大数据分析

**中图分类号:**TP311

**文献标识码:**A

**文章编号:**1673-629X(2019)08-0142-05

doi:10.3969/j.issn.1673-629X.2019.08.027

## Breast Cancer Risk Prediction Analysis Based on Apache Spark and Random Forest Algorithm

MIAO Li-zhi<sup>1,2,3</sup>, DIAO Ji-yao<sup>4</sup>, LOU Chong<sup>4</sup>, CUI Jin-dong<sup>4</sup>

(1. School of Geographical and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;

2. Jiangsu Engineering Laboratory for Smart Analysis of Healthy Big Data and Location Based Services, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;

3. Engineering Research Center of Ubiquitous Network Health Service System of Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

4. School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

**Abstract:**Modern medicine is developing towards intelligent health. Under this background, to improve the detection and prediction of breast cancer risk, we use big data analysis and multiple weak classifiers based on random forest model to integrate the results of decision trees to obtain incidence of disease. The pipeline learning method is used to train the model. We also carry out pathogenic factor analysis and result prediction based on the pipeline learning. Meanwhile, the influencing factors with higher weight are studied by Pearson correlation coefficient and Superman rank correlation coefficient, to improve the monitoring risk of breast cancer. The experiment shows that among the relevant parameters of the nucleus of breast cancer pathogenic cells, the Perimeter, Texture and Concave points have a greater impact on the pathogenesis of breast cancer and are more likely to cause the lead to the disease. The prediction accuracy of the model based on the pipeline training method can reach 99.04%, which will provide a certain reference for the discovery of breast cancer risk.

**Key words:**Apache Spark; random forest; disease prediction; machine learning; intelligent health; big data analysis

收稿日期:2018-09-10

修回日期:2019-01-15

网络出版时间:2019-03-27

基金项目:国家自然科学基金(41471329);南京邮电大学国自基金孵化项目(NY218084)

作者简介:苗立志(1981-),男,博士,副教授,硕导,研究方向为位置服务与智慧健康、分布式地理空间信息处理;刁继尧(1993-),男,硕士研究生,研究方向为泛在网络智慧健康。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1624.034.html>

## 0 引言

近年来,随着以深度学习为代表的大数据分析技术的不断发展和成熟,出现了大数据分析技术与医疗健康领域开始深度结合的迹象。科技巨头抢占市场、各路资本大量涌入,癌症成为了热点方向。众多智能健康项目兴起,如 IBM 的沃森人工智能系统研究提高护理水平<sup>[1]</sup>、谷歌风投资金的 1/3 进入医疗健康与生命科学领域<sup>[2-3]</sup>、微软的最新医疗健康项目 Hanover<sup>[3]</sup>等。

随着社会和经济的发展,由于不健康的生活方式和饮食习惯,以及电离辐射等因素,在中国,癌症的健康负担逐年增长,每年超过 160 万人诊断为癌症,12 万因癌症而死亡。与其他大多数国家一样,乳腺癌也成为了中国女性最常见的癌症;每年中国乳腺癌新发数量和死亡数量分别占全世界的 12.2% 和 9.6%。针对这一严重的社会现实,迫切需要开展有关乳腺癌发病风险的研究,包括发病原因分析和基于历史数据进行乳腺癌风险的预测。

现如今大数据分析技术日趋成熟,应用逐渐广泛。大数据分析是在强大的支撑平台上运行分析算法发现隐藏在大数据中潜在价值的过程<sup>[4]</sup>。从异构数据源抽取和集成的数据构成了数据分析的原始数据,而大数据分析的核心问题是如何对这些数据进行有效的表达、解释和学习<sup>[5]</sup>。

