复杂高维医学数据挖掘与疾病风险分类研究

林晓丽

北京科技大学

**加密论文编号：\_\_\_\_\_\_\_\_\_\_\_\_\_**

**密　　　　级：\_\_\_\_\_\_\_\_\_\_\_\_\_**

论文题目：复杂高维医学数据挖掘与疾病风险分类研究

s20140842

学　 　号：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

林晓丽

作　 　者：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

信息与通信工程

专 业 名 称：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2016年12月20日

复杂高维医学数据挖掘与疾病风险分类研究

Study on the data mining and risk classification for high-dimensional complex medical data

研究生姓名：林晓丽

指导教师姓名：皇甫伟

北京科技大学计算机与通信工程学院

北京100083，中国

Master Degree Candidate： Xiaoli Lin

Supervisor： Wei Huangfu

School of Computer and Communication Engineering

University of Science and Technology Beijing

30 Xueyuan Road，Haidian District

Beijing 100083，P.R.CHINA

分类号：\_\_\_\_\_\_\_\_\_\_\_\_ 密　　级：\_\_\_\_\_\_\_\_\_\_\_\_\_\_

１０００８

TP391

ＵＤＣ：\_\_\_\_\_\_\_\_\_\_\_\_ 单位代码：\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**北京科技大学硕士学位论文**

复杂高维医学数据挖掘与疾病风险分类研究

**论文题目：**

林晓丽

**作者：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**指 导 教 师： 皇甫伟 单位： 北京科技大学**

**指导小组成员： 单位：**

**单位：**

**论文提交日期：**2016年 12月 20日

**学位授予单位：北 京 科 技 大 学**

致 谢

两年半的研究生生涯即将结束，在研究课题期间有过怀疑、有过否定也收获了独属于自己的成就感，庆幸自己能为乳腺癌课题贡献一点小小的力量，也庆幸自己得到了许多人的帮助和关心。

首先感谢我的导师皇甫伟老师，从论文的开题到终稿，您给予了我无数的指导和鼓励。谢谢您曾抽出宝贵的休息时间为我反复修改论文；谢谢您在我对课题感到迷茫时的鼓励和安慰；还谢谢您在我偷懒时地不断鞭策。很庆幸能成为您的学生，也很骄傲能成为您的学生。

这篇论文的完成还离不开山东大学第二医院的余之刚主任、王斐医生、刘丽媛医生及其他医生们的帮助指导，感谢您们为课题所做的前期努力，同时感谢您们在繁忙的工作之余还能给予论文宝贵的指导建议。此外还要感谢孙冬柏校长、张晓峰老师为了课题项目落实的不断奔波。

感谢我的三位舍友杨露、王月月、邓淑宁，缘分让我们聚在一个宿舍，研究生生涯有你们的陪伴是我的幸运。

感谢杨心竹和边庚磊，我始终记得两年半前我们的第一次见面。谢谢心竹，致我们一起熬过的无数夜晚。谢谢边边——我的健身伙伴和第四位“舍友”。

感谢我的其他实验室小伙伴们：乔喜慧、谢雪琴、张欢、张柏林、段宝升、张弘、文波、李丹阳、邢逸康、阿恒、李佳轩、黄鹤林、王浩彬，有你们的陪伴实验室很温暖。

感谢我们吃喝小分队其他成员：孙晓晖、庄辉、黄赐、谭鑫鑫、王敏、刘娜，你们充实了我的研究生生活，希望毕业后大家还能有机会重聚。

感谢我的家人，爷爷，很抱歉常常忘记打电话给您；奶奶，很抱歉您生病不能陪在身边；爸爸，很抱歉我还常常惹您生气。谢谢您们在我读书期间的支持和鼓励，希望我能成为你们的骄傲。

最后感谢我的男朋友李梓骥，感谢五年半的大学时光有你相伴。

摘 要

以大数据为代表的新一代信息技术浪潮渗透在包括医疗卫生、健康管理在内的诸多领域，有力地改变着传统医学的统计分类方法和思维模式，并可能为人类提供高效准确的数据挖掘和疾病风险分类评估能力。然而，医学数据，尤其是人群队列的医学调研数据，通常观测大量属性，单个样本呈现出个体差异性，使这些医学数据具有高维复杂的特征，其数据挖掘和疾病风险分类在具有重大研究价值的同时，也面临着更大的技术挑战。

论文选取乳腺癌人群队列调研数据为研究对象。乳腺癌是全球女性发病率最高的恶性肿瘤类型，早期的疾病风险分类可以提前介入治疗从而极大地降低发病率和死亡率。欧美国家的乳腺癌风险分类模型不适合用于中国女性，，亟需建立适合中国国情的低漏警率、低成本和易推广的乳腺癌风险分类模型。因此，乳腺癌人群队列数据挖掘和疾病风险分类具有重大而迫切的现实价值，同时在数据类型上也具有典型的代表性。

论文围绕以乳腺癌人群队列数据为代表的复杂高维医学数据，从数据挖掘和疾病风险分类方面的主要工作和创新包括：

（1）针对医学数据的高维性和非平衡性特征，论文提出了单类F-score特征选择法进行特征筛选，并建立了基于单类F-score特征选择法的朴素贝叶斯分类模型。该模型仅用8个特征即获得漏警率9.1%、受试者工作特征曲线(ROC曲线)下面积0.776的效果，优于其它对比算法，且已具备一定的临床指导价值，表明单类F-score特征选择具备良好地应对高维非平衡数据分类的能力。

（2）针对单类F-score特征选择在筛选特征子集时未考虑特征之间相关性的问题，论文提出了改进型的单类F-score诱导的基于遗传算法的特征选择法。实验结果证明改进后的模型ROC曲线下面积达到0.823，具备更好的分类效果。

（3）为了支持疾病风险分类模型在我国巨大人口基数中的推广，论文进一步研究了上述分类算法的显性化和分级的疾病风险显性化评估。前者实现了论文所提出的两类分类算法的闭式解表达，将分类算法中隐式的分类过程以查表计算的方式具象化。后者提出了一类树状结构的分级疾病风险评估框架，以概率的形式给出了分级模型不同分类结果下显示的患病风险量化值。

论文以单类F-score特征选择为基础，建立了一系列医学数据的挖掘和分

类算法，在乳腺癌人群队列数据上的效果已具备临床指导价值，并可以推广到相似疾病的数据挖掘和风险分类。

关键词： 朴素贝叶斯分类，单类F-score特征选择，遗传算法，分级模型

Study on the data mining and risk classification for high-dimensional complex medical data

**Abstract**

A new generation of information technology represented by big data has penetrated into health care, health management and many other fields. It effectively changes the statistical classification method and thinking pattern of traditional medicine, and provides the outstanding capability of data mining and disease risk assessment capabilities for human beings. The medical survey data of the population cohort are complex and high dimensional, which contain a huge number of attributes and individual differences. It is of great significance for data mining and disease risk classification, and facing technical challenges at the same time.

The cohort study data of breast cancer are chosen as the research data. Breast cancer has the highest incidence of all the malignant tumors in women worldwide. The breast cancer risk classification model can help reduce the incidence rate of breast cancer. It is necessary to build an efficient classification model to perform accurate and economical diagnoses. Only the respondents classified into the high risk group need further checks to determine the breast cancer patients. The classification model must have a low false-negative rate, must be low-cost and also can easy to be extended.

The main work and innovations in the field of data mining and disease risk classification of this paper include:

（1）Aiming at the characteristics of high dimension and imbalance of medical data, this paper proposes a one-class F-score feature selection method for feature selection, and establishes a Naive Bayesian classification model based on one-class F-score feature selection method. The experiment results show that, with the presented method, the false-negative rate is decreased to 0.09 and the area under the receiver operating characteristic curve (AUC) is 0.776 with 8 features selected only. Compared with related methods, our method leads to the lowest false-negative rate and the lowest number of features selected and has a certain clinical value. It shows that the one-class F-score feature selection is capable of dealing with high dimensional balance data classification.

（2）This paper proposes an improved one-class F-score induction feature selection based on genetic algorithm. The experimental results of the improved model showed that the AUC reached 0.823 and obtain a better classification effect.

（3）In order to support the promotion of the disease risk classification model in China's huge population base, this paper focus on the closed-form solution of the aforementioned classification algorithm and the explicit assessment of the classification of disease risk in the further studies. On the one hand, this paper proposed a closed-form formulation to describe the classification process. On the other hand, we proposed a method to express the risk probability of illness by tree structure based on probability.

This paper sets up a series of data mining and classification algorithms on medical data base on the one-class F-score feature selection. It has the clinical guidance value on breast cancer and can be extended to data mining and risk classification of similar diseases.

**Key Words： naive Bayesian classifier, one-class F-score feature selection, genetic algorithm, hierarchical model**

**目 录**

[致 谢 I](#_Toc469732748)

[摘 要 III](#_Toc469732749)

[**Abstract** V](#_Toc469732750)

[1 引言 1](#_Toc469732751)

[1.1 课题研究背景及意义 1](#_Toc469732752)

[1.2 研究内容 3](#_Toc469732753)

[1.3 论文组织结构 4](#_Toc469732754)

[2 研究综述 6](#_Toc469732755)

[2.1 疾病风险分类模型的基础知识 6](#_Toc469732756)

[2.1.1 表浅数据与疾病的相关性 6](#_Toc469732757)

[2.1.2 乳腺癌风险分类模型的研究进展 7](#_Toc469732758)

[2.2 非平衡数据分类的研究现状 8](#_Toc469732759)

[2.2.1 数据的非平衡问题 8](#_Toc469732760)

[2.2.2 非平衡数据的分类算法 9](#_Toc469732761)

[2.3 特征选择的研究现状 11](#_Toc469732762)

[2.3.1 特征选择的概述 11](#_Toc469732763)

[2.3.2 特征子集的生成方法 12](#_Toc469732764)

[2.3.3 特征的评价准则 13](#_Toc469732765)

[2.4 本章小结 14](#_Toc469732766)

[3 基于单类F-score特征选择的朴素贝叶斯分类模型 16](#_Toc469732767)

[3.1 数据来源及处理 16](#_Toc469732768)

[3.2 分类模型的评估标准 16](#_Toc469732769)

[3.3 基于单类F-score特征选择的朴素贝叶斯分类 17](#_Toc469732770)

[3.3.1 传统F-score特征选择法 18](#_Toc469732771)

[3.3.2 单类F-score特征选择 19](#_Toc469732772)

[3.3.3 朴素贝叶斯分类 20](#_Toc469732773)

[3.3.4 NBCOF模型的设计流程 21](#_Toc469732774)

[3.4 NBCOF模型实验结果 22](#_Toc469732775)

[3.5 本章小结 25](#_Toc469732776)

[4 基于遗传算法的NBCOF模型 26](#_Toc469732777)

[4.1 遗传算法 26](#_Toc469732778)

[4.1.1 遗传算法的基本原理 26](#_Toc469732779)

[4.1.2 基于遗传算法的特征选择 27](#_Toc469732780)

[4.2 基于遗传算法的单类F-score特征选择 28](#_Toc469732781)

[4.2.1 算法实现过程 28](#_Toc469732782)

[4.2.2 算法性能优化 30](#_Toc469732783)

[4.3 实验结果 33](#_Toc469732784)

[4.4 本章小结 35](#_Toc469732785)

[5 乳腺癌风险分类模型的显式表达及分级分类模型的建立 37](#_Toc469732786)

[5.1 朴素贝叶斯模型的显式表达 37](#_Toc469732787)

[5.1.1 朴素贝叶斯的判别过程 37](#_Toc469732788)

[5.1.2 分类模型显式化表达 39](#_Toc469732789)

[5.2 乳腺癌风险分类分级模型的建立 44](#_Toc469732790)

[5.2.1 分级模型概述 44](#_Toc469732791)

[5.2.2 模型的分级意义与分级结构 45](#_Toc469732792)

[5.2.3 分级模型的实验结果 48](#_Toc469732793)

[5.3 本章小结 50](#_Toc469732794)

[6 总结与展望 52](#_Toc469732795)

[6.1 总结 52](#_Toc469732796)

[6.2 展望 52](#_Toc469732797)

[参考文献 55](#_Toc469732798)

[附录A 乳腺癌调查问卷 61](#_Toc469732799)

[作者简历及在学研究成果 71](#_Toc469732800)

[独创性说明 73](#_Toc469732801)

[关于论文使用授权的说明 73](#_Toc469732802)

[学位论文数据集 75](#_Toc469732803)

1. 引言
   1. 课题研究背景及意义

数据挖掘技术近年来引起了信息产业界的极大关注，随着计算机在医疗系统的广泛应用，医学数据挖掘研究正进入快速增长阶段，将数据挖掘技术应用到医学信息数据库中，可以发现其中的医学诊断规则和模式，辅助医生进行疾病诊断。世界卫生组织曾宣告：“癌症中1/3可以预防，1/3通过早发现、早治疗可以痊愈，剩下的1/3可以通过适当的治疗改善生活质量和延长生命时间。”这意味着疾病发现越早治愈率越高。理论上若直接对筛查者使用专业医疗诊断设备可以快速确诊是否患病。由于我国人口基数大，医疗资源相对缺乏且分布极不均衡，采用大规模的、基于医疗设备检测进行疾病风险分类的手段由于成本高、消耗医疗资源过大、覆盖范围有限而缺乏现阶段的可实施性。因此，在我国医疗资源和医疗投入相对不足的条件下，需要寻找新的手段在减低医疗成本和经济成本的同时实现人群的疾病风险分类。下文以乳腺癌为例，详细叙述建立疾病风险分类模型的意义及要求。

