# 基于Fisher线性判别分析对乳腺微钙化性质的预测研究*

汪家清① 张 鑫② 曹 彤③* 王能才③ 张海英③

[文章编号] 1672-8270(2022)02-0005-05 [中图分类号] R318 [文献标识码] A

[摘要] **目的**：使用基于机器学习的Fisher线性分类判别方法，对分割的乳腺微钙化数据进行线性变换，预测乳腺微钙化的性质。**方法**：基于Fisher线性分类判别分析原理，建立预测判别模型对乳腺微钙化的良、恶性进行分类。选取在医院行乳腺癌筛查的432例患者的原始数据，将原始数据中的30项569条乳腺癌特征数据为输入变量，以乳腺微钙化良、恶性的预测准确率为输出变量，建立乳腺微钙化分类判别模型。**结果**：将测试样本代入训练后的Fisher线性判别模型中，其预测乳腺微钙化的良、恶性分类准确率达到93.86%，受试者工作特征(ROC)曲线下面积(AUC)值为0.99，模型的分类性能良好。**结论**：建立的Fisher线性判别模型对乳腺微钙化良、恶性的预测分类效果较好，能够方便快捷地为乳腺疾病的临床诊断起辅助作用。

[关键词] 乳腺微钙化；Fisher线性判别；线性变换；预测分类；机器学习

**DOI: 10.3969/J.ISSN.1672-8270.2022.02.002**

**Prediction study on the nature of breast microcalcification based on Fisher linear discriminant analysis/ WANG Jia-qing, ZHANG Xin, CAO Tong, et al//China Medical Equipment,2022,19(2):5-9.**

**[Abstract] Objective:** To predict the nature of breast microcalcification by linear transformation for the segmented microcalcification data of breast on the basis of using Fish linear classification discriminant method of machine learning. **Methods:** To establish predictive discriminant model for classing benign and malignant microcalcification of breast on the basis of the principle of Fisher linear discriminant analysis. The original data of 432 patients who underwent the screening of breast cancer in hospital were selected, and the 569 characteristic data of breast cancer of 30 items in original data were used as input variable, and the predictive accuracies of benign and malignant microcalcifications of breast were used as output variable in establishing the classification discriminant model of breast microcalcifications. **Results:** After the test samples were put into the trained Fisher linear discriminant model, the accuracy of predicting benign and malignant microcalcifications of breast reached 93.86%, and the area under curve (AUC) value of the receiver operating characteristic (ROC) curve was 0.99. The classification performance of this model was favorable. **Conclusion:** The established Fisher linear discriminant model has a favorable effect on the prediction and classification of benign and malignant microcalcifications of breast, and it can conveniently and quickly play an auxiliary role for the clinical diagnosis of breast diseases.

**[Key words]** Breast microcalcification; Fisher liner discriminant; Linear transformation; Prediction and classification; Machine learning

**[First-author's address]** Specialist Department of Breast, Jiugang Hospital of Jiayuguan City, Jiayuguan 735100, China.

*基金项目：第三批甘肃省民生课题(1303FCMA018)“远程会诊医疗信息服务平台建设与示范应用”

①嘉峪关市酒钢医院乳腺专科 甘肃 嘉峪关 735199
②兰州理工大学电气工程与信息工程学院 甘肃 兰州 730050
③解放军联勤保障部队第九四〇医院信息科 甘肃 兰州 730050
*通信作者：76444058@qq.com
作者简介：汪家清，女，(1972- )，本科学历，副主任医师，从事乳腺疾病的规范化诊疗及乳腺癌的早期筛查诊断及治疗研究工作。