大数据分析相关内容包括可视化分析、数据挖掘、预测分析、语义分析及数据质量管理。文中主要采用大数据分析中的预测分析,目前常见的预测方法主要有两类:分析预测法<sup>[6]</sup>和技术预测法<sup>[7]</sup>。国内已有使用分析预测法开展的相关研究,并取得了较好的效果。如徐兵河等<sup>[8]</sup>利用分析预测法,对石蜡包埋组织的基因表达谱,分析预测局部晚期乳腺癌的化疗反应,对 ER 相关基因、细胞增殖及免疫相关基因的表达水平定量分析可预测接受新辅助化疗的女性乳腺癌患者获得 pCR 的可能性。李秀央等<sup>[9]</sup>利用分析预测法探讨流行性乙型脑炎发生率与预测因子的关系,最终得到预测值与实际发生率很接近,仅相差 0.026 4/10 万,准确率为 97.94%。张爱霞等<sup>[10]</sup>利用回归分析预测法对伤亡事故进行了预测,证明了回归分析预测法是一种有效的事故发生趋势预测法。技术预测法是一种通过相关技术进行预测的方法。如林毅超<sup>[11]</sup>利用基于人工神经网络的技术预测法对股价做出预测,选用股市实时指标作为人工神经网络的输入变量,经过循环 13 次训练 96 组数据和预测 1 个股价,结果显示平均预测误差率为 3.4%,绝对偏差在 USD0.27-1.94 之间。王兴旺等<sup>[12]</sup>提出了一种基于多种类型信息计量分析的前沿技术预测方法,通过设定不同权值的计算方式,

获得更为精准的预测结果,并以车联网技术为例进行了实证。

文中采用技术预测法范畴的随机森林算法,构建通过基于 Spark 技术的大数据预测机制;选取乳房部位细胞属性,创建相应的数据集,并提取相应的特征向量来建立分类模型。将数据集分为两部分:70% 作为训练数据训练模型,30% 作为测试数据测试模型;其中采用管道学习方法来训练数据。

## 1 Spark 模型与随机森林算法

### 1.1 Spark 模型

Spark<sup>[13]</sup>是加州大学伯克利分校的 AMP 实验室所开发的集群模式的计算平台,其框架的构建以内存计算为基础。Hadoop 中的计算平台是 MapReduce<sup>[14]</sup>,缺点是运行缓慢,运行程序时需要复制额外的信息序列化和磁盘 I/O,带来的时间和空间开销代价比较大;但适合对离线的任务进行分解。Spark 模型基于内存计算,而且每一个 Job 的执行是基于构建的 stage 有向无环图。Spark 模型运行速度快,且适合进行大规模信息处理。文中利用弹性分布式数据集(resilient distributed dataset, RDD)<sup>[15]</sup>对数据进行相应的操作,选取 Spark Standalone<sup>[16]</sup>集群模式开发,具有较高的容错性和较快的开发速度。

### 1.2 随机森林算法

为了避免单棵决策树容易出现过拟合的现象,并提高预测精度,文中采取了随机森林算法:利用机器学习的集成学习思想,通过构造多个弱分类器最终合成成为一个强分类器,在有效减少过拟合现象的同时,提高预测精度。

随机森林是用多棵决策树对样本进行训练并预测的一种分类器。每个决策树模型  $h(X, \beta_k)$  都有一票投票权来选择最终的分类结果。分类决策公式如下:

$$H(X) = \arg \max_Y \sum_{i=1}^k I(h_i(x) = Y)$$

其中,  $H(X)$  表示随机森林分类结果;  $h_i(x)$  表示单个分类结果;  $Y$  表示分类目标;  $I(\cdot)$  表示示性函数。

该式为随机森林的分类问题,即取各个决策树结果的多数为最终结果。而对于随机森林的回归问题,则可以选取各个决策树结果的期望作为最终结果。

## 2 预测模型构建

### 2.1 建模流程

为实现多个乳腺癌影响因子中,对权重较大的因子,首先需要构建预测模型。具体方法如下:设置  $K$  个弱分类器,使用 Gini 系数<sup>[17]</sup>计算类别纯度,将相似的样本放在同一个弱分类器中,采用 K-means 聚类算

法<sup>[18]</sup>进行训练,并使用均值组合方式。在模型训练完成后,使用另外一组构建好特征的样本,经过模型训练,最后评估模型。

整个建模过程分为两步:训练和测试,如图 1 所示。在训练阶段,主要是根据计算好特征的样本,划分

好  $K$  个弱分类样本后,再进行随机森林训练。训练完成后,测试数据应用训练好的预测模型可得到预测值,将预测值与实际值做运算可得到模型的精度值,进而评估模型的性能。