乳腺癌是全球女性发病率最高的恶性肿瘤，对女性的健康和生活质量产生了严重地威胁[1]。据相关研究预测，2015年美国的新发乳腺癌病例为231,840例，乳腺癌致死病例为40,290例[2]。同期中国的新发乳腺癌病例为268,600例，致死病例为69,500例[3]。与欧美发达国家不同，我国人口基数大，乳腺癌发病率持续上升且呈现年轻化趋势，患者确诊时多属于晚期，对中国女性个人、家庭和社会均产生了严重的影响。乳腺癌是可以预防和治愈的，根据长期临床经验，只要做到及早预防、及时检查、及时治疗，乳腺癌是可以预防和治愈的。目前欧美发达国家已建立一套完整的乳腺癌防治措施：开展全人群健康调查——基于危险因素对人群进行乳腺癌风险分类——对诊断为高风险的人群实施干预措施——最终达成恢复健康的目的。如在美国进行推广的Gail-2模型，是基于7个特征包括年龄，初潮年龄，首次妊娠年龄，乳腺活检次数和一级亲属中患乳腺癌的人数建立的[4]。Gail-2模型包含一套调查表和对应的计算公式，使用者填写调查表后模型会根据公式计算调查者的患病风险大小。美国于1988年在全国范围内广泛推广Gail-2模型，对患癌风险高的人群采取药物干预措施。根据调查研究显示，自推广Gail-2模型的十年间，通过对高危人群进行药物干预，美国的乳腺癌发病率由141例/100,000下降为124例/100,000[5,6]。但Gail-2模型的数据来源于白色人种，由于遗传背景、乳房结构、激素水平和生活习惯的差异，模型并不适合对中国女性进行乳腺癌风险分类评估[7]。而上海建立的乳腺癌风险评估模型、南京医科大学研发的改良Gail模型等，因功效有限及地域特征等原因限制了其在社区、基层医疗机构的广泛推广应用。因此，亟待建立符合中国国情的低医疗成本的乳腺癌风险分类模型。鉴于我国医疗资源相对不足且分布不均的情况，本文将寻求建立一个低成本的疾病风险分类模型

在已有医学研究中，个人体征信息、环境信息、遗传信息等信息可以作为在一定程度上反映个体是否患病的依据。这些信息可以在不借助医疗设备或医疗人员参与的情况下获取，具有低成本、易获取的特点。基于我国现有国情，本文将这类数据建立一个低成本的乳腺癌风险分类模型。

研究使用的数据选取山东大学第二医院（以下简称山大二院）于2008年收集的流行病学调查数据。数据中样本根据受访者的医学诊断（钼靶检查、超声波检查及乳房穿刺）结果被分为乳腺癌类和未患病类。调查数据是典型的复杂高维非平衡数据[8]，不仅不同类别（多为两类）间的样本数量比例相差较大，而且样本涉及的特征数较多。传统的医学统计学方法在处理两类样本比例相差过大的数据时会对数据先筛选再分析建模，如医学上的疾病风险分类模型多采用1:2或1:1的病理对照实验，即把实验数据中患病和非患病比例人为地删减为2:1或1:1，但实际生活中大多数疾病的发病率远低于这个比例，基于这种方法建立的分类模型可能在实际应用时存在缺陷。因此针对现有调查数据样本涉及特征数多且两类样本间数量差异过大的问题，为实现低医疗成本和低经济成本的乳腺癌风险分类模型，模型采用如图1-1所示方式建立。

图1-1中分类模型基于现有调查数据建立，模型将受访者分为低风险个体和高风险个体，其中仅高风险个体需要接受后续医学诊断确定是否患有乳腺癌。乳腺癌风险分类模型需要满足以下需求：

（1）低漏警率：即乳腺癌患者被误分类为低风险个体的概率要小。患者被误分类为低风险个体意味着该患者可能错过治疗乳腺癌的最佳时机，这个误分类成本远远高于将健康人误分为高危个体。因此模型的分类结果需要保证具有低漏警率的特点。

（2）低成本：即模型的建立成本和后期推广成本要低。调查数据涉及大量特征，过多的特征将造成建模计算的复杂度增加、数据储存成本的增加以及后期推广模型时经济成本的增加。因此，在保证模型分类效果的同时，需要尽可能的减少特征数的使用。

（3）易推广：乳腺癌风险分类模型建立的初衷是为了在当前有限的医疗资源下降低女性患癌概率和死亡概率。模型的筛查对象不仅是医疗资源相对充裕的大城市人群，还是偏远乡区里缺少医疗设备和医师人员的人群。这类筛查人群往往家庭经济水平较低、获取信息落后，还未完全进入互联网时代。因此模型不仅需要适合在互联网上推广，还需要适合偏远地区的推广。一般的机器学习算法，如神经网络算法[9]，在进行数据分类时需要经过复杂的计算过程，这意味着模型在面向互联网未普及地区推广时，需要借助配套的移动设备或调查人员，无形中增加了筛查成本。因此乳腺癌风险分类模型的设计可类似于Gail-2模型能够以公式化或显式化的方式表达，方便模型的推广和使用。



**图1-1 乳腺癌风险分类模型**

* 1. 研究内容

本课题的研究目的是利用现有流行学调查数据建立一个低漏警率、低成本且易推广的乳腺癌风险分类模型。研究内容从课题研究目的入手，可分为以下几个方面：

（1）为解决高维非平衡数据造成分类器的高漏警率问题，本文从非平衡数据的特征选择着手，提出了一种适用于非平衡数据的单类F-score特征选择方法。通过单类F-score特征选择法筛选出对少数类辨别能力强的特征子集，能有效提高分类器对少数类的分类正确率，降低模型的漏警率。

（2）针对单类F-score特征选择在筛选特征子集时未考虑特征间相关性的问题，实验对单类F-score特征选择进行优化，提出了基于遗传算法[10]的单类F-score特征选择。通过遗传算法优化后的单类F-score特征选择筛选出的特征子集在分类模型中具有更好的类识别能力，进一步降低了分类模型的漏警率。此外，针对遗传算法在迭代过程中耗时大的问题，研究采用建表索引和并行化计算的方法，有效降低了遗传算法的迭代时间，提高了算法的整体性能。

（3）针对模型显示化的要求，本文进一步研究了上述模型中分类算法的显性化，基于贝叶斯判别理论[11]将分类算法中隐式的分类过程以查表计算的方式具象化。此外，论文还研究了分级的疾病风险显性化评估，通过分级中逐级筛掉低风险人群的方式，产生一类树状结构的分级风险评估图，实现分级模型下不同分类结果对应患病风险的“可视化”。

* 1. 论文组织结构

论文的整体安排如下：

第一章作为绪论部分，首先对论文的研究背景及研究意义进行了简要的阐述。之后将研究中面临的主要问题和所做的工作进行简单的梳理，通过系统地归纳总结帮助读者了解论文中面临问题的本质和相对应的创新方法。最后对论文的大体结构进行简略介绍，方便读者了解整篇文章的体系架构。

第二章对课题相关的研究内容进行了详细阐述总结。除了介绍医学上关于疾病风险分类模型尤其是乳腺癌风险分类模型的研究现状外，还详细阐述了数据挖掘领域中非平衡数据分类和特征选择这两方面的研究发展现状，并对非平衡数据分类的常用方法、特征选择的常用算法进行了总结和分析。

第三章首先阐述了关于模型分类效果的评估标准。之后针对非平衡数据分类和特征选择两个问题，本章提出了一种新的特征选择方法：单类F-score特征选择法。并在此基础上建立了一个基于单类F-score特征选择法的朴素贝叶斯分类模型。实验结果显示，单类F-score特征选择法相比其它算法，能更有效地提高朴素贝叶斯分类器的分类效果。

第四章中实验选择遗传算法对单类F-score特征选择法的进一步优化。提出了基于遗传算法的单类F-score特征选择，实验结果显示基于遗传算法优化的特征选择法建立的朴素贝叶斯分类模型具有更好的分类效果。同时针对遗传算法计算开销大、耗费时间长的问题，本章对算法进行了性能优化，最终大大提高了算法的运行速度。

第五章针对模型的易推广要求，深入了解贝叶斯判别理论，详细分析了朴素贝叶斯分类器的分类判决过程，并将抽象模型的分类过程以具体的表格和公式展示，实现了模型的显式化。此外，本文提出了一种“可视化”样本患病风险的方法，通过建立逐级筛减低风险人群的分级模型及其对应的树状结构风险评估图，为进入模型的样本提供不同分类结果下对应的患病风险概率。

第六章对研究工作进行了总结分析，并对下一步的研究方向和计划进行详细地阐述。

1. 研究综述
   1. 疾病风险分类模型的基础知识
      1. 表浅数据与疾病的相关性

表浅数据是指在不借助医疗设备或医疗人员参与的条件下可获取的数据，如调查者的身高、体重、日常行为记录、带遗传性的家族疾病史等。这类数据相比直接的医疗诊断，不能直接的、完整的、确定性的反映个体的健康状态。但已有的医学研究表明，个体的表浅数据与其是否患有某些疾病存在相关性。通过有效的数据挖掘手段，表浅数据可以作为在一定程度上反映个人的健康状况的依据。例如，瑞典的研究表明，吸烟可能导致胰腺癌与肺癌的家族聚集。在研究较成熟的糖尿病风险预测模型中，Herman[12]等在1995年报道了糖尿病危险因素分类树法，将糖尿病可能的危险因素囊括为年龄、腰围、妊娠期糖尿病史、身高、种族(民族)、高血压史、家族史及运动情况（这些特征项均属于表浅数据范围），使用Logistic回归[13]模型分析每个特征与糖尿病前期或糖尿病状况的关系，并利用分类树模型将这些信息归类。最终该风险预测模型实现对美国20—70岁人群糖尿病筛检的灵敏度达78％。特异度达65％[14]。此外，目前被最广泛应用的模型之一的芬兰糖尿病风险评分法，即选取年龄、体质指数、腰围、腰臀围比、血压、糖尿病家族史、饮食和运动习惯等表浅数据设计调查问题，每题都有评分标准，当分数达到一定数值时再让个体进行口服葡萄糖耐量试验，以确认其处于糖尿病期或糖尿病前期。该模型预报10年糖尿病发病风险的敏感度为73％，特异度为83％[15]。此外，国内有学者使用腰围、腰臀围比、体质指数预测等表浅信息预测患糖尿病的风险。

由此可见，个体的表浅数据中隐藏着大量对医学研究有价值的信息。而随着互联网时代的到来，乳腺健康相关的表浅数据信息获取更为方便且可获得数据类型更多，这些表浅数据往往多达成百上千项特征。海量的数据为我们带来更多更准确的信息的同时，也易在数据分析时造成维数灾难。因此如何有效利用海量表浅因素进行数据挖掘工作以期准确预测乳腺癌患病风险，不仅能有效节约医疗资源，还易于在人群中大规模推广。

* + 1. 乳腺癌风险分类模型的研究进展

乳腺癌是女性常见的恶性肿瘤之一，而乳腺癌发现越早，被治愈的可能性就越大，因此如何提高乳腺癌早期检出率是防治该疾病的关键点[16]。针对这一问题，欧美发达国家建立了相应的乳腺癌风险分类模型用于女性的乳腺癌筛查工作，且已在实际运用中取得了一定的成效。传统分类模型采用统计学方法，以表浅数据和基因数据为基础，分析筛者是否患有某种疾病或患病概率大小。下面介绍几种应用较为广泛的用于乳腺癌筛查的模型。

（1）Gail模型

Gail等人基于美国BCDDP（Breast Cancer Demonstration Project）提供的28,478例女性表浅数据，通过病例对照多变量Logistic回归分析建立了Gail模型。模型共纳入5个特征：年龄、初潮年龄、首次妊娠年龄、乳腺活检次数以及以及亲属中患乳腺癌的人数。之后，Anderson等人[17]对模型进行改良建立了Gail-2模型，模型效果优于Gail模型。美国于1988年在全国范围内广泛推广Gail-2模型，对患癌风险高的人群采取药物干预措施。根据调查研究显示，自推广Gail-2模型的十年间，通过对高危人群进行药物干预，美国的乳腺癌发病率由141例/100,000下降为124例/100,000。但是Gail模型基于白色人种数据建立，相关研究显示，Gail模型不适合用于美国黑人女性[18]和亚洲黄色人种[19]的乳腺癌风险分类。

（2）Claus模型

Claus[20]模型的数据来源于美国癌症与类固醇激素的多中心病例对照研究，模型纳入的特征包括：年龄、患乳腺癌的一级和二级亲属人数及发病年龄。模型侧重考虑乳腺癌家族史与女性患乳腺癌的相关性。之后van Asperen等人[21]对Cluas模型进行了修改，将乳腺癌家族史、卵巢癌的家族史和BRCA基因检测结果使用线性回归分析。修改后的Cluas模型筛查效果更优。但与Gail模型类似，Claus模型的数据基于美国白人女性，因此是否适用于其他种族还有待研究。

（3）Tyerr-Cuzick模型

Tyerr-Cuzick模型[22]中数据来源于国际乳腺癌干预研究IBIS（International Breast Intervention Study）,模型使用贝叶斯统计分析方法。模型纳入的特征包括：体重、绝经年龄、激素代替性治疗、乳腺良性基本上、小叶原位癌以及基因BRCA1/2。研究显示，Tyerr-Cuzick模型的准确度高于Gail模型和Claus模型[23]。但Tyerr-Cuzick模型需要获取调查者的基因信息，有别于Gail模型仅需获取调查者的表浅数据，导致该模型的推广程度低于Gail模型。

（4）BRCAPRO模型

BRCAPRO模型[24]是最具代表性的基因型模型。模型纳入的特征包括：基因BRCA1和BRCA2的突变频率、患乳腺癌和卵巢癌的一级和二级亲属人数。模型采用贝叶斯定理对筛查者患癌风险进行评估。但Kang等人[25]的研究发现，BRCAPRO模型并不适合用于韩国女性，即BRCAPRO模型也与其它模型一样，存在种族的差异。

以上模型在各自国家的乳腺癌筛查中提供了有效的参考依据且发挥了积极的作用。而我国乳腺癌预防及筛查工作起步较晚，尚未出现用于乳腺癌筛查的有效风险分类模型。由于遗传背景、乳房结构、激素水平和生活习惯的差异，在美国得到广泛推广的Gail-2模型并不适合用于中国女性乳腺癌风险分类。有研究利用Gail-2模型对山东省23,000例普查对象进行分类，发现ROC曲线下面积仅0.564[26]，远未达到医用标准（一般医用标准下模型的ROC曲线下面积需0.7以上）。同时，国内现阶段的乳腺癌风险分类模型多是基于Gail模型的改良版，因功效、地域特征等因素限制了其在社区、基层医疗机构间的广泛推广应用，且医学上的疾病风险分类模型多采用病理对照1:2或1:1实验，即把实验数据人为地删减为2:1或1:1的比例，这并不符合现实中乳腺癌在中国女性人群中的发病概率，因此亟待建立适合我国国情的乳腺癌风险分类模型。

* 1. 非平衡数据分类的研究现状
     1. 数据的非平衡问题

非平衡数据的处理是实际研究中的一项重要工作。在很多实际问题中，比如网络入侵、文本分类、医疗诊断等，分类处理的对象大多是不平衡数据集，即某些类别的样本要比其他类别的样本数目多[8]。如现实中患有乳腺癌的人群总是极少数的。因此在乳腺癌数据中，划分为乳腺癌的样本数必会远远少于未患病的样本。为便于研究分析，在非平衡数据中，样本数较少的类通常被称为少数类或稀疏类，而样本数较多的类被称为多数类[8]。