策略研究[J].科技经济导刊,2021,29(13):36-37.
[12] Zhao M,Wang X.A Synthetic Approach for Data center Power Consumption Regulation towards Specific Targets in Smart Grid Environment[J]. Energies,2021,14(9):62-65.
[13] 朱丽,张彦道.工业化数据中心助力行业低碳发展[J].通信世界,2021(9):46-48.
[14] 杨蓉.微软开展数据中心液体冷却实验[J].计算机与网络,2021,47(8):11.
[15] Brodie P,Velkova J.Cloud ruins:Ericsson's Vaudreuil-Dorion data centre and infrastructural abandonment[J].Information Communication and Society,2021,24(2):1-17.
[16] 张鹏.急诊医疗管理信息系统的设计与应用[J].机电信息,2021(12):49-51.
[17] Saiyad A,Patel A,Fulpagare Y,et al.Predictive modeling of thermal parameters inside the raised floor plenum data center using Artificial Neural Networks[J].J Build Engineer,2021,42(1):83-86.
[18] 顾明辰,寇建秋,孙赟,等.医院急诊预检分诊系统接口优化方案设计与实现[J].中国卫生信息管理杂志,2021,18(2):248-252.
[19] 魏喜莲.云数据中心网络安全设备部署研究[J].铁道通信信号,2021,57(4):41-47.
[20] Park SJ,Kim JY,Yoon YH,et al.Analysis of the Adequacy of Prehospital Emergency Medical Services Use of Patients Who Visited Emergency Departments in Korea from 2016 to 2018:Data from the National Emergency Department Information System[J].Emerg Med Int,2021,58(16):142-146.
[21] 杨明川,刘倩,赵继壮.人工智能数据中心研究[J].信息通信技术与政策,2021,47(4):1-7.
[22] 王少伟,孙咸江.医院数据中心建设研究[J].电子世界,2020(21):27-28.
[23] 刘辉兰,马继锋,张海波,等.火神山医院基于私有云的数据中心建设实践与思考[J].中国数字医学,2020,15(5):16-18.
[24] 罗继军,朱明春.医院数据中心(IDC)的规划设计[J].智能建筑,2020(4):37-40.
[25] 马军,闫若玉,王斌,等.基于混合云架构的医院数据中心的建设[J].中国医疗设备,2019,34(1):95-97.
[26] 温煜,唐丹,徐双平,等.基于大数据架构的医院数据中心管理[J].中国数字医学,2018,13(12):15-16,5.

收稿日期：2021-08-18



(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

乳腺癌是在全世界女性中最常见的恶性肿瘤，是导致女性死亡的主要原因之一[1-3]。微钙化可能是乳腺癌的早期症状，通过对乳腺微钙化性质的分析，可以判断其良、恶性，从而为个性化治疗计划和疾病的预后提供重要的依据[4-9]。

机器学习被广泛用于解决预测分类的问题，Fisher线性判别分析是一种经典的线性学习方法，其在人脸识别、市场营销和生物医学研究中都有应用[10-14]。为此，本研究通过构建判别模型对乳腺微钙化的良恶性进行分类，为乳腺疾病的临床诊断提供强有力的辅助手段。

## 1 资料与方法

### 1.1 数据资料

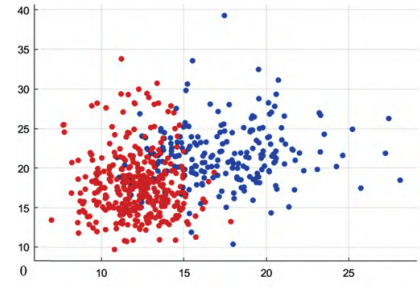
选取2017年10月至2020年5月于甘肃省嘉峪关市酒钢医院乳腺专科行乳腺癌筛查的432例患者的原始数据进行研究，其中女性430例，男性2例；年龄23～80岁，平均年龄(48.25±0.15)岁。所有患者行乳腺超声检查，不能明确乳腺钙化性质的遂行MRI检查，由MRI医师对结果的感兴趣区域(region of interest，ROI)进行分析和提取，最终共获得569条乳腺癌特征数据集。

### 1.2 数据集

在569条乳腺癌特征数据集样本中有肿瘤样本(正样本)与无肿瘤样本(负样本)数目不平衡，其正样本357条，负样本212条。疑似区域的特征指标包括圆度、径向长度的平均值和标准差、灰度熵、灰度均值、灰度标准差、肿块面积、平均分形维数、分形维数标准差、光度惯性动力、各向异性、轮廓梯度熵、平滑度、偏度和峰度及基于灰度共生矩阵的纹理特征等，共30个维度。特征指标的最大值与最小值见表1。