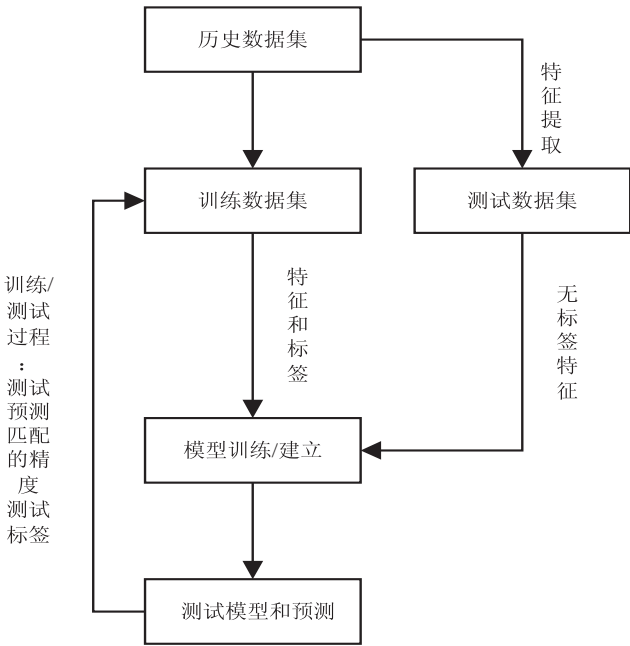


图 1 建模流程

2.2 模型构建

针对乳腺癌发病的多个影响因素展开研究,数据采用威斯康星临床科学中心的相关原始数据( <http://dataju.cn/Dataju/web/datasetInstanceDetail/21> )。该数

据包含了细胞核特征的 10 个属性,主要包括:Radius、Texture、Perimeter 等,如表 1 所示。其中用 fractal dimension 属性值来表示乳腺癌的阴性、阳性。

表 1 细胞核特征变量

序号	属性	示例	说明
1	Radius	5.0	平均半径
2	Texture	1.0	灰度值的标准差
3	Perimeter	1.0	周长
4	Area	1.0	面积
5	Smoothness	2.0	平滑度(半径长度的局部变化)
6	Compactness	1.0	致密性(周长 <sup>2</sup> /面积-1)
7	Concavity	3.0	轮廓凹面部分的严重程度
8	concave points	1.0	轮廓凹面部分的数目
9	symmetry	1.0	是否对称
10	fractal dimension	-1	分形维数

(1) 影响因子特征向量构建。  
数据集中每条样本采用两个类别进行标记:-1 (阴性)和 1 (阳性),每个样本的特征包含如下字段:  
在数据的属性中 fractal dimension<sup>[19]</sup>(分形维数)表示是否患病(-1 或 1)。  
特征:{“radius”,“texture”,“perimeter”,“area”,“smoothness”,“compactness”,“concavity”,“concave

points”,“symmetry”,“fractal dimension”},并基于 VectorAssembler 方法对每个维度的特征做变换,使用 StringIndexer 方法返回 Dataframe,并增加标签列 label。其中数值 1 表示阳性,数值 0 表示阴性,如表 2 所示。  
(2) 训练随机森林分类器。  
按照树训练一个随机森林分类器,主要有以下参数:

表 2 数据文件

	Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity	Concave points	Symmetry	Fractal dimensions	features	label
1	5.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0	-1.0	[5.0,1.0,1.0,...]	0.0
2	5.0	4.0	4.0	5.0	7.0	10.0	3.0	2.0	1.0	-1.0	[5.0,4.0,4.0,...]	0.0
3	3.0	1.0	1.0	1.0	2.0	2.0	3.0	1.0	1.0	-1.0	[3.0,1.0,1.0,...]	0.0
4	6.0	8.0	8.0	1.0	3.0	4.0	3.0	7.0	1.0	-1.0	[6.0,8.0,8.0,...]	0.0
5	4.0	1.0	1.0	3.0	2.0	1.0	3.0	1.0	1.0	-1.0	[4.0,1.0,1.0,...]	0.0
6	8.0	10.0	10.0	8.0	7.0	10.0	9.0	7.0	1.0	1.0	[8.0,10.0,10.0,...]	1.0
7	1.0	1.0	1.0	1.0	2.0	10.0	3.0	1.0	1.0	-1.0	[1.0,1.0,1.0,...]	0.0
8	2.0	1.0	2.0	1.0	2.0	1.0	3.0	1.0	1.0	-1.0	[2.0,1.0,2.0,...]	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...