传统的分类器大多基于样本数据是平衡的这一假定条件，因此在面对非平衡数据的分类问题时，这些分类器常常是失效的。因为在分类时传统分类器倾向于获取分类结果的最大准确率，这会造成分类器在非平衡数据分类判决时总会倾向于多数类，而忽略少数类[8]。而矛盾的是，少数类的数据中往往携带更多有意义的信息。对于传统的分类模型，如在乳腺癌数据分类中，假设100个样本有1例为乳腺癌患者，剩余99例为普通人。若分类模型将100人全部分类为普通人，则模型的分类准确率为99%，但模型无法有效识别乳腺癌患者，这显然不是乳腺癌分类模型的真正目的并不具备实际意义，只有能对少数的乳腺癌人群进行准确的分类，模型才具有实际使用价值。因此对于非平衡数据的分类而言，如何有效地提高少数类的分类准确率和分类器的整体性能已成为数据挖掘领域的一个热点[27]。

* + 1. 非平衡数据的分类算法

根据当前非平衡数据分类的研究内容，可以将研究分为数据层面、算法层面和集成组合层面三个方向。

（1）数据层面

数据层面的方法是通过数据重采样来改变数据的类分布，使得原本不平衡的数据变得平衡，从而降低数据的不平衡性对分类器的影响从而提高分类效果[28]。数据重采样分为过采样、欠采样或综合两种方式的采样方法[29]。

数据过采样是通过增加少数类的样本数量来降低数据的不平衡性。一般过采样中增加的样本采用直接复制少数类样本的形式，这种方法的优点是能较好的保存原始样本的数据信息，缺点是新增样本数量过多时可能出现分类器过分拟合训练数据的现象[30]。针对这一现象，有人提出相应的改进方法：通过某种方式产生新的合成少数类样本。如SMOTE(synthetic minority over- sampling technique)算法[31]将少数类样本的近邻假设为少数类，采用线性插值的方式产生新样本。这种方法可以有效避免分类器的过拟合问题，但新增样本并非来源于现实数据，基于医学数据的真实性的使用原则，将这种构造少数类的方式应用于乳腺癌数据中不太适合。

数据欠采样是通过减少多数类样本来降低数据的不平衡性。常见的欠采样方法是随机地删减多数类的样本，这种方法的缺点是易造成原始数据中多数类数据的信息丢失而导致分类结果对多数类不利[30]。为避免这一现象有研究选择将远离分类边界或出现重复的样本删掉，这类方法获得的数据在分类器的分类效果中表现更好，但仍避免不了有效信息的丢失情况。

混合采样即将过采样和欠采样两种方式混合使用，如有人提出SMOTE+Tomek Links算法和SMOTE+ENN算法[32]。总的来说数据层面的方法一般从两个方面进行改进：1.采样样本的选择问题；2.采样方式的选择问题。但总体来说，数据层面的方法都存在着两个无法避免的问题：1.样本增加导致分类器过拟合且易引入噪声数据；2.样本减少导致有效信息丢失。

（2）算法层面

算法层面是通过优化现有算法或提出新算法来解决非平衡数据的分类问题。算法层面的研究分为代价敏感学习、单类学习和分类器优化三个方面。

代价敏感学习是处理非平衡数据分类问题的主要方法之一。一般分类算法在测试分类器性能时，不论哪一类的样本被误分类均被记做一个误分类，即错误的代价是等价的。但是在非平衡数据中两类样本被误分的代价并不相等，如误将乳腺癌患者划分为未患癌类的代价是病人被耽误治疗甚至死亡，而健康人群被误分为患癌人群的代价只是进行一次专业的医学诊断而已。代价敏感学习就是为解决这一问题而出现的。所谓代价敏感学习就是针对不同类误分的成本不同，设置不同的惩罚系数，误分的成本越高，对应的惩罚系数越大[33]。如类别A和类别B误分的代价设置为1:4，则一个含有10个A类和3个B类的区域会被判别为B类。这将有助于分类器识别出更多的少数类样本。但代价敏感学习同样存在一些问题：1.惩罚系数的设置没有准确的认知，多数研究是采用经验法或实验测试法获取的；2.易受样本的不平衡比例影响，如比例为1:35和1:100的数据并不适用同一惩罚系数。

单类学习即仅考虑单类（多指少数类）数据的分类问题。其中使用最广泛的是One-class SVM (support vector machines)算法，它的核心思想是使用核函数将某一类的训练数据映射到某一特征空间W，并形成一个尽可能小的封闭曲线。若测试数据落在曲线内则被视为与训练数据属于同一类[34]。单类学习方法在实际生活中有较广泛的应用，如波音公司用Brute系统来监控生成漏洞[35]，Shrink系统使用卫星雷达影像来发现漏洞[36]等。

分类器优化即通过修改分类器自身的参数或函数改变分类器在分类时偏向多数类这一情况。如能对样本所属类别进行评分的朴素贝叶斯分类器，通过调节分类的门限值可以使分类器在判决样本类别时偏向于少数类[37]。通过改变支持向量机（SVM）核函数的方法同样可以解决非平衡数据的分类问 题[38]。此外还有通过修改最近邻算法（kNN）的权重距离函数来弥补数据的非平衡性[39]。

（3）集成组合层面

集成组合层面有两类常用算法，一类是以迭代为核心的Boosting算法，另一类是以多分类判别为核心的分类器组合算法。

Boosting算法属于迭代型算法，算法的主要思路是迭代过程中根据训练结果对被误分的训练样本施以较大的权重，使下一代分类器在分类时更关注这类样本。非平衡数据分类中被误分类的样本往往属于少数类，因此Boosting算法可用来处理非平衡数据的分类问题。Boosting算法不需数据的先验知识，只需要调整算法的迭代次数即可[40]。但Boosting算法并非专门用于解决非平衡数据的分类问题，因此不论是被误分的少数类样本还是多数类样本设置的权重值都是一样的，而随着迭代过程的累积，被误分的多数类样本的权重会变大，这将降低分类器对少数类样本的识别率。

分类器组合法的基本思想是基于训练数据建立一组分类器，使用同一测试数据通过这组分类器可以获得多个分类结果，模型通过投票法判断测试样本属于哪一类别来提高模型分类的准确率[41]。这种方法在面对比例失衡较大的数据分类时，少数类的分类准确率提升有限且会大大降低多数类的分类正确率。

* 1. 特征选择的研究现状
     1. 特征选择的概述

理论上使用的特征越多，分类器的分类能力越好，但是在实际分类时，对于有限的训练数据，过多的特征不仅会降低模型训练的速度，还会因为引入噪声信息导致分类器对训练数据的“过适应”问题[42,43]。因此涌现出大量关于降低样本特征的方法，常用的方法可以分为两类：特征提取和特征选择。

特征提取是将原有特征空间以某种映射方式获得新的特征维数，从而实现降低特征集合维数的目的，如主成分分析(Principle Component Analysis，PCA)方法[44]。但这类方法有一个缺点：经过映射后新特征的理解性变差。但在多数情况下，特征的可理解性具有重要意义，如医学中分类模型使用的特征往往是导致某疾病的强关联因素。此外，数据的特征提取过程是基于全部的原始数据得到的，特征提取这种方法并没有降低数据的收集成本。

特征选择是指从一组特征集合中去除冗余或不相关的特征，选择出最优的特征子集，该子集不仅可以以较低的特征空间维数保留大部分的分类信息还能使分类结果达到最优[45]。即选择的特征子集数量尽量小，且能提高分类器的分类效果。特征选择方法一般分为过滤法和封装法[46]。过滤法不需要模型的学习过程而是从分析训练数据入手，根据特征数据的分析结果判断出是否过滤该特征。封装法则是以模型的学习结果为判断依据进行特征选择。过滤法因为计算简单且不需要经过模型训练等特点被较多研究采用，但过滤法在特征选择中未将特征之间的相互作用纳入考虑范围，相反封装法恰好能解决这一问题。因此大多数研究在特征选择中考虑将两者方法结合使用。

* + 1. 特征子集的生成方法

从特征选择的定义出发，特征选择的基本框架如图2-1所示。图2-1中子集的选择方式一般可以可以分为穷举法、启发法以及随机法[47]三类，具体分类如图2-2所示。



**图2-1 特征选择过程**

（1）穷举法

穷举法指遍历特征空间中所有特征的组合，选取最优特征子集的方法。可进一步将生成的特征子集细分为完备集和非完备集。若设特征集合中特征数量为N，则穷举法的计算复杂度为。穷举法的优点是一定能获取最优的特征子集，但代价是由于穷举法的盲目性，寻优的过程易造成时间的耗费和计算复杂度的增加[48]，所以在处理实际问题时实用性不大。

（2）启发法

启发法是一种具有较强主观性的算法，实际使用中常通过人工设置的规则，采样重复迭代的方式产生一个递增的特征子集。若设特征集合中特征数量为N，则启发法的计算复杂度为。这种方法相较穷举法，计算复杂度得到大大较低，因此实际中该方法得到较为广泛的应用。常见的方法有向前选择、向后选择、组合选择以及基于实例的选择。启发法具有计算简单且训练速度快的优点，但缺点也同样明显，即最终选出的特征子集往往只是较优解而非最优解。

（3）随机法

随机法是指随机生成一组特征子集，按照子集评价标准判断该子集是否符合所选要求。若设特征集合中特征数量为N，则随机法的计算复杂度为。随机法可以分为完全随机法和概率随机法。完全随机是指特征子集的选择是一个随机过程，不存在导向性。概率随机法是指特征子集的选择需要根据已设定的某种方式进行，如根据概率进行选择。常用的随机方法有遗传算法、模拟退火算法[49]、粒子群算法[50]。这类方法涉及参数设置，且参数设置的好坏决定最终是否能得到最优解。与启发法类似，随机法最终选出的特征子集只是一个较优解而不是最优解。



**图2-2 特征选择方法分类**

* + 1. 特征的评价准则

特征选择方法一般分为过滤法和封装法，下面对这两种方法对应的评价准则进行详细阐述。

（1）过滤法的特征评价准则一般采用直接计算特征数据和类之间的相关性。根据计算函数的类型可以将过滤法的评价准则分为：距离度量、一致性度量、信息度量以及依赖性度量。

距离度量是以“距离”的方式来度量特征分辨两类能力的优劣。常用的距离度量方式有欧式距离、平方距离以及Chebychev距离等。如二分类问题中，若特征A在两类中的条件概率差值大于特征B的，则判断特征A好于特征B；若差值为0，则判断特征A与特征B等同。基于距离度量的算法有：Relief及其变种ReliefF[51]，分支定界和BFF[52]等。

一致性度量[53]的核心思想是找到与全集合区分类别能力一样的特征最小子集合。具体方法为已知两个样本，若两个样本特征集合对应的值均相同则称两个样本是一致的，若不同则是不一致的。最终通过特征的不一致率来确定最小特征子集。一致性度量的优点是处理噪声能力强、计算快速且能去除冗余信息。而缺点是对噪声数据过于敏感且限定了特征的取值类型只能是离散型。

信息度量的核心思想是找出与两类或多类分类相关性最强的特征子集且子集间的相关性最小[54]。信息度量一般采用信息增益或互信息。信息增益是通过先验不确定性和特征的后验不确定性之间的差异来筛选特征子集。而互信息是通过判断特征与类别之间的依存关系的强弱来筛选特征子集。常见的信息度量算法有：BIF[55]、MIFS[56]、mRMR[57]、FCBF[58]和CMIM[59]。近几年基于信息熵的特征选择方法是一大热门研究，因为信息熵不要求假定数据分布为已知，不要求特征的数据类型且能够量化特征之间的不确定度和非线性关系。

依赖性度量即判断特征在类别间的可分离性大小。统计上许多系数均可被采用，如Fisher分数、概率误差、线性可判定分析和Pearson相关系数等。如Ding[60]使用F-Statistic相关系数度量mRMR在连续特征与类别和已选特征的相关性。Hall[61]则提出了一种既考虑特征的类区分能力，又考虑特征间冗余大小的度量方法。总的来说依赖性度量采用的具体方法需要视特征的数据类型而定。

（2）封装法有模型学习的过程，因此封装法的特征评价准则一般采用模型的分类错误率大小，其中学习模型选择的是最终被用于分类的模型。相较过滤法的特征评价准则，封装法的优点是选出的特征子集数量更少，用于分类器的效果也更好。而缺点则是在获取特征子集的速度上较慢，时间复杂度较高。

* 1. 本章小结

本章针对乳腺癌风险分类模型的医学背景和关键问题，选取三个具有代表性的方向进行了深入研究，它们分别是疾病风险分类模型的医学基础知识、非平衡数据分类的研究现状和特征选择的研究现状。

以乳腺癌风险分类模型为背景，从医学角度介绍了疾病与表浅数据的关系，并详细阐述了疾病风险分类模型尤其是乳腺癌风险分类模型的研究现状。同时罗列了应用较广泛的几种已有乳腺癌模型，并对几种模型进行了总结分析。最后指出针对现有乳腺癌风险分类模型不适合中国女性筛查的情况下建立适合中国女性乳腺癌筛查模型的重要性。

针对建立模型面临的两个挑战之一——非平衡数据的分类问题，本章首先介绍了非平衡数据的概述以及这种现象对分类器的分类效果产生的影响。同时总结了现有研究中非平衡数据分类的发展现状，此外还对几种常用解决方法进行了详细的阐述并指出各自的优缺点。

针对研究面临的另一个挑战——特征选择问题，本章首先对特征选择进行了概述介绍，其次总结了生成特征子集的几种方式、各自优缺点以及目前的研究方向。最后阐述了特征选择中如何对特征进行评价及其对应的评价准则，为后续章节提出新的特征选择法提供了理论依据和实验可行性。

1. 基于单类F-score特征选择的朴素贝叶斯分类模型
   1. 数据来源及处理

本文使用的数据来源于山东大学第二医院于2008年开展的一项覆盖三省一市的乳腺癌横断面流行病学调查。数据内容除传统基础卫生信息外，还涵盖调查者生活习惯、乳腺癌家族史和各类疾病史等大量信息。调查数据包含超过100,000个样本，每个样本包含100余项特征，如体征数据、饮食习惯数据、乳腺癌家族史等。数据根据医学诊断结果被分为两类，一类为患癌类，另一类为非患类，其中患癌类样本数仅200余人。