### 1.3 数据分布

表1显示，在特征指标中，面积的方差特别大，在后面的实验中，需要考虑是否进行数据标准化处理，以及加入修正加权阈值算法。把数据集导入到Matlab软件中，可以得到数据的分布情况。乳腺微钙化良恶性数据分布见图1。




注：图中红色为乳腺微钙化良性分布；蓝色为乳腺微钙化恶性分布。

图1 乳腺微钙化良恶性数据分布

## 2 Fisher线性判别分析原理

Fisher线性判别分析常用于解决高维数据的二分类问题，其基本思想是将高维数据点投射到低维空间(如一维直线上)，以解决维数过高引起的“维数灾难”[14-16]。Fisher线性判别分析基本原理见图2。

### 2.1 Fisher线性判别分析法

给定一个数据矩阵$D=[d_1,d_2,\cdots,d_n]^T\in R^{n\times m}$，其中$d_i$是一个$m$维向量，使数据矩阵$(D)$划分为$k$个类为$D^T=[D^T_1,D^T_2,\cdots,D^T_K]$，其中$D_i\in R^{n_i\times m}$且

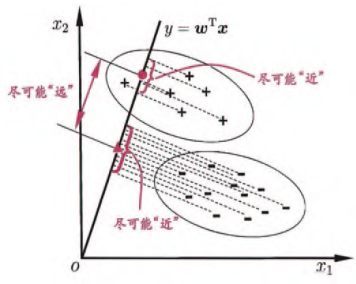
表1 乳腺癌数据特征的30个维度最大值与最小值

| 序号 | 最小值 | 最大值 | 序号 | 最小值 | 最大值 | 序号 | 最小值 | 最大值 |
|---|---|---|---|---|---|---|---|---|
| 1 | 6.981 | 28.11 | 11 | 0.112 | 2.873 | 21 | 7.93 | 36.04 |
| 2 | 9.71 | 39.28 | 12 | 0.36 | 4.885 | 22 | 12.02 | 49.54 |
| 3 | 43.79 | 188.5 | 13 | 0.757 | 21.98 | 23 | 50.41 | 251.2 |
| 4 | 143.5 | 2501 | 14 | 6.802 | 542.2 | 24 | 185.2 | 4254 |
| 5 | 0.053 | 0.163 | 15 | 0.002 | 0.031 | 25 | 0.071 | 0.223 |
| 6 | 0.019 | 0.345 | 16 | 0.002 | 0.135 | 26 | 0.027 | 1.058 |
| 7 | 0.0 | 0.427 | 17 | 0.0 | 0.396 | 27 | 0.0 | 1.252 |
| 8 | 0.0 | 0.201 | 18 | 0.0 | 0.053 | 28 | 0.0 | 0.291 |
| 9 | 0.106 | 0.304 | 19 | 0.008 | 0.079 | 29 | 0.156 | 0.664 |
| 10 | 0.05 | 0.097 | 20 | 0.001 | 0.03 | 30 | 0.055 | 0.208 |

注：表中1～30为特征序号，分别为径向长度、圆度、周长、面积、光滑度、紧凑度、凹度、凹点s、对称性、分型维度，以及以上指标的平均值、标准差和方差。



(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net




注：图中$x_2$为第二类样本的集合；y为线性判别函数。

图2 Fisher线性判别分析原理

$\Sigma^{k}_{i=1}n_i=n$。传统线性判别分析的目标是计算最优线性变换$G\in R^{m\times 1}$，使得原始空间的类结构保留在低维空间中，因此，最优线性变换$(G)$将$m$维空间每个$D$中的向量$d_i$映射到$l$维空间中的向量$y_j$，即$G:d_i\in R^m\rightarrow y_j=G^Td_i\in R(l<m)$。