maxDepth:每棵树的最大深度。增加树的深度可以调高模型的效果,但会延长训练时间。

maxBins:连续特征离散化时选用的最大分桶个数,并且决定每个节点如何分裂。

Impurity:计算信息增益的指标。

auto:在每个节点参与分裂时是否自动选择参与特征的个数。

Seed:随机数生成种子。

文中参数设置为: maxDepth:3; maxBins:20; auto:

“auto”;Seed:5 043。

3 实验分析

3.1 乳腺癌影响因子相关度分析

实验数据集共有 683 条数据,其 fractal dimension 属性值表示乳腺癌的阴/阳性。为了分析影响因素与致病性之间的相关度,选取皮尔逊相关系数<sup>[20]</sup>和 Spearman 等级相关系数<sup>[21]</sup>来分别表征,并将各个属性的相关程度进行排序,如表 3 所示。

表 3 属性相关度排序

系数排序	皮尔逊相关系数结果		Spearman 等级相关系数结果	
1	Perimeter	0.734	Perimeter	0.853
2	Texture	0.733	Texture	0.852
3	Concave points	0.732	Concave points	0.795
4	Radius	0.688	Smoothness	0.767
5	Smoothness	0.679	Compactness	0.732
6	Compactness	0.655	Radius	0.676
7	Area	0.568	Area	0.614
8	Concavity	0.538	Concavity	0.604
9	symmetry	0.413	symmetry	0.507

通过计算各个属性与致病性的相关度,从表 3 可以看出,Perimeter、Texture、Concave points 影响因子对于乳腺癌的影响程度较大。细胞核周长、纹理组织和凹点对于乳腺癌的致病性具有较好的特征表述,将对乳腺癌的检测与发病机制相关研究具有较好的借鉴意义。

3.2 训练预测

文中采用管道学习训练模型,管道在参数网格上不断爬行,自动完成模型优化。用管道训练得到的最优模型进行预测,预测结果有 683 条数据,其中 rawPrediction 是特征和系数的组合值,probability 是每个类别计算出来的概率,prediction 是最终的类分配,

如表 4 所示。将 label 标签值与 prediction 标签值进行比较得到模型的预测精度值是 99.01%,其中包含准确预测条数 677 条。

3.3 结果分析

根据预测结果可以计算预测模型的相关指标值,其中 MSE 实验值为 2.5%,表明预测数据与实际值之间的误差较小;MAE 实验值为 2.5%,表明平均绝对误差较小;RMSE 实验值为 15.9%,表明预测值与原始数据值的误差为 0.159,误差值较小;R-Squared 值为 89.3%,与 1 较接近,表明预测数据与原始数据拟合度较高。



表 4 实验预测结果

	Radius	Texture	Perime- ter	Area	Smoo- thness	Compa- ctness	Conca- vity	Concave points	Symmetry	Fractal dimension	features	label	rawPredi- ction	proba- bility	predi- ction
1	5.0	1.0	1.0	1.0	2.0	1.0	3.0	1.0	1.0	-1.0	[5.0,1.0, 1.0,...]	0.0	19.310	0.965	0.0
2	5.0	4.0	4.0	5.0	7.0	10.0	3.0	2.0	1.0	-1.0	[5.0,4.0, 4.0,...]	0.0	19.918	0.996	0.0
3	3.0	1.0	1.0	1.0	2.0	2.0	3.0	1.0	1.0	-1.0	[3.0,1.0, 1.0,...]	0.0	19.918	0.996	0.0
4	4.0	1.0	1.0	3.0	2.0	1.0	3.0	1.0	1.0	-1.0	[4.0,1.0, 1.0,...]	0.0	19.918	0.996	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6															
8	4.0	8.0	8.0	5.0	4.0	5.0	10	4	1	1	[4.0,8.0, 8.0,...]	1.0	19.938	0.965	1.0
3															