首先进行数据预处理，对数据进行清理，剩余109314例样本，每个样本包含156项特征。之后向医学领域专家咨询，删掉不适合用于乳腺癌风险分类模型的特征，最终选定的特征项为103项。特征信息最终包括受访者的体征信息如身高体重指数（BMI）、腰围分层大小，生理信息如初潮年龄、生育年龄、绝经年龄，日常饮食偏好信息如食用豆制品、红肉的频率、是否饮酒，以及日常运动频率信息、生活满意度评分信息和乳腺癌家族史等信息。

经过以上处理后，实验数据包含109314例样本，每个样本包含103项特征。其中确诊的乳腺癌样本为282例，患癌样本数与非患癌样本数比例为1:381。最后对数据进行归一化处理，将诊断为乳腺癌的样本划分为少数类，未患癌样本划分为多数类。数据以7:3的比例随机分为训练数据和测试数据，训练数据用于建立分类模型，测试数据用于对分类模型进行评估。

* 1. 分类模型的评估标准

准确率是分类模型中常用的评估标准之一，但是面对非平衡数据的分类问题时，使用准确率来评价分类性能是不适合的。如非平衡数据中，少数类占数据的1%，剩余99%为多数类。若模型将所有数据识别为多数类，则准确率能达到99%，但属于重要信息的稀有类被完全错分。这样的分类模型虽然准确率很高，但是没有实际价值。因此需要其它的参数作为非平衡数据下分类模型的评估标准。

建立一个混淆矩阵表示两类分类问题，如表3-1所示。假设正类样本表示乳腺癌患者，负类样本表示普通人，TP和TN分别表示被正确分类的患者与普通人数量，而FN和FP分别表示被误分类的两类样本数量。

**表3-1 两类分类的混淆矩阵**

|  |  |  |
| --- | --- | --- |
| 类别 | 预测为正类 | 预测为负类 |
| 实际为正类 | TP | FN |
| 实际为负类 | FP | TN |

分类模型使用以下几种参数进行评估：

1.灵敏度（Sensitivity）

医学上称为灵敏度，一般也可称为查全率或召回率（Recall）[62]。在模型中表示被诊断为患癌的样本中被正确分类为高风险类的样本比例。

sensitivity (%)  (3-1)

2.漏警率（False-negative rate）

在模型中表示被诊断为患癌的样本中被误分类为低风险类的样本比例。

false-negative rate (%)  (3-2)

3.特异性（Specificity）

在模型中表示被诊断为未患癌的样本中被正确分类为低风险类的样本比例。

specificity (%)  (3-3)

4.ROC曲线下面积（AUC）

ROC（Receiver Operating Characteristic）曲线是反映灵敏度和特异性连续变量的综合指标[63]。ROC曲线选取TPR为纵坐标、FPR为横坐标。由于曲线不受样本的分布和比例影响，ROC曲线下面积（AUC）的大小可以作为非平衡数据下模型分类效果的评判标准。曲线下面积越大，则表示分类的效果越好。

 (3-4)

 (3-5)

为使模型的分类效果达到低漏警率这个要求，在实验中，漏警率应越小越好，与之相反的是灵敏度（即1-漏警率）需要越大越好。此外为减少被误分为高危类的普通人群数量，特异性需要尽可能大。

* 1. 基于单类F-score特征选择的朴素贝叶斯分类

本节提出了一个基于单类F-score特征选择的朴素贝叶斯分类器，用于处理多特征下的非平衡数据的分类问题。后续实验中将分类器简称为NBCOF（a naive Bayesian classifier based on the one-class F-score feature selection）。

* + 1. 传统F-score特征选择法

F-score特征选择法属于混合封装式和过滤式的特征选择法，其特征子集的选择方式是基于向后选择的启发式选择。F-score方法用于衡量特征在两类之间分辨能力，在二分类问题上具有简单有效的优点[64]。算法具体描述如下：

设训练样本集为，，将两类数据样本分为正类样本和负类样本，正类样本数表示为，负类样本数表示为，则样本的第个特征的F-score值F(i)为：

 (3-6)

其中，为第个特征在总样本中的平均值，和为第个特征分别在两个类中的平均值，如图3-1所示。而表示第k个正类样本的第i个特征值，则表示第k个负类样本的第i个特征值。公式(3-6)中，分子可以看作两类的近似类间距之和，而分母表示总类内样本的协方差。因此，F-score值越大，即表示类间距离远，类内紧密，则此特征在两类中的分辨能力越强[64]。



**图3-1 二分类中的传统F-score特征选择**

传统的F-score是一种简单直观且有效的特征选择方法，目前已有较为广泛的应用。如B.Sarojini将F-score特征选择法结合SVM分类器用于‖型糖尿病预测，谢娟英将F-score特征选择法进行改进，将其由二分类应用推广至多分类应用。但是这些方法并不适合解决非平衡数据的分类问题。因为当正类样本数量与负类样本数量相差过大时，F-score算法会倾向于选择能识别多数类的特征而忽略少数类。这种现象并不符合分类模型的要求，因此需要对算法进行优化。

* + 1. 单类F-score特征选择

传统F-score特征选择方法在二分类问题上有较好的辨别能力，但是当正负类样本的数据比例处于非平衡状态时，特征选择的子集往往偏向于表达多数类的信息，这种情况将加剧分类器在分类时忽略少数类的现象，导致模型最终的分类效果无法实现低漏警率的目标。因此面对非平衡数据的特征选择问题时，应充分考虑少数类的信息在所选特征子集中的表达问题。受单类SVM算法的启发，本文提出了一种新的特征选择法：单类F-score特征选择法。这种方法的核心思想是：在计算F-score值时偏向于考虑少数类样本的状态，从而筛选出分辨少数类样本能力更好的特征子集，对应的示意图为图3-2。



**图3-2 二分类中的单类F-score特征选择**

针对以上思路，单类F-score特征选择法可以描述如下：

假设训练样本集为，其中，将两类数据样本分为少数类样本和多数类样本，少数类样本数为，多数类样本数为，则样本的第个特征的F-score值F(i)可定义为：

 (3-7)

其中，为第个特征在总样本中的平均值，和为第个特征分别在两个类中的平均值。而表示第k个少数类样本的第i个特征值。公式(3-7)中分子可以视为多数类与少数类之间的类间距离，而分母不考虑多数类，仅表示少数类内样本的协方差。分子越大，表示该特征在多数类和少数类中 的类间距离越大。而分母越小，意味着该特征在少数类内越密。因此特征的F-score值越大，代表该特征在分类时识别出少数类的能力越好。分别使用单类F-score算法和传统F-score算法对特征集合进行处理，两种算法的计算结果如图3-3所示。



**图3-3 两种F-score值的比较**

在图3-3中，横坐标表示各个特征在调查问卷中的序号，纵坐标表示特征的F-score数值大小，为方便观察数据点的变化，纵坐标的F-score值取对数。从图中可以看出，同一特征下，两种F-score方法计算的值并不相同。某些特征的传统F-score值很小，但单类F-score值却很大，意味着这类特征具有较强的分辨少数类的能力。

* + 1. 朴素贝叶斯分类

贝叶斯分类是基于贝叶斯定理的一种分类算法，而朴素贝叶斯分类则是在贝叶斯分类的基础上，假定特征向量的各分量间相对于决策变量是相对独立的[65]。

假设表示事件H发生的概率，表示发生事件H的条件下发生事件X的概率。表示发生事件X的条件下发生事件H的概率。贝叶斯定理可以表示为：

 (3-8)

由此朴素贝叶斯分类可被描述为：

设一个数据样本可用一个N维特征向量表示，用来描述样本中N个属性的N个值。类分类变量C有m个类别，取值设为。因此，给的一个未知类别的样本X，后验概率最高的类即为X的预测类，根据公式(3-8)有：

 (3-9)

公式(3-9)中，可视为常数，则后验概率的最大值求解可转换为求的最大值。根据朴素贝叶斯分类的定义，特征向量的各分量间相对于决策变量是相对独立的。则有：

 (3-10)

因此求解后验概率的最大值可转为求最大值。根据训练数据，有：

 (3-11)

 (3-12)

其中表示训练样本在第i类中的样本数，表示训练样本总数，表示训练样本在第i类中第k个特征下取值为的样本数。

* + 1. NBCOF模型的设计流程

相比传统的F-score特征选择方法，单类F-score特征选择法能提高分类器识别少数类的能力。通过对传统F-score特征选择流程图[66]的改进，图3-4展示了NBCOF的流程设计图，算法的整体流程可描述如下：

* 首先计算每个特征的单类F-score值，并根据数值的大小进行降序排序。
* 完成排序后，选择前h个特征作为特征子集，使用朴素贝叶斯分类器进行建模并记录下分类器的评估结果。
* 按照特征的排列顺序，每次向特征子集中加入一个特征，使用朴素贝叶斯分类器进行建模并依次记录下分类器的评估结果，直到全部特征被选入特征子集中。
* 循环结束后，从各模型的评估结果中选出分类结果最优的模型，该模型的特征子集即为被选出的最优特征子集。



**图3-4 NBCOF设计流程图**

* 1. NBCOF模型实验结果

本文实验均基于RStudio3.2.3与MATLAB R2014a实现，电脑配备CPU主频为3.10GHz，内存8GB。

根据图3-4所示的流程图，将已排序的特征依次放入循环中，不同的特征子集建立不同的分类模型，对各模型的分类效果进行评估，最终评估结果如图3-5所示。



**图3-5 特征子集评估结果**

图3-5中，横坐标表示被选入分类器的特征数量，其中特征被选入前已按照各自的单类F-score 值进行降序排列。随着被选入的特征数量的增加，模型的敏感度、特异性和AUC均在增加，即模型的分类效果越来越好。当h为8时，模型的敏感度、AUC和特异性均达到最大值，其中敏感度为0.902，AUC为0.776，特异性为0.508，如表3-2所示。随着特征数的增加，三条曲线均开始不同程度地下降，这种现象有可能是后续被选入的特征对分类器的分类无用甚至可能干扰了分类器的分类，导致分类效果的下降。

**表3-2 各分类模型评估结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 算法 | 敏感性 | 特异性 | 漏警率 | AUC | 被选特征数 |
| 朴素贝叶斯分类(NB) | 0.764 | 0.509 | 0.236 | 0.730 | 103 |
| NB+传统Fscore | 0.772 | 0.572 | 0.228 | 0.740 | 86 |
| NB+Fscore+SMOTE | 0.571 | 0.549 | 0.429 | 0.565 | 103 |
| NBCOF | 0.902 | 0.508 | 0.098 | 0.776 | 8 |

表3-2中，将NBCOF模型与其它分类模型进行比较。实验中被选择的模型有：朴素贝叶斯分类器、基于传统F-score特征选择的朴素贝叶斯分类器、基于传统F-score特征选择和SMOTE算法的朴素贝叶斯分类器以及NBCOF模型。其中，经实验验证，SMOTE算法选择对少数类以1:5比例进行过采样，对多数类以5:1进行欠采样。从表中数据可以看出，NBCOF模型在五个参数中均表现优异，尤其在敏感性、漏警率以及被选特征数这三个参数上。而传统F-score特征选择法在处理非平衡数据分类问题上并没有明显的优势。

表3-3展示了使用单类F-score特征选择法选出的八个特征：乳头是否破溃、年龄、是否绝经、现在的身体状况、怀孕次数、腰围分层、臀围尺寸以及是否存在乳腺肿块。

**表3-3 特征子集**

|  |  |
| --- | --- |
| 特征 | 单类F-score值 |
| 乳头是否破溃 | 5.115 |
| 年龄 | 0.771 |
| 是否绝经 | 0.354 |
| 现在的身体状况 | 0.238 |
| 怀孕次数 | 0.176 |
| 腰围分层 | 0.159 |
| 臀围尺寸 | 0.158 |
| 是否存在乳腺肿块 | 0.145 |

图3-6为NBCOF分类模型的ROC曲线图。其中横坐标表示FPR，纵坐标表示TPR。美国广泛使用的Gail-2模型，使用与本实验相同的调查数据进行分析，AUC仅0.564[67]，而NBCOF模型达到0.776。这表明NBCOF模型相比Gail-2模型，更适合中国女性的使用。

1

**图3-6 NBCOF模型的ROC曲线**

* 1. 本章小结

乳腺癌调查数据中乳腺癌患者和非乳腺癌患者的样本数存在严重的数据不平衡问题。本章针对这一问题，提出了基于单类F-score特征选择的朴素贝叶斯分类模型（NBCOF模型）。模型将受访人群分为两类，一类为低风险患病类，另一类为高风险患病类，仅高风险人群需接受进一步医疗诊断。由分析结果可知，单类F-score特征选择算法相比传统F-score，不仅其筛选的特征子集数量更少，特征子集所建立的分类模型也具有更低的漏警率。

由于特征子集的生成采取依次选入的形式，未考虑特征之间可能存在的相关性影响，所以模型需要进行进一步优化工作。

1. 基于遗传算法的NBCOF模型
   1. 遗传算法
      1. 遗传算法的基本原理

受达尔文的进化论的启发，遗传算法（Genetic Algorithm，简称GA）的核心思想是“物竞天择，适者生存”。遗传算法是一种启发式的寻优算法，通过计算模型模拟出生物进化的自然选择和变异过程，在模拟过程中从初始种群开始，一代一代地寻求问题最优解，直到循环迭代次数达到预设值或满足收敛条件时停止迭代。遗传算法的流程图如图4-1所示。算法的整个过程可以概括为选择、繁殖以及变异，在寻优问题中，每个可能的最优解都被编码为一个个体，一定数量的个体组成了群体。通过预设的适应度函数（一般即求解最优值的函数）对群体中的每个个体进行评估，通过一定的选择方法，淘汰适应度低的个体，同时将筛选出的适应度高的个体进行繁殖和变异，以上个体组合为一个新的群体。由于新群体中的个体继承了上一代的优势，因此适应度更高。随着一次次的迭代，种群将朝着最优解的方向进化。



**图4-1 遗传算法流程图**

遗传算法模拟了生物进化和遗传的过程，具有不同于传统优化算法的医学优点：

* 算法具有较好的自组织和自适应能力，初始阶段只需对问题的可能解进行编码，而不是问题参数本身。因此遗传算法可用来解决很多非结构化问题，具有一定的智能性。
* 搜索时不要求优化函数可导，也没有函数连续性的要求。
* 算法实现简单，易于推广和使用。
  + 1. 基于遗传算法的特征选择

由于遗传算法具有一定的自适应性和智能性，在特征选择中，遗传算法常用来解决寻找分类模型最优特征子集问题，属于特征选择里的随机选择法。下面详细介绍一下使用遗传算法进行特征选择的过程：