两个散射矩阵的定义为公式1和公式2：

$$S_b=\frac{1}{n}\Sigma^{k}_{i=1}\Sigma_{d\in D_i}(c_i-c)(c_i-c)^T=\frac{1}{n}\Sigma^{k}_{i=1}n_i(c_i-c)(c_i-c)^T \quad (1)$$

$$S_t=\frac{1}{n}\Sigma^{n}_{i=1}(d_i-c)(d_i-c)^T \quad (2)$$

其中$c_i=\frac{1}{n_i}\Sigma_{d\in D_i}d$为$i^{th}$的平均值，$c=\frac{1}{n}\Sigma_{d\in D}d$为整个数据集的平均值，在低维空间中，作为线性变换的结果，其散射矩阵改变为公式3：

$$S_b^L=G^TS_bG,\ S_t^L=G^TS_tG \quad (3)$$

散射矩阵的计算可以通过使用前体$H_b$和$H_t$来进行简化，其计算为公式4和公式5：

$$H_b=\frac{1}{\sqrt{n}}(\sqrt{n_1}(c_1-c)\cdots\sqrt{n_k}(c_k-c)) \quad (4)$$

$$H_t=\frac{1}{\sqrt{n}}(D^T-ce^T) \quad (5)$$

式中$e=[1,\cdots,1]^T\in R^n$，则散射矩阵$S_b$和$S_t$可以表示为公式6：

$$S_b=H_bH_b^T,\ S_t=H_tH_t^T \quad (6)$$

经判别分析的优化，可以获得最佳变换$G$，其计算为公式7：

$$argmax=\{trace((S_t^L)^{-1}S_b^L)\} \quad (7)$$

### 2.2 基于权重因子改进的Fisher线性判别法

在原判别函数的基础上引入权重因子，把分式型判别函数改为差式判别函数，建立模型使得差值最大化。

由经典判别法得到模型：$max=G^TS_bG/G^TS_tG$，引入权重因子，把权重因子ρ代入模型，于是得到变换后的模型为：$max=\rho G^TS_bG-(1-\rho)G^TS_tG$，其简化后的计算为公式8：

$$max=G^T[\rho(S_b+S_t)-S_t]G \quad (8)$$

可得模型的解$G$为$\rho(S_b+S_t)-S_t$的最大特征值对应的特征向量。取$\rho\in(0,1)$，利用回代正确率来选定ρ的最佳取值，得到最优判别函数。

## 3 Fisher线性判别模型建立

本研究以数据集中30个标签变量作为预测变量输入，取恶性乳腺钙化为1，良性乳腺钙化为2，作为因变量输出。由表1可以看出参数指标面积的方差非常大，在不能确定其对患病影响的情况下，需要考虑是否需要对数据集进行标准化处理。将数据分为训练集与测试集，其中测试集数据为随机抽取总数据集中的20%。利用Python软件对乳腺钙化良、恶性数据进行分析，得到Fisher线性判别模型。Fisher线性判别模型分类器实现步骤见图3。

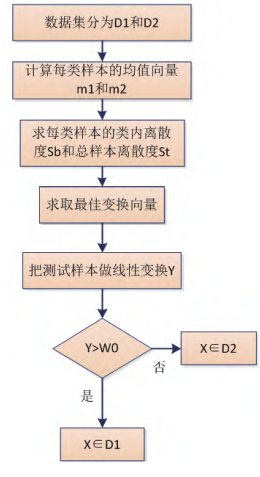



图3 Fisher线性判别模型分类器设计

## 4 结果

### 4.1 参数计算

依据已经划分好的数据集，通过正反标签的训练数据集计算得到均值向量、样本内离散度矩阵及样本间离散度矩阵。利用正反标签的样本内离散度矩阵求得总样本内离散度矩阵，之后利用总样本内离散度矩阵求得被投影向量。根据被投影向量将原特征向量均



(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

投影到一维直线上，采用阈值分割的方法加权修正得到分割阈值的质量。

4.2 模型预测结果

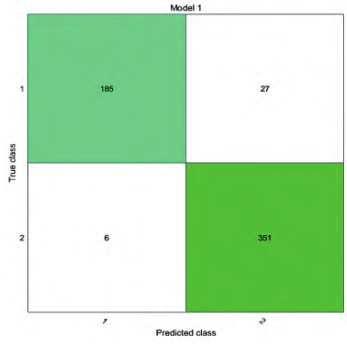
由于无法确定是否要进行数据标准化，首先在实验中加入了数据标准化，并且设置加权阈值观察实验结果。实验的正确预测样本数与准确率的对比结果见表2。