4 结束语

采用基于 Spark 和随机森林算法的机器学习训练方法,研究了管道学习训练预测方法,并将其用于乳腺癌的预测场景,实现分析管道训练模型预测精度值为 99.01%,表明对于乳腺癌的预测有着较高的准确率。同时,通过相关度分析获得了与乳腺癌相关度较高的三个影响因子 Perimeter、Texture、Concave points,可以用来作为乳腺癌预防和防复发的重要指标。通过分析乳房细胞核的特征变量的方法,可以在很大程度上降低医患双方的医疗成本,提高医院的工作效率,具有较高的准确率。

参考文献:

[1] 顾坚磊,江建平,田 园,等. 人工智能技术的应用:罕见病临床决策系统的需求、现状与挑战[J]. 第二军医大学学报,2018,39(8):819-825.

[2] MEGHANA K. Google ventures CEO:our focus on the life sciences will grow[N]. MedCityNews,2015-12-07.

[3] 钟文艳. 美国智能医疗产业发展现状分析[J]. 全球科技经济瞭望,2017,32(6):38-44.

[4] 陶雪娇,胡晓峰,刘 洋. 大数据研究综述[J]. 系统仿真学报,2013,25(S):142-146.

[5] 程学旗,靳小龙,王元卓,等. 大数据系统和分析技术综述[J]. 软件学报,2014,25(9):1889-1908.

[6] 王 琪,张洪伟. 基于 Spark 计算模型的随机森林的电话量预测研究[J]. 成都信息工程学院学报,2015,30(5):445-449.

[7] 陈 蓉. 话务量分析和多种预测模型的比较研究[D]. 北京:北京邮电大学,2008.

[8] 徐兵河,张绪超. 石蜡包埋组织的基因表达谱分析预测局部晚期乳腺癌的化疗反应[J]. 循证医学,2007,7(3):138-140.

[9] 李秀央,陈 坤,赵克勤. 用基于联系数的主因子分析预测

法预测流行性乙型脑炎[J]. 中华流行病学杂志,2005,26(3):218-220.

[10] 张爱霞,朱 明,赵 亮. 用回归分析预测法预测伤亡事故[J]. 河北理工大学学报:自然科学版,2007,29(4):11-13.

[11] 林毅超. 股价变动的神经网络技术预测法研究[D]. 广州:暨南大学,2002.

[12] 王兴旺,董 珏,余婷婷,等. 基于多种类型信息计量分析的前沿技术预测方法研究[J]. 情报杂志,2018,37(10):70-75.

[13] 王 磊,时亚文. 基于 Spark 的大数据计算模型[J]. 电脑知识与技术,2016,12(20):7-8.

[14] 高莉莎,刘正涛,应 毅. 基于应用程序的 MapReduce 性能优化[J]. 计算机技术与发展,2015,25(7):96-99.

[15] 李 星,李 涛. 基于 Spark 的推荐系统的设计与实现[J]. 计算机技术与发展,2018,28(10):194-198.

[16] GENG Yushui,YAN Xianzhao. Spark standalone mode process analysis and data skew solutions[C]//Proceedings of 2017 IEEE 2nd information technology, networking, electronic and automation control conference. [s. l.]:IEEE,2017.

[17] 刘星毅. 一种新的决策树分裂属性选择方法[J]. 计算机技术与发展,2008,18(5):70-72.

[18] 唐 浩,杨余旺,辛智斌. 基于 MapReduce 的单遍 K-means 聚类算法[J]. 计算机技术与发展,2017,27(9):26-30.

[19] LIU Yanyan,GONG Yanming,WANG Xin,et al. Volume fractal dimension of soil particles and relationships with soil physical-chemical properties and plant species diversity in an alpine grassland under different disturbance degrees[J]. Journal of Arid Land,2013,5(4):480-487.

[20] 姜亚斌,邹任玲,刘 建,等. 表面肌电信号的下肢痉挛信号特征分析与识别[J]. 电子科技,2017,30(11):38-41.

[21] 何艳频,孙爱峰. Spearman 等级相关系数计算公式及其相互关系的探讨[J]. 中国现代药物应用,2007,1(7):72-73.