（1）编码

在特征选择中，遗传算法的每个个体代表着一个特征子集。根据特征的编号，每个特征可看作一个编码位，对其进行二进制编码。若第i个编码位为1，则表示第i个特征被选中，若为0 ，即表示特征未被选中。因此，每个个体都是一条长度为特征数量的二进制编码。

（2）种群初始化

种群有多个个体组成，初始化种群指随机产生一组个体。一般来说，为产生不同数量的特征子集，先随机产生个体中含“1”的个数，再随机把“1”放入不同编码位中。

（3）适应度函数

遗传算法在进化中仅以适应度函数为依据，通过判断群体中个体的适应度来进行导向性搜索。特征选择的目的是选择较少的特征使分类器获得更好的分类效果。因此研究中一般选择分类准确率大小作为适应度评估。也有研究使用分类准确率和特征数量这两个参数作为适应度评估标准[68]。

（4）选择、交叉和变异

选择算子是指根据个体的适应度大小，从种群中选出优良的个体的算法。适应度越大的个体，被选中的概率就越大。常用的选择算子有：轮盘式选择算子、具有排名的转盘式选择算子、随机一致选择算子和锦标赛选择算子[69]。

交叉算子是指对两个父代个体的部分编码位进行交换，重新生成两个新个体的算法[10]。交叉的目的是保留种群优良“基因”的同时，在下一代种群中产生新个体，以维持种群的多样性。交叉操作按照一定的概率（交叉概率）在种群中随机选择两个个体完成。根据交叉的编码数目，交叉算子可以分为：单点交叉、多点交叉和均匀交叉。

变异算子是指以一个很小的概率随机改变种群中个体的某些编码位数值。变异能够有效地保证种群的多样性，避免遗传算法在寻找最优解时陷入局部最优的困境[10]。根据变异的编码数目，变异算子可以分为基本位变异、均匀变异、二元变异和高斯变异。

* 1. 基于遗传算法的单类F-score特征选择

文章第三部分介绍了一种适合于多维、非平衡数据分类问题的特征选择法——单类F-score。由实验结果可知，改进算法后筛选出的特征子集建立的分类模型具有更好的分类效果，但同时算法也存在一定的局限性，由于特征子集的生成采取向后依次选入的形式，没有考虑特征之间的相关性。特征子集的选择很有可能并非模型的最优特征子集，因此选择遗传算法对原有特征选择算法进行优化。

* + 1. 算法实现过程

定义：个体采用二进制编码形式，从左至右编码位对应特征编号，“1”表示对应特征被选中，如01001即表示第二、五位特征被选中。

（1）初始种群生成函数

采用单类F-score值作为启发，不同于传统遗传算法，初始种群中个体采用启发式生成。算法的核心思想是：特征的单类F-score值越大，则该特征被选中的概率越大。具体选择过程描述为：

1）计算各特征被选中的概率。

首先计算各特征的单类F-score值，将它们按照F-score值的大小进行降序排列。排位越靠前的特征，被选中的概率越大。这种特征选择的方式类似于选择算子中具有排名的转盘式选择算子。根据已有选择算子的公式进行修改[69]，若设排在第i位的特征被选中的概率为，则有：

 (4-1)

其中，i为特征的排名顺序，n为特征的总数量，a,b是常数，，通常a=1.1。

2）对特征进行选择。

先产生[0,1]范围内的一个随机数r，若概率满足：

 (4-2)

则第i位特征被选择，相应编码位变为1。

3）生成个体

为确保特征子集的多样性，每个个体中“1”的个数应该不完全相同且数目不易过多。设置变量max\_，max\_表示个体中可生成“1”的最大数量。采用随机法在[1,max\_]中选中一个数值num，该数值表示当前个体应根据特征选择算法选出最多num个特征数。采用二进制方法对个体进行编码，被选中特征对应的编码位为1，未选中位补0，编码长度等于特征的总数量。

4）生成初始种群

初始种群设为pop，种群中个体数为m，特征总数量为n。采用迭代m次的方法可生成m个个体，这些个体组成初始种群，而pop可看作一个m\*n的矩阵。

（2）适应度函数

上一章中已说明，在非平衡数据分类中，分类准确率不适合用来评估模型的分类效果。因此实验中选择ROC曲线下面积（AUC）的大小作为适应度的评估标准。算法流程如图4-2所示，具体算法可描述为：将每个个体放入适应度函数中，根据个体中“1”的位置确定特征子集。将包含特征子集的数据按照7:3比例随机分为两份，其中7成数据作为训练数据，用于朴素贝叶斯分类器的建模，剩余3成数据作为测试数据用来计算模型的AUC，即个体的适应度数值。



**图4-2 适应度算法流程图**

（3）选择算子

根据文献，锦标赛选择算子能较快地搜寻到最优解[69]。为减少计算开销实验采用锦标赛选择算子进行个体选择。算法具体描述如下：

在选择时，从种群中随机选出x个个体，这x个个体中适应度最高的个体即为被选中个体，可放入下一代种群中。若个体选择数为y，将上述过程重复y次即可。

（4）交叉算子

交叉算子选择两点均匀交叉方式。算法具体描述如下：

假设交叉概率为，在[0,1]之间随机地产生一个数r，若，则从种群中随机选出两个个体作为父代。首先随机产生两个交叉位和，再在1、2、3之间随机选择一个数字，若选中1，则在上交换；若选中2，则在上交换；若选中3，则在上交换。图4-3展示随机选中数字2时个体的交换方式。



**图4-3 个体交叉示意图**

（5）变异算子

变异算子选择多点变异方式。算法具体描述如下：

假设变异概率为，针对种群中一个个体，随机地在[0,1]之间产生一个数r，若，则个体进行变异。设置变异点数目为A，随机地在[1,103]中选A个数作为变异位置，最后对变异位取反即可。

* + 1. 算法性能优化

遗传算法在寻优过程中个体需要进行较多计算，尤其在计算个体适应度时。而实验中种群的规模一般较大，这意味着遗传算法的寻优需要消耗大量的计算开销和时间开销。为了有效缩短算法的执行时间，本文采用以下方法对算法进行性能优化：

（1）懒惰计算（Lazy Evaluation）

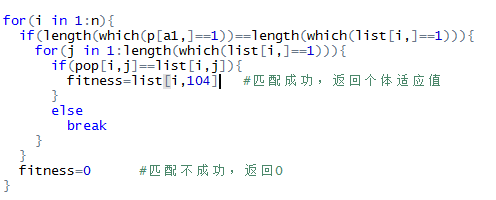
懒惰计算又称为懒性计算，是计算机编程中的一个概念，它的目的是要最小化计算机要做的工作。

遗传算法中每个个体都需要计算对应的适应度数值。而在种群的迭代中，优良个体会被保留下来，这些个体在上一代或上几代时已计算过适应度。若建立一张表并存储每个个体及其对应的适应度，如图4-4所示，则迭代中遇到重复出现的个体时，可以直接在表中进行个体匹配获取对应的适应度数值，而无需重新对个体计算一次。算法如图4-5所示，具体步骤可描述为：

1. 设个体为p[a1,]，其适应度大小为fitness，链表为list[n,m]，首先判断链表是否为空，若否，继续步骤2。
2. 将待求适应度个体与链表中个体进行循环匹配，首先判断两种含“1”个数是否相等，若相等再判断含“1”位置是否相同，若位置相同则fitness为链表中对应个体的适应度数值。若不相同，则匹配链表中下一个个体，直到循环结束仍未匹配成功，则fitness为0。



**图4-4 建立链表**



**图4-5 懒惰计算**

（2）并行计算

遗传算法对个体进行适应度计算时，具有一定的相对独立性，即每个个体理论上可以进行独立地计算，这提供了算法并行计算的可能。所谓并行计算，是指将待解决问题分为若干个独立部分，每部分使用独立地处理器进行计算，各处理器计算完成后再将结果返回给用户，实现多个处理器共同处理同一问题的过程[70]。个体适应度评价的并行化如图4-6所示，并行算法使用R语言中“doParallel”程序包实现，具体步骤为：

1. 获取当前电脑设备可用核数cl.cores：

cl.cores <- detectCores()

1. 创建cl个核构建并行环境：

cl <- makeCluster(cl.cores)

1. 使用%dopar%实现多核的并行计算，同时使用foreach函数对并行结果进行合并整理为fit：

fit <- foreach(i=1:n,.combine='append') %dopar% fitness{

%计算个体适应度的函数

}

上述语句中，n为种群个数；combine为合并方式的设置，因为函数的返回值是单个元素，所以才有元素合并方式——“append”；%dopar%后紧随待并行函数fitness。

1. 并行计算完成后，关闭并行环境

stopCluster(c1)

使用以上两种方法对遗传算法进行优化，并行核数为4，优化前后算法运行时间如图4-7所示。图中横轴表示遗传算法当前正在进行第几次迭代，纵轴表示迭代一次所需的时间，单位为分钟。从图中可以看出，遗传算法未优化时，每迭代一次的时间接近1小时。而算法进行优化后，除计算初始种群适应度时耗时27.18min外，其余每代种群适应度的计算耗时均在15min以下，个别种群甚至仅需4min。这是因为计算初始种群时索引链表为空，优化算法仅依靠并行化减少计算耗时，且实验时并行核数为4，故初始计算适应度时优化算法的耗时接近原算法的1/4（因为合并数据等计算开销，一般会大于1/4）。随着链表开始更新数据，个体的适应度可以通过查询链表获取后，算法的计算耗时会进一步降低，故曲线开始呈下降趋势。且随着迭代次数的增加，链表中储存的个体也会增加，相应地链表查询获取适应度成功概率也会增加，故曲线整体呈下降趋势。



**图4-6 个体适应度计算并行化**



**图4-7 算法运行时间**

* 1. 实验结果

实验参数设置：

* 种群中个体数量设置为50。
* 个体长度为103位，采用二进制编码。
* 初始个体中含“1”个数不超过25个。
* 交叉概率为0.9，变异概率为0.1。
* 终止条件：迭代至第40代。

基于遗传算法的单类F-score特征选择实验结果如图4-8所示。图中横轴表示算法的迭代次数，左边纵轴表示每一代种群中最优个体的适应度大小，即对应的特征子集建立的朴素贝叶斯分类模型的AUC数值。而右边纵轴表示每一代种群中最优个体对应的特征子集合中的特征数。表4-1中展示了特征子集中各特征的详细信息。从图4-8可以看出，随着迭代次数的增加，种群中最优个体的适应度逐渐增大。当迭代次数达到23代以后，最优个体适应度的曲线逐渐趋于平稳，最终在第33代时达到最大值。与之对应的表示特征子集数的曲线则没有明显的变化规律，当种群中最优个体适应度达到最大值时，特征子集数为22。

**表4-1 最优特征子集**

|  |  |  |
| --- | --- | --- |
| 特征 | 排序编号 | 单类F-score值 |
| 年龄 | 2 | 0.771 |
| 现在的身体状况 | 4 | 0.238 |
| 怀孕次数 | 5 | 0.176 |
| 臀围尺 | 7 | 0.158 |
| 是否存在乳腺肿块 | 8 | 0.145 |
| 体重 | 15 | 0.070 |
| 睡眠满意度 | 17 | 0.041 |
| 1年后生活满意度期望 | 26 | 0.031 |
| 锻炼频度 | 29 | 0.026 |
| 是否患糖尿病 | 30 | 0.025 |
| 是否为急性子 | 45 | 0.017 |
| 是否患冠心病 | 55 | 0.010 |
| 是否乳头溢液 | 56 | 0.010 |
| 平均睡眠时间 | 63 | 0.008 |
| 五年前生活满意度 | 65 | 0.008 |
| 新鲜豆类食用频率 | 67 | 0.007 |
| 是否痛经 | 71 | 0.006 |
| 三级亲属患乳腺癌人数 | 87 | 0.001 |
| 是否吸烟 | 92 | 0.0005 |
| 户籍 | 96 | 0.0004 |
| 身高 | 102 | 0.00004 |



**图4-8 基于遗传算法的单类F-score特征选择的实验结果**

表4-2重新整理了各模型的分类效果，从表中可以看出采用遗传算法优化后的NBCOF模型的敏感性为0.910，特异性为0.542，漏警率为0.09，模型的AUC达到0.823。分类效果明显好于其它模型，而模型对应的特征子集数仅次于NBCOF模型。由此可以证实，采用遗传算法优化后的NBCOF模型，弥补了单类F-score特征选择中未考虑特征间相关性的问题。新的特征选择方法能帮助模型获得更好的分类效果。

**表4-2 各分类模型评估结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 算法 | 敏感性 | 特异性 | 漏警率 | AUC | 被选特征数 |
| 朴素贝叶斯分类(NB) | 0.764 | 0.509 | 0.236 | 0.730 | 103 |
| NB+传统Fscore | 0.772 | 0.572 | 0.228 | 0.740 | 86 |
| NB+Fscore+SMOTE | 0.571 | 0.549 | 0.429 | 0.565 | 103 |
| NBCOF | 0.902 | 0.508 | 0.098 | 0.776 | 8 |
| 基于遗传算法的NBCOF | 0.910 | 0.542 | 0.09 | 0.823 | 22 |

* 1. 本章小结

本章针对上一章提出的NBCOF模型中单类F-score特征选择法的不足，选择遗传算法加以改进，优化了特征子集的选择方式。模型中采用的遗传算法基于单类F-score改进，其中个体生成的算法部分采用启发式生成法，即单类F-score值越大，对于特征被选入特征子集的可能性就越大。这种启发式的生成个体方法能加快遗传算法的寻优速度。而实验结果证明，改进后的遗传算法的实验结果明显优于传统遗传算法。同时实验还建立了基于遗传算法的NBCOF模型，优化后的模型在分类效果上比其它模型更优，且仅需要22个特征数。

1. 乳腺癌风险分类模型的显式表达及分级分类模型的建立

为实现模型的另一目标——易推广，本章最后将对已有分类模型进行分析，实现模型的显式表达。此外，本章提出了乳腺癌风险分级分类模型，模型可以显示化的评估不同分类结果下个体患病风险的大小。

* 1. 朴素贝叶斯模型的显式表达

本文选择朴素贝叶斯分类器作为分类模型，其中一个重要原因是模型易于显式化。显式化即将一种隐式的结果以一种便于理解的形式进行表达。模型的显示化可以理解为：模型的分类过程可以用一组简单的公式或表格来展示。寻求模型显示化的目的是为了便于后期模型的推广工作，尤其是偏远落后地区的推广。因为这些地区经济发展较弱且中老年人居多，纸质手册自筛查的方式往往比一个互联网调查问卷普及更广，因此推广时需要一个易于显式化的模型。而机器学习中大多数分类器并不符合这一要求。如果将这些分类器进行数据分类时看作一个黑盒，即输入数据输出分类结果，那么中间的分类过程由于大多数分类算法结构复杂往往难以用简单的方式还原。而朴素贝叶斯分类器属于线性分类器，它的分类过程易于显示化，因此本文最终选择朴素贝叶斯分类器作为基础分类模型。