表2 有无修正加权阈值
对正确预测样本的实验结果对比(个)

| 加权阈值 | 分割阈值 | 总样本数 | 正确预测样本数 | 准确率(%) |
|---|---|---|---|---|
| 标准化与修正加权阈值 | 0.0049 | 114 | 71 | 62.28 |
| 标准化与普通加权阈值 | −1.0419 | 114 | 90 | 78.95 |
| 修正加权阈值 | 0.0050 | 114 | 69 | 60.53 |
| 非修正加权阈值 | 3.4498 | 114 | 107 | 93.86 |

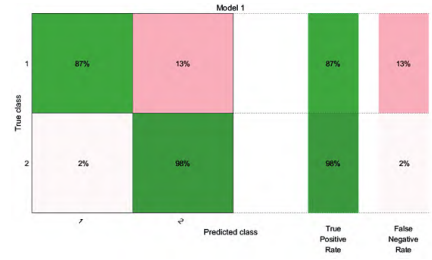
通过不同数据处理方式的对比结果可以看出，将114个测试样本代入训练后的判别模型中，在不采用数据标准化，同时不采用修正加权阈值的情况下，分割阈值为3.4498，此时，分割阈值最大，样本的正确预测数最多，准确率为93.86%。在不采用数据标准化，同时采用修正加权阈值的情况下，分割阈值为0.0050，此时，样本的预测准确率最低，为60.53%。因此，基于此数据集，在Fisher线性判别模型的构建中，不需要对数据进行标准化处理，而添加修正加权阈值函数对预测效果也有负面影响。

在不使用数据标准化和修正加权阈值的情况下，数据预测正确与错误的样本数以及预测正负类率见图4和图5。




注：图中浅绿色为正样本的正确预测样本数；深绿色为负样本的正确预测样本数；白色为正负样本的错误预测样本数。

图4 正负样本的实际结果与预测结果




注：图中绿色为正样本的正确预测率；粉色为正样本的错误预测率；白色为负样本的错误预测率。

图5 正负样本的实际与预测正负类预测率

4.3 模型性能评估与对比

4.3.1 性能评估

采用10倍交叉验证的方法，把数据多次重新划分后代入判别模型进行训练，最后将测试集代入训练好的模型，得到了较好的分类准确率，其最高分类准确率为94.7%，此时采用受试者工作特征(receiver operating characteristic，ROC)曲线和ROC曲线下面积(area under curve，AUC)评估分类器的性能，得到了较好的效果。分类器的ROC曲线见图6。

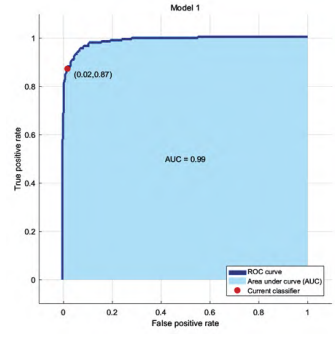



图6 正负样本预测准确率的受试者工作特征曲线

图6显示，ROC曲线覆盖面积较大，AUC为0.99，表明由Fisher线性判别分析建立的模型的分类效果较好，能准确预测乳腺微钙化的良、恶性。当预测分类准确率处于最佳临界值，即假正类率为0.02，真正类率为0.87时，灵敏度和特异度均较高，假正类与假负类也最少，模型性能最优。

4.3.2 性能对比

在乳腺病灶数据处理以及预测分类的研究中，可以有很多不同的方法来创建模型，本研究选择了几种比较常见的算法来进行建模，使用同样的数据进行训练，同样的评估数据集进行验证，确保其公平性和准确性。最终得到的模型性能对比结果见表3。



(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表3 不同模型预测乳腺微钙化良恶性的准确率对比(%)

| 模型 | 错误率 | | 综合准确率 |
|---|---|---|---|
| | 第Ⅰ类 | 第Ⅱ类 | |
| Bayes判别模型 | 13.3 | 8.3 | 88.9 |
| Fisher判别模型 | 12.7 | 1.7 | 94.7 |
| Logistic回归模型 | 20.0 | 12.5 | 83.3 |
| KNN算法模型 | 16.5 | 12.5 | 89.4 |
| SVM分类模型 | 13.6 | 8.5 | 90.2 |
| CART分类回归树模型 | 22.4 | 12.7 | 87.5 |