* + 1. 朴素贝叶斯的判别过程

朴素贝叶斯分类器是一种基于假设独立的贝叶斯定理的简单概率分类器。根据第三章关于朴素贝叶斯模型的介绍可知，朴素贝叶斯分类器在分类时认为数据的各特征是相互独立的。本节将在第三章知识的基础上，介绍朴素贝叶斯分类器是如何进行分类判决的。

理论上，概率分类器可以看作一个条件概率模型，则朴素贝叶斯分类器可以表示为，其中C为类分类有m个值，条件表示有n个特征。对一个样本分类的过程即可理解为比较中各类别对应的概率大小，则问题转换为求。贝叶斯定理可以将其表示为：

 (5-1)

公式(5-1)中，分母可视为是一个固定值Z，因此大小仅与分子有关，使用链式法则可以将公式(5-1)写为：

 (5-2)

由于模型假定各特征间相互独立，故公式(5-2)可以简化为：

 (5-3)

现在以乳腺癌数据分类为例进行下一步公式推导，设数据分类的两个类别S1，S2分别代表患癌样本类和非患癌样本类，想要判断样本D在哪一类，则需要先确定和的大小，根据公式(5-3)以比值的形式进行比较，有：

 (5-4)

分类器的判别规则采用最大似然比判别。设分类器的判决门限为，若，则样本D属于S1类，若则属于S2类。

为方便模型中参数的求解，我们对公式(5-4)两边取对数，有：

 (5-5)

公式（5-5）可进一步写为

 (5-6)

则分类判别规则转换为判断与之间的关系。若大于则判为样本属于S1类，若小于则判为属于S2类。

* + 1. 分类模型显式化表达

根据上一小结对朴素贝叶斯分类判决过程的描述可知，只要知道模型数据中各特征的条件概率以及两类概率及分类器门限值，即可以表格和公式的形式将分类模型表示出来。

根据已有调查数据，可统计出特征各属性在患病情况下的条件概率对数值和非患病情况下的条件概率对数值。将 整理为，患病先验概率对数值与未患病先验概率对数值之差整理为******,则仅需比较

对应，把表5-1和表5-2中各特征的对数值整理为表5-4中的。对应，将表5-3中两类概率对数值（负数）的差值整理为表5-4中的******。当样本进行分类时，可以用查表的形式获取样本特征中具体属性的对应数值，并判断 ******与的大小关系，若******，则样本判为高风险患病类；若******，则样本判为低风险患病类。

* 1. 乳腺癌风险分类分级模型的建立
     1. 分级模型概述

分级是一种常用于揭示数据的真实发展规律的方法。分级方法在处理属性数据上具有一定的优势，因此有研究将“分级”的概念用于建立处理复杂数据的分类模型上。常见的分级模型的流程图如图5-1所示。



**图5-1 分级模型的流程图**

分级模型的建立离不开分级评价标准的确定，根据现有研究分级的评价方法可以分为三类：分级精度评价、总体精度评价以及信息评价。所谓分级精度评价即分级时确保各级模型的分类结果达到最优；总体精度评价是指以模型的总体分类结果达到最优为目的；而信息评价是指分级模型产生的信息以信息熵表示时达到最大。

* + 1. 模型的分级意义与分级结构

常用分类模型在分类时仅判别样本属于哪一类。以第三章NBCOF模型为例，模型的敏感度为0.902，特异性为0.508，漏警率0.098。当样本放入NBCOF模型后模型会输出分类判决结果：属于高风险患病类或低风险患病类。如果想进一步获取患病风险大小的信息，可以根据模型的分类评估参数和模型使用的数据量判断。若乳腺调查数据为100,000例样本，其中乳腺癌患者为200人，敏感度0.902即乳腺癌患者基本被分类准确仅20人被误分为低风险患病类，而特异性0.508即未患病中有50,000人被正确分类为低风险患病类，剩余50,000人则被误分为高风险乳腺癌患病类。若一个筛查者被分类为高风险患病类，则筛查者的患病风险将由未分类前的200/100,000增加到180/50,000，相反若筛查者被分为低风险患病类，则她的患病风险将由未分类前的200/100,000降低到20/50,000。这个数据对于筛查者本人来说比直接的分类结果更有意义，但一个分类器仅能表示一次患病风险改变值。若存在多个分类器则能组合输出不同分类结果下对应的不同风险值，实现了分类结果的风险“可视化”。

本章提出的乳腺癌分级分类模型是为了显示化分类样本患病的风险大小。图5-2展示了分级模型的简要流程图。从图中可以看出，随着分级模型中级数增加，数据中未患病样本所占比例将越来越少。故被越靠后的模型分类为高风险患病类的样本，其对应的患病风险也越高。因此各级模型需要具有较好的分类效果，才能尽可能多地筛除数据中未患病样本数，分级模型的分级评价标准采用分级精度评价。



**图5-2 乳腺癌分级分类模型的简要流程图**

根据上文提出的设计思路和模型评价准则，乳腺癌风险分类分级模型的算法流程图如图5-3所示。

模型流程具体描述如下：

输入：乳腺癌调查数据

输出：判断数据属于高风险患病类/低风险患病类

步骤一：对数据进行预处理，随机取70%的数据作为训练数据。

步骤二：采用基于遗传算法的单类F-score特征选择进行最优子集搜寻。

步骤三：采用朴素贝叶斯分类器作为最优子集的评估模型，评估标准为模型的AUC大小。

步骤四：判断迭代次数是否达到最大值，若未达到，进入步骤二，若已达到，进入步骤五。

步骤五：通过以上迭代获得最优特征子集

步骤六：使用特征子集建立第一级分级模型。

步骤七：使用一级模型对全部数据分类，将被分类为低风险患病的样本从数据中删除，剩余数据进入二级分类模型。

步骤八：重复一级分类模型的特征选择方法和建模方法，建立第二级分类模型。

步骤九：将数据放入二级模型中，筛掉数据中被分类为低风险患病的样本。

步骤十：进入下一级模型。



**图5-3 乳腺癌风险分级分类模型流程图**

由图5-3所示的流程图可知，分级分类模型中每级分类模型会对样本数据进行分类判断，若判为高风险患病类则需进入下一级模型，若判为低风险患病类则分类过程结束。因此随着分类级数的增加，数据中真正患病的样本比例增大，对应的患病风险也会增加。由于每级分类模型可以给出至少一个患病风险评估值，故最终分级模型可达到如图5-4所示的风险评估效果。



**图5-4 分级模型评估样本患病风险流程图**

* + 1. 分级模型的实验结果

限于文章篇幅有限，实验选取两级分类模型展示分级模型如何实现个体患病风险概率的显式化。根据图5-3的分级模型流程图，使用基于遗传算法的单类F-score特征选择进行最优子集筛选，最终建立的一级模型即第四章中的实验模型。一级模型的对应数据如表5-5和5-6所示。则个体对应的患病风险值如表5-7所示。

**表5-5 一级分类模型的评估参数**

|  |  |
| --- | --- |
| 参数 | 数值 |
| 特异性 | 0.542 |
| 敏感性 | 0.910 |

**表5-6 一级分类模型分类结果**

|  |  |
| --- | --- |
| 类别 | 人数 |
| 患乳腺癌 | 282 |
| 未患乳腺癌 | 109032 |
| 共计 | 109314 |

**表5-7 一级模型分类对应的患病概率**

|  |  |
| --- | --- |
| 样本分类结果 | 患病风险概率 |
| 分类为乳腺癌患者 | 0.0051 |
| 分类为非乳腺癌患者 | 0.00047 |
| 未分类前 | 0.0026 |

根据一级模型的分类结果，二级模型的数据如表5-8所示。

**表5-8 二级分类模型分类结果**

|  |  |
| --- | --- |
| 类别 | 人数 |
| 患乳腺癌 | 257 |
| 未患乳腺癌 | 49937 |
| 共计 | 50194 |

基于一级分级模型删减后数据，二级模型同样使用基于遗传算法的NBCOF模型建立，模型对应的评估参数如表5-9所示。故二级分类模型中个体基于分类结果对应的患病风险如表5-10所示。

**表5-9 二级分类模型的评估参数**

|  |  |
| --- | --- |
| 参数 | 数值 |
| 特异性 | 0.406 |
| 敏感度 | 0.833 |

**表5-10 二级模型分类对应的患病概率**

|  |  |
| --- | --- |
| 样本分类结果 | 患病风险概率 |
| 分类为乳腺癌患者 | 0.0070 |
| 分类为非乳腺癌患者 | 0.0022 |
| 未分类前 | 0.0051 |

整理以上表格数据，可以将二级分类模型中各判决结果对应的患病风险概率显示为一个树形结构，具体形式如图5-5所示。



**图5-5 树形结构图**

从图5-5可以看出，当个体进入二级分类模型后，通过一级分类模型会有两种判别结果。设个体未进入分类模型前的患病基准风险为1，若一级模型判为高风险患病类，则该个体患病风险约为2。使用同样的方式，若一级模型判为患病低风险类，则该个体的患病风险约为1/6。验证了上文提出的“若个体被判患病高风险类，则其患病的概率将比未分类前高”的结论。同理，在二级分类模型中，若二级模型判个体为高风险患病类，则个体的患病风险将继续上升，风险值约为2.7。若判个体为低风险患病类，则个体的患病风险将变为4/5。由此可见，树形结构实现了个体进入分级模型后，将不同判决结果对应不同的患病风险量化数值，帮助筛查者对模型的判别结果有更客观的理解。

* 1. 本章小结

本章针对模型需要具备易推广性的特点，采用贝叶斯判别法分析朴素贝叶斯分类器的判别过程，将分类器的分类过程显式化的以查表和公式计算的方式呈现，这便于模型的后期推广工作。同时本章提出了使用分级的疾病风险显性化评估方法，利用一类树状结构的分级疾病风险评估框架，以概率的形式给出了分级模型不同分类结果下显示的患病风险量化值。

1. 总结与展望
   1. 总结

基于我国现阶段医疗资源匮乏且严重不平衡的现象，本文提出疾病风险分类模型需要达到三个目标：低漏警率、低成本以及易推广。为实现这一研究目标，本文以单类F-score特征选择为基础，建立了一系列医学数据的挖掘和分类算法，在乳腺癌人群队列数据上的效果已具备临床指导价值，并可以推广到相似疾病的数据挖掘和风险分类。

本文的贡献可总结为以下几个方面：

（1）针对数据的高维非平衡性造成分类模型高漏警率及高成本问题，本文提出单类F-score特征选择方法，该方法能筛选出对少数类辨别能力强的特征，在解决非平衡数据带来高漏警率的同时，有效减少模型中特征数量的使用。实验结果证实，相比其它算法，NBCOF模型的仅选用8个特征实现了漏警率最低仅9.1%，ROC曲线下面积最大0.776，且模型已具备一定的临床指导价值。

（2）针对已提出的单类F-score特征选择法未考虑特征间相关性的问题，本文提出基于遗传算法的单类F-score特征选择，利用受单类F-score值启发的遗传算法搜索特征集合中的最优特征子集。实验证明优化后的单类F-score特征选择选出的特征子集具有更好的类识别能力，再次降低了分类模型的漏警率。此外，针对遗传算法在迭代过程中耗时大的问题，研究采用建表索引和并行化计算两种方法，有效降低了遗传算法的单次迭代时间，提高了算法的整体性能。

（3）为了支持疾病风险分类模型在我国巨大人口基数中的推广，本文提出两个显性化表示方式。一种是分类算法的显性化，通过深入分析朴素贝叶斯分类器的分类判决过程，以表格和公式的形式具象化表示出来。另一种是分级模型的疾病风险显性化，基于一个树状结构图将分级模型不同分类结果以概率的形式显性化表示。

* 1. 展望

从模型数据角度出发，现有疾病风险分类模型是以乳腺癌队列数据为基础，下一步研究工作可考虑将论文中的一系列算法及显示化方法应用于其他疾病数据中。

从模型算法角度出发，现有模型均基于朴素贝叶斯分类器，下一步研究工作可以考虑采用其它分类器进行疾病风险分类建模。

参考文献

1. Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012. CA Cancer J Clin. 2015, 65(2):87-108.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
2. Desantis C E, Fedewa S A, Mpsh A G S, et al. Breast cancer statistics, 2015: Convergence of incidence rates between black and white women[J]. Ca A Cancer Journal for Clinicians, 2015, 66(1):31–42.
3. Chen W, Zheng R, Baade P D, et al. Cancer statistics in China, 2015[J]. Ca A Cancer Journal for Clinicians, 2016, 66(2):115–132.
4. Gail M H, Brinton L A, Byar D P, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually.[J]. Jnci Journal of the National Cancer Institute, 1990, 81(24):1879-86.
5. Gustafsso K, Aronsson G, Marklund S, et al. The World Bank. Int J Behav Med. 2014. 21(1):77-87.
6. Gustafsso K, Aronsson G, Marklund S, et al. Population and social conditions. Int J Behav Med. 2014. 21(1): 77-87.
7. Liyuan L. A pilot study on risk factors and risk assessment score screening model for high-risk population of breast cancer[D]. Shandong University, 2010:16.
8. He H, Garcia E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9):1263-1284.
9. Demuth H B, Beale M H, De Jess O, et al. Neural Network Design[C]// Wseas International Conference on Circuits. Martin Hagan, 2014.
10. Lei L, Yang D. A GA-based feature selection and parameters optimization for support vector regression.[C]// International Conference on Natural Computation, Icnc 2011, Shanghai, China, 26-28 July. 2011:335-339.
11. Kerr N L, Maccoun R J, Kramer G P. Bias in judgment: Comparing individuals and groups[J]. Psychological Review, 1996, 103(4):687--719.
12. Herman WH,Smith PJ,Thompson TL,et a1.A new and simple question naire to identify rmple at increased risk for undlagnosed diabetes[J].Diabetes Care,1995,18(3):382-387．
13. Ahmed M H, Abdu T A M. Logistic Regression Analysis[M]. Springer Netherlands, 2008.
14. Ruige JB,Neeling JN,Kostense PJ,et a1.Pefronnanee of an NIDDM screning questionnaire based on symptoms and risk factors[J].Diabetes Care.1997.20(4):491-496．
15. Jaana LD,PiIjo IP,Markku P,et al.Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention:follow-up of the Finnish Diabetes Prevention Study[J].Lancet,2006,368(9548):1673-1679．
16. 郑莹, 吴春晓, 张敏璐. 乳腺癌在中国的流行状况和疾病特征[J]. 中国癌症杂志, 2013, 23(8):561-569.
17. 赵洁玉, 徐卫云. 乳腺癌风险评估及预测模型的研究进展[J]. 临床外科杂志, 2013, 21(7):566-568.
18. Adamscampbell L L, Makambi K H, Palmer J R, et al. Diagnostic accuracy of the Gail model in the Black Women's Health Study.[J]. The Breast Journal, 2007, 13(4):329–331.
19. Wen Y C. Validation of the Gail model for predicting individual breast cancer risk in a prospective nationwide study of 28,104 Singapore women[J]. Breast Cancer Research, 2013, 14(1):1-12.
20. Claus E B, Risch N, Thompson W D. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction[J]. Cancer, 1994, 73(3):643-51
21. van Asperen C J, Jonker M A, Jacobi C E, et al. Risk estimation for healthy women from breast cancer families: new insights and new strategies.[J]. Cancer Epidemiology Biomarkers & Prevention, 2004, 13(1):87-93.
22. Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors[J]. Statistics in Medicine, 2012, 10(23):1111-30.
23. Jacobi C E, Bock G H D, Siegerink B, et al. Differences and similarities in breast cancer risk assessment models in clinical practice: which model to choose?[J]. Breast Cancer Research and Treatment, 2009, 115(2):381-90.
24. Hall M, Reid J L, Pruss D, et al. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer.[J]. Cancer, 2009, 115(10):2222–2233.
25. Kwong A, Wong C H, Suen D T, et al. Accuracy of BRCA1/2 mutation prediction models for different ethnicities and genders: experience in a southern Chinese cohort.[J]. World Journal of Surgery, 2012, 36(4):702-13.
26. Liyuan L. A pilot study on risk factors and risk assessment score screening model for high-risk population of breast cancer[D]. Shandong University, 2010:16.
27. D. Hand, H. Mannila and P. Smyth, 2001.“Principles of data mining”, MIT.
28. Chawla N V, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1):1-6
29. Maloof M A. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown[J]. ICML-2003 Workshop on Learning from Imbalanced Data Sets II, 2010.
30. Kubat M. Addressing the Curse of Imbalanced Training Sets: One-Sided Sampling[C]// Proceedings of the International Conference on Machine Learning (ICML-97. 1997.
31. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1):321-357.
32. 古平, 欧阳源遊. 基于混合采样的非平衡数据集分类研究[J]. 计算机应用研究, 2015(2):379-381.
33. Tomar D, Agarwal S. Prediction of defective software modules using class imbalance learning[J]. Applied Computational Intelligence & Soft Computing, 2016, 2016(1):1-12
34. Erfani S M, Rajasegarar S, Karunasekera S, et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning[J]. Pattern Recognition, 2016, 58:121-134
35. PATRICIA RIDDLE, RICHARD SEGAL, OREN ETZIONI. REPRESENTATION DESIGN AND BRUTE-FORCE INDUCTION IN A BOEING MANUFACTURING DOMAIN[J]. Applied Artificial Intelligence, 1994, 8(1):125-147.
36. Kubat M, Holte R, Matwin S. Learning When Negative Examples Abound[C]// European Conference on Machine Learning. Springer-Verlag, 2000:146-153.
37. Weiss G M. Mining with rarity: a unifying framework[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1):7-19.
38. Wu G, Chang E Y. Class-boundary alignment for imbalanced dataset learning[J]. Icml Workshop on Learning from Imbalanced Data Sets, 2003:49--56.
39. Barandela R, Sánchez J S, Garcı́A V, et al. Strategies for learning in class imbalance problems ☆[J]. Pattern Recognition, 2003, 36(3):849-851.
40. Joshi M V, Kumar V, Agarwal R. Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements[M]// Evaluating boosting algorithms to classify rare classes: comparison and improvements. 2001:257-264.
41. Kiryu T. Introduction to Data Mining[J]. 2010, volume 16(472):127-130(4).
42. Tahani H, Keller J M. Information fusion in computer vision using the fuzzy integral[J]. IEEE Transactions on Systems Man & Cybernetics, 1990, 20(3):733-741.
43. Keller J M, Gader P, Tahani H, et al. Advances in fuzzy integration for pattern recognition[J]. Fuzzy Sets & Systems, 1994, 65(2-3):273-283.
44. K. J. Friston, C. D. Frith, P. F. Liddle,等. Functional Connectivity: The Principal-Component Analysis of Large (PET) Data Sets[J]. Journal of Cerebral Blood Flow & Metabolism Official Journal of the International Society of Cerebral Blood Flow & Metabolism, 1993, 13(1):5-14.
45. 西奥多里蒂斯[希腊]. 模式识别: 第2版[M]. 电子工业出版社, 2004.
46. Liu H, Motoda H, Setiono R, et al. Feature Selection: An Ever Evolving Frontier in Data Mining.[J]. 2010, 10:4-13.
47. 王娟, 慈林林, 姚康泽. 特征选择方法综述[J]. 计算机工程与科学, 2005, 27(12):68-71.
48. Chandrashekar G, Sahin F. A survey on feature selection methods [J]. Computers & Electrical Engineering, 2014, 40(1):16-28.
49. Selima S Z, Alsultanb K. A simulated annealing algorithm for the clustering[J]. Pattern Recognition, 1991, 24(10):1003-1008.
50. Trelea I C. The particle swarm optimization algorithm: convergence analysis and parameter selection[J]. Information Processing Letters, 2003, 85(6):317-325.
51. Kononenko I, Kononenko I. Analysis and extension of RELIEF[C]// The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 1994.
52. Xu L, Yan P, Chang T. Best first strategy for feature selection[C]// International Conference on Pattern Recognition. IEEE, 1988:706-708 vol.2.
53. 姚旭, 王晓丹, 张玉玺,等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2):161-166.
54. 徐燕, 李锦涛, 王斌,等. 基于区分类别能力的高性能特征选择方法[J]. 软件学报, 2008, 19(1):82-89.
55. Jain A K, Duin R P W, Mao J. Statistical Pattern Recognition: A Review[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000, 22(1):4-37.
56. Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on Neural Networks, 1994, 5(4):537-550.
57. Sulaiman M A, Labadin J. Feature selection based on mutual information[C]// International Conference on It in Asia. IEEE, 2015.
58. Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research, 2004, 5(12):1205-1224.
59. Fleuret, Fran&#. Fast Binary Feature Selection with Conditional Mutual Information[J]. Journal of Machine Learning Research, 2004, 5(3):1531-1555.
60. Ding C, Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data[C]// Bioinformatics Conference, 2003. Csb 2003. Proceedings of the. IEEE, 2003:523.
61. Hall M A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning[C]// Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2000:359-366.
62. 陈思, 郭躬德, 陈黎飞. 基于聚类融合的不平衡数据分类方法[J]. 模式识别与人工智能, 2010, 23(06):772-780.
63. Fawcett T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8):861-874.
64. Ilango B S, Ramaraj N. A hybrid prediction model with F-score feature selection for type II Diabetes databases[C]//Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India. ACM, 2010: 13.
65. Zhang M L, Peña J M, Robles V. Feature selection for multi-label naive Bayes classification[J]. Information Sciences An International Journal, 2009, 179(19):3218-3229.
66. Akay M F. Support vector machines combined with feature selection for breast cancer diagnosis[J]. Expert Systems with Applications, 2009, 36(2):3240-3247.
67. Liyuan L. A pilot study on risk factors and risk assessment score screening model for high-risk population of breast cancer[D]. Shandong University, 2010:16.
68. 邢文训. 现代优化计算方法[M]. 清华大学出版社, 1999.
69. 沈崇圣. 遗传算法中常用选择算子在MATLAB中的实现[J]. 上海应用技术学院学报(自然科学版), 2003, 3(3):199-202.
70. 张林波. 并行计算导论[M]. 清华大学出版社, 2006
71. 乳腺癌调查问卷

**问卷1**

**一、基本情况**

住院病案号**（就诊号)**

1您的户口所在地 (1)城市 (2)农村

2您的民族：

3您的职业：(1)农民，(2)工人，(3)教师， (4)公务员， (5)个体经商者，(6)司机，(7)服务业，(8)公司职员，(9)家庭主妇，（10）医护人员，（11）学生 （12）其它

4您的教育水平：上学年数 年

4.1最高学历: (1)小学及以下 (2)初中 (3)高中或初中中专(职业高中) (4)大学 (5)研究生

5 您的家庭人口数 ，家庭月平均收入

（1）＜1000元 （2）1000-1999元 （3）2000-2999元 （4）3000-4999元 （5）≥5000元

5.1 与其他人(村里或同事)相比，您家的经济状况如何？

(1) 很好 (2) 比较好 (3)一般 (4) 比较差 (5) 很差

5.2与其他人(村里或同事)相比，您家的社会地位如何？

(1) 很好 (2) 比较好 (3)一般 (4) 比较差 (5) 很差

6您目前的医疗费用来源: (1)完全自费 (2)医疗保险(3)公费医疗(4)新型农村合作医疗

7婚育状况

7.1婚姻状况：(1)单身，从未结婚； (2)已婚并生活在一起； (3)已婚但分居三个月以上； (4)离婚； (5) 寡居 （6）再婚 （7）未婚同居 （8）其它

7.2 共生育 次； 怀孕 次； 流产 次；若流产，首次流产年龄为 岁(周岁)

7.3足月妊娠次数 次 88=不适用

7.4 首次足月生育年龄为 岁(周岁) 88=不适用

7.5孩子个数？ 个， 88=不适用

7.5.1男孩 个， 88=不适用

7.5.2女孩 个, 88=不适用

7.5.3孩子是否足月产： (1)是 (2)否 88=不适用

7.5.4孩子是否母乳喂养？ (1)是 (2)否 88=不适用

7.5.5 若是，累积母乳喂养时间 月。 88=不适用

7.6 您是否常用避孕药的办法控制生育? (1)是 (2) 否 88=不适用

7.6.1若是，共用了 年。 88=不适用

8 生理状况

8.1 您的初潮年龄 岁(周岁) ； 您月经是否有规律？ (1)有 (2)无

8.2 您经期持续天数为 天 ， 月经间隔天数为 天

8.3 您有无痛经： (1)有 (2)无

8.4 您是否绝闭经？(1)是 (2)否 若绝闭经，绝经年龄 岁(周岁)，

若绝闭经，您是否使用雌激素替代用品： (1)是 (2)否 88=不适用

若绝闭经，您的体重是否增加? (1)是 (2)否 若增加,是多少? kg

**二、疾病与家族史**

**(一)疾病史**

9.1 您 (1是 2否) 曾经诊断为乳腺癌?

若是，是 年 月, 由 医院诊断的。

9.2 您 (1是 2否) 曾经诊断为乳腺良性肿瘤？

若是，是 年 月, 由 医院诊断的。

若是，是否为不典型增生（1是 2否）

9.3您 (1是 2否) 曾经诊断为乳腺增生症？

若是，是 年 月, 由 医院诊断的。

9.4您 (1是 2否) 曾经诊断为乳头溢液？

若是，是 年 月, 由 医院诊断的。

9.5您 (1是 2否) 曾经诊断为患有副乳？

若是，是 年 月, 由 医院诊断的。

9.6您 (1是 2否) 曾经诊断为乳头内陷？

若是，是 年 月, 由 医院诊断的。

9.7您是否曾经诊断为患有其它妇科肿瘤？

9.7.1子宫颈癌：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.7.2 卵巢癌：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.7.3卵巢囊肿：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

**9.8您是否曾经诊断为患有其它慢性疾病？**

9.8.1糖尿病：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.8.2高血压：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.8.3 冠心病：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.8.4肾炎：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.8.5精神障碍：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

9.8.6脑血管病：

(1)是 (2) 否 若是，是 年 月, 由 医院诊断的。

**(二)乳腺癌家族史**

10.1 一级亲属内(父亲、母亲、儿子、女儿、兄弟姐妹，共 个人)是否患有乳腺癌：

(1)有 (2)无 若有，请指出是谁 。

10.2 二级亲属内(叔、伯、姑、舅、姨、祖父母、外祖父母，共 个人)是否患有乳腺癌：

(1)有 (2)无 若有，请指出是谁 。

**三、生活习惯**

**11 吸烟**

11.1 您吸烟吗?