注：表中KNN为最邻近结点算法；SVM为支持向量机；CART为决策树。

表中算法包括线性算法(Logistic、Fisher)和非线性算法(Bayes、KNN、SVM、CART)，通过结果对比可以得出，基于Fisher线性判别分析建立的预测模型相对于其他算法模型的分类准确率更高，模型性能最优。

## 5 结论

本研究的理论基础为Fisher线性判别分析，选取乳腺数据集中的30个标签变量指标作为输入变量，建立了乳腺微钙化分类的经典判别模型，最后输出变量为乳腺微钙化的良、恶性。实验结果以及模型性能评估对比表明，建立的Fisher线性判别模型对乳腺微钙化分类是可行的，并且准确率高，易于操作，实用性强，对医生的临床诊断有一定的辅助作用。经过实验验证，数据样本指标中的面积对结果无任何影响，所以不需要采用数据标准化以及修正加权阈值算法，并且处理后其他数据指标的特征差异被削弱，影响预测结果。由于数据指标维数过大，在后续实验中应加入主成分分析，来提高分类的效率与准确率。

建立的Fisher线性判别分析模型是建立在已有指标样本的基础上，对于其他样本的预测分类效果可能会有误差。在今后的研究中，还需增加训练数据样本，提高准确性，使乳腺钙化分类模型更加完善。

## 参考文献

[1] Scimeca M,Urbano N,Toschi N,et al.Precision medicine in breast cancer:From biological imaging to artificial intelligence[J].Semin Cancer Biol,2021,72:1-3.

[2] 张倩,宋蕾,侯金香,等.不同分子分型的非特殊型浸润性乳腺癌超声及钼靶特征分析[J].中国医学装备,2020,17(7):70-74.

[3] 潘斌杰,朱剑锋,徐日庆.基于PSO-SVM算法的软土复合固化剂最优配比[J].江苏大学学报(自然科学版),2021,42(3):339-345.

[4] 沈荣波.基于机器学习的乳腺钼靶图像肿块检测技术研究[D].武汉:华中科技大学,2019.

[5] Das A,Nair MS,Peter SD.Kernel-based Fisher discriminant analysis on the Riemannian manifold for nuclear atypia scoring of breast cancer[J].Biocybern Biomed Eng,2019,39(3):728-741.

[6] Suhail Z,Denton ERE,Zwiggelaar R.Classification of micro-calcification in mammograms using scalable linear Fisher discriminant analysis[J].Med Biol Eng Comput,2018,56(8):1475-1485.

[7] 徐琰,胡保全.浅谈人工智能在乳腺癌领域的应用进展[J].中华乳腺病杂志(电子版),2017,11(5):257-261.

[8] 霍双红.基于机器学习的乳腺肿瘤识别[D].太原:中北大学,2017.

[9] 张营,周浩,徐志宾.Fisher判别分析法区分MRI中乳腺病灶性质的研究[J].河南外科学杂志,2017,23(2):42-45.

[10] 李静.基于深度学习的乳腺癌早期诊断研究[D].杭州:杭州电子科技大学,2017.

[11] 李喆.乳腺影像肿瘤诊断中的数据分析[D].天津:天津大学,2017.

[12] 刘倩倩.判别分析和Logistic回归模型在原油和燃料油种类鉴别中的应用研究[D].青岛:中国海洋大学,2013.

[13] 辛芳芳.基于Fisher分类器和计算智能的遥感图像变化检测[D].西安:西安电子科技大学,2011.

[14] 宋光辉,林勇,谢晓燕,等.超声造影对乳腺良恶性肿块Fisher判别的分析研究[J].山东医药,2009,49(44):37-38.

[15] 吉国力,陈舒婷,张延坤,等.逐步回归与判别分析的应用研究——在乳腺疾病建模中的应用[J].厦门理工学院学报,2006(2):22-26.

[16] 张静.乳腺MSCT形态学与功能学检测结果的Fisher判别[D].上海:第二军医大学,2005.

收稿日期：2021-08-15



(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net