(1)从来不吸 (2)偶尔 （3）经常*（****选“从来不吸”跳至问题11.4****）*

11.2 您第1次吸烟的年龄为 岁(周岁)，从开始吸烟到现在共吸烟 年(若期间戒烟，请扣除)。

11.3 在吸烟期间，每天平均吸烟 支。

11.4 与您共同生活或工作的家人或同事是否经常吸烟？ (1)是， (2)否

**12 饮酒**

12.1 您喝酒吗？（1）从不 （2）偶尔 (3)经常*（****选“从不”跳至问题13.1****）*

12.2 若喝酒, 共喝酒 年。

12.3 第一次喝酒(不是抿一下酒)时，您多少岁？ 岁(周岁)

12.4您曾经喝醉过多少次？ (1)没有 (2) 1-2次 (3) 3-5次 (4) 6-9次 (5)10次或以上

12.5 您喝酒的种类为：(1)白酒 (2)啤酒 (3)红葡萄酒 (4)白葡萄酒 (5)其它

12.6您每次饮酒量为多少? ，平均每月喝 次。

**13 饮食(最近一年):**

**1=几乎每天都吃（5-7天）, 2=每周3-4天, 3=每周1-2天, 4=几乎从来都不吃**

13.1 豆制品（豆腐、豆浆等）:

1 2 3 4

13.2新鲜豆类（毛豆、蚕豆等）

1 2 3 4

13.3 红肉类（猪肉、羊肉、牛肉）:

1 2 3 4

13.4 奶制品:

1 2 3 4

13.5玉米类:

1 2 3 4

13.6胡萝卜:

1 2 3 4

13.7煎炸食品（油炸肉、油炸面食）:

1 2 3 4

13.8有色蔬菜或水果:

1 2 3 4

13.9蒜类:

1 2 3 4

13.10火腿肠:

1 2 3 4

13.11腌制食品（咸菜、咸蛋、咸肉等）:

1 2 3 4

14 您经常饮茶吗？

(1)是 (2)否 若是，饮茶种类为：(1)绿茶; (2)红茶; (3)其它

15 您经常饮咖啡吗？ (1)是 (2)否

**16 睡眠状况（最近一年）：**

16.1您一天(包括晚上与白天)一般睡 小时, 这样的情况有 年 月。

16.2您经常**失眠**吗? (1)是 (2)否

16.3您晚上经常做噩梦吗？ (1)是 (2)否

16.4您有**早醒**的情况吗？ (1)是 (2)否

16.5您有**晚睡**的情况吗？ (1)是 (2)否

16.6您对您睡眠的总体满意程度如何？

(1)非常满意 (2)比较满意 (3)不满意 (4)非常不满意

**17 体育活动情况**

17.1 您参加体育锻炼吗？

(1)从不 (2)偶尔 (3)经常

17.2如果您参加锻炼，您目前的锻炼频度？

(1)小于3次/周 (2)大于等于3次/周

17.3您目前的运动量（全天可累计）？

(1)小于30分钟/日 (2)大于等于30分钟/日

**18 个人性格**

18.1 您是个内向的人吗?  (1) 是 (2)一般 (3)否

18.2 您是个非常热情的人吗？ (1) 是 (2)一般 (3)否

18.3 您是个急性子的人吗？ (1) 是 (2)一般 (3)否

**19 生活满意度**

**1=非常满意2=满意 3=一般 4=不满意5=非常不满意 88=不适用**

19.1 您对您**现在**的生活:

1 2 3 4 5 88

19.2 您对您**现在**收入:

1 2 3 4 5 88

19.3 您对您**现在**的身体状况:

1 2 3 4 5 88

19.4 您对您**现在**的住房条件:

1 2 3 4 5 88

19.5 您对您**现在的**婚姻状况(已婚):

1 2 3 4 5 88

19.6 您对您**现在**的就医条件:

1 2 3 4 5 88

19.7 您对您**现在**的邻里关系:

1 2 3 4 5 88

19.8 您对您**半年前**的生活:

1 2 3 4 5 88

19.9 您对您**1年前**的生活:

1 2 3 4 5 88

19.10 您认为您**5年前**的生活:

1 2 3 4 5 88

19.11 您认为您**将来**1年后的生活:

1 2 3 4 5 88

19.12 您认为您**将来5年后**的生活是:

1 2 3 4 5 88

**四、服药、使用化学试剂史**

20.1是否常用染发剂： (1)无 (2)有； 若有， 使用时间 年 月

20.2有无服用雌激素药：(1)无 (2)有； 若有， 服用时间 年 月

20.3有无服用扩血管药：(1)无 (2)有； 若有， 服用时间 年 月

20.4有无服用降脂药： (1)无 (2)有；

20.41若有，服用类别： (1)他汀类， 服用时间 年 月

(2)烟酸类， 服用时间 年 月

(3)贝特类， 服用时间 年 月

(4)胆酸螯合树脂类， 服用时间 年 月

(5)其他， 服用时间 年 月

20.5有无服用降压药： (1)无 (2)有；

20.51若有，服用类别：(1)利尿剂， 服用时间 年 月

(2)血管紧张素转换酶抑制剂，服用时间 年 月

(3)β受体阻滞剂， 服用时间 年 月

(4)钙通道阻滞剂， 服用时间 年 月

(5)血管紧张素Ⅱ受体阻滞剂，服用时间 年 月

(6)α受体阻滞剂， 服用时间 年 月

20.6有无服用降糖药： (1)无 (2)有；

20.61若有，服用类别：(1)二甲双胍类， 服用时间 年 月

(2)磺脲类， 服用时间 年 月

(3)α糖苷酶抑制剂， 服用时间 年 月

(4)噻唑烷二酮， 服用时间 年 月

(5)胰岛素

20.7有无服用抗凝药： (1)无 (2)有；

20.71若有，服用类别：(1)肝素，服用时间 年 月 (2)香豆素类，服用时间 年 月

20.8有无服用抗血小板药： (1)无 (2)有；

20.81若有，服用类别：

(1)阿司匹林， 服用时间 年 月

(2)双嘧达莫， 服用时间 年 月

(3)噻氯匹定， 服用时间 年 月

(4)芬氯咪唑， 服用时间 年 月

(5)血小板糖蛋白Ⅱb/Ⅲa 受体阻滞剂， 服用时间 年 月

20.9有无服用抗抑郁药：(1)无 (2)有； 若有

20.91若有，服用类别：

(1) 单胺氧化酶抑制剂， 服用时间 年 月

(2) 三环类抗抑郁药， 服用时间 年 月

(3) 选择性5-HT再摄取抑制剂类抗抑郁药 ， 服用时间 年 月

(4) 与异构位点结合的5-HT再摄取抑制剂， 服用时间 年 月

(5) 5-HT与NE再摄取抑制剂类抗抑郁药， 服用时间 年 月

(6) NE与DA再摄取抑制剂 ， 服用时间 年 月

(7) 选择性NE再摄取抑制剂, 服用时间 年 月

(8) NE能和特异性5-HT能抗抑郁药, 服用时间 年 月

(9) 5-HT2A受体拮抗剂及5-HT再摄取抑制剂, 服用时间 年 月

(10)草药和非典型抗抑郁药, 服用时间 年 月

20.10有无实施介入手术:(1)无 (2)有 (3)心脏搭桥手术

**五、乳腺相关知识**

21.1 您平常是否关注乳房保健? （1）不关注 （2）比较关注 （3）非常关注

21.2 您是否知道乳房自检方法? （1）是 （2）否

21.3您自我检查过乳房吗? （1）是 （2）否

21.4您是否定期去专门的医院做乳腺检查吗？ （1）是 （2）否

21.5您是了解乳腺癌这种疾病？ （1）很了解 （2）知道一点 （3）否*（****选“否”跳至问题21.6****）*

21.5.1您对乳腺癌的认识主要来自于哪种途径？

(1) 广播电视专题节目 (2) 平时读书看报纸或杂志 (3) 亲戚或朋友患乳腺 (4) 通过互联网了解

(5) 宣传单或宣传手册 (6) 参加过专题讲座或报告等活动 (7) 其他

21.6您是否认为筛检可以早期发现乳腺癌？ （1）是 （2）否

21.7您是否认为早期发现乳腺癌可提高生存率？ （1）是 （2）否

21.8您是否接受过乳腺癌的筛检？ （1）是 （2）否

a．您是否接受过临床触诊？ （1）是 （2）否

b．您是否做过X线检查？ （1）是 （2）否

c. 您是否做过乳腺的超声检查？ （1）是 （2）否

21.9有关乳腺癌症状相关知识

a．您是否认为乳房局部不适是乳腺癌出现时的信号?

（1）是 （2）否 （3）不知道

b．您是否认为乳房有肿块是乳腺癌出现时的信号?

（1）是 （2）否 （3）不知道

c．您是否认为腋窝淋巴结肿大是乳腺癌出现时的信号?

（1）是 （2）否 （3）不知道

d．您是否认为乳头内陷是乳腺癌出现时的信号?

（1）是 （2）否 （3）不知道

e．您是否认为乳头溢液是乳腺癌出现时的信号?

（1）是 （2）否 （3）不知道

21.10乳腺癌相关因素

a．您是否认为初次月经在12岁之前可能会导致乳腺癌的发生?

（1）是 （2）否 （3）不知道

b．您是否认为未生育或晚育可能会导致乳腺癌的发生?

（1）是 （2）否 （3）不知道

c．您是否认为绝经年龄晚可能会导致乳腺癌的发生?

（1）是 （2）否 （3）不知道

d．您是否认为长期饮酒可能会导致乳腺癌的发生?

（1）是 （2）否 （3）不知道

e．您是否认为高脂饮食可能会导致乳腺癌的发生?

（1）是 （2）否 （3）不知道

f．您是否认为长期服用雌激素类药物可能会导致乳腺癌的发生?

（1）是 （2）否 （3）不知道

g．您是否认为有乳腺癌家族史的人更易患乳腺癌?

（1）是 （2）否 （3）不知道

**问卷2**

**视诊**

1皮肤

(1) 正常 (2) 酒窝症 (3) 橘皮样 (4) 酒窝症和橘皮样兼有

1.1若有异常 (1) 左侧 (2) 右侧 (3) 双侧

2乳头 (1) 正常 (2) 凹陷 (3) 糜烂 (4) 上抬

2.1若有异常 (1) 左侧 (2) 右侧 (3) 双侧

3溢液 (1) 是 (2) 否

3.1若有 (1) 左侧 (2) 右侧 (3) 双侧

3.2溢液乳孔 (1) 单个 (2) 多个

3.3溢液性状 (1) 血性 (2) 脓性 (3) 清水样 (4) 黄色 (5)乳汁样

4副乳 (1) 有 (2) 无

**触诊**

1腺体 (1) 正常 (2) 局部增厚 (3) 粗大颗粒状 (4) 条索状 (5)兼有

1.1腺体压痛 (1) 有 (2) 无

2肿块 (1) 有 (2) 无

2.1若有 (1) 左侧 (2) 右侧 (3) 双侧

2.2肿块大小 (1) ＜1cm (2) 1-2cm (3) 2-5cm (4)＞5cm

2.3肿块质地 (1) 硬 (2) 韧 (3) 软

2.4肿块形状 (1) 圆形 (2) 椭圆形 (3) 分页状 (4) 不规则

2.5 肿块边缘 (1) 清楚 (2) 不清楚

2.6 活动度 (1) 大 (2) 中等 (3) 小

2.7 压痛 (1) 有 (2) 无

3腋淋巴结 (1) 融合成团 (2) 触及肿大 (3) 未触及 (4) 其它

3.1 侧别 (1) 左侧 (2) 右侧 (3) 双侧

4锁骨上淋巴结 (1) 融合成团 (2) 触及肿大 (3) 未触及 (4) 其它

4.1 侧别 (1) 左侧 (2) 右侧 (3) 双侧

**5建议** (1) 月经2-3天后复诊 (2) B超 (3) 钼靶 (4) B超+钼靶

**6初步诊断** (1) 未见异常

(2) 乳腺增生

(3) 乳腺良性疾病：①导管扩张②乳头溢液③乳头状瘤④腺瘤⑤增生并腺瘤形成

(4) 良性肿瘤 ： ①乳头状瘤 ②腺瘤 ③增生并腺瘤形成

(5) 可疑恶性

(6) 乳腺癌

(7) 其它

**其 它 体 格 检 查**

1 身高 **□□□.□**厘米 (cm);

2 体重 **□□□.□**公斤(kg)

3 腰围 **□.□**尺, 或**□□□.□**厘米;

4 臀围 **□.□**尺, 或**□□□.□**厘米;

5 血压： 收缩压**□□□** mmHg； 舒张压**□□□**mmHg。

6 血糖（空腹or其它）：**□□□**mmol/L

7 甘油三酯：**□□□**mmol/L

8 总胆固醇：**□□□**mmol/L

9 乳腺B超结果：（附报告单）

10乳腺钼靶结果：（附报告单）

11病理结果： （附报告单）

12.1手术方式：(1)改良根治术 (2) 根治术 (3)保乳手术 (4)改良根治+再造

12.2前哨淋巴结活检术 **：**(1)有 (2) 无

13.1新辅助化疗：(1)有 (2) 无

13.11若有方案为：(1)紫杉类 (2) 蒽环类 (3)环磷酰胺 (4)铂类 (5)5-Fu类 (6)长春碱类

13.12周期数： (1) 1个 (2) 2个 (3) 3个 (4) 4个 (5) 5个 (6) 6个

13.2辅助化疗：(1)有 (2) 无

13.21若有方案为：(1)紫杉类 (2) 蒽环类 (3)环磷酰胺 (4)铂类 (5)5-Fu类 (6)长春碱类

13.22周期数： (1) 1个 (2) 2个 (3) 3个 (4) 4个 (5) 5个 (6) 6个

14辅助内分泌治疗：(1)有 (2) 无

15放射治疗：(1)有 (2) 无

作者简历及在学研究成果

1. 作者入学前简历

|  |  |  |
| --- | --- | --- |
| 起止年月 | 学习或工作单位 | 备注 |
| 2010年09月至2014年06月 | 在北京科技大学通信工程专业攻读 |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. 在学期间从事的科研工作
2. 在学期间所获的科研奖励
3. 在学期间发表的论文

[1] **Xiaoli Lin**, Wei Huangfu, Fei Wang, Liyuan Liu, Keping Long. A breast cancer risk classification model based on the features selected by a novel F-score index for the imbalanced multi-feature dataset[c]. 8th International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (Cyberc 2016), October 13-15, 2016, Chengdu.

独创性说明

本人郑重声明：所呈交的论文是我个人在导师指导下进行的研究工作及取得研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得北京科技大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

签名：\_\_\_\_\_\_\_\_\_\_\_ 日期：\_\_\_\_\_\_\_\_\_\_\_\_

关于论文使用授权的说明

本人完全了解北京科技大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

**（保密的论文在解密后应遵循此规定）**

签名：\_\_\_\_\_\_\_\_\_\_\_ 导师签名：\_\_\_\_\_\_\_\_\_\_\_ 日期：\_\_\_\_\_\_\_\_\_\_\_\_

学位论文数据集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **关键词\*** | **密级\*** | **中图分类号\*** | **UDC** | **论文资助** |
|  |  |  |  |  |
| **学位授予单位名称\*** | | **学位授予单位代码\*** | **学位类别\*** | **学位级别\*** |
| 北京科技大学 | | 10008 |  |  |
| **论文题名\*** | | **并列题名** | | **论文语种\*** |
|  | |  | |  |
| **作者姓名\*** |  | | **学号\*** |  |
| **培养单位名称\*** | | **培养单位代码\*** | **培养单位地址** | **邮编** |
|  | | 10008 | 北京市海淀区学院路30号 | 100083 |
| **学科专业\*** | | **研究方向\*** | **学制\*** | **学位授予年\*** |
|  | |  |  |  |
| **论文提交日期\*** |  | | | |
| **导师姓名\*** |  | | **职称\*** |  |
| **评阅人** | **答辩委员会主席\*** | | **答辩委员会成员** | |
|  |  | |  | |
| **电子版论文提交格式** 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） **推荐格式：**application/msword；application/pdf | | | | |
| **电子版论文出版（发布）者** | | **电子版论文出版（发布）地** | | **权限声明** |
|  | |  | |  |
| **论文总页数\*** | **提示：**  在答辩后给学校上交最终正式论文时，请删除本提示框并填写本页。  在答辩及答辩前，请保持本页及本框不变；建议将本页其他提示框删除。  填写此页时，请删除此信息提示框。 | | | |
| 共33项，其中带\*为必填数据，为22项。 | | | | |