改进随机森林的集成分类方法预测结直肠癌存活性

王宇燕[1]，王杜娟[1]，王延章[1]，Yaochu Jin[1,2]
1 大连理工大学 管理与经济学部，辽宁 大连 116023
2 英国萨里大学 计算机系，吉尔福德 萨里 GU2 7XH

摘要：癌症是人类死亡的主要原因之一，许多国家在癌症方面的支出占医疗总支出的很大比例。癌症存活性预测作为癌症预后的一项重要工作，可以辅助医生做出更精准的诊疗决策，进而降低癌症治疗成本。近年来，基于数据驱动的癌症存活性预测方法逐渐得到应用，而预测的准确性是评价预测方法性能的主要指标，因此提高癌症存活性预测方法的准确性一直是一个活跃的研究领域。

结直肠癌是一种具有高发病率和高死亡率的癌症，为了提高结直肠癌存活性预测的准确性，利用遗传算法对随机森林进行改进，提出基于GA-RF的集成分类方法。该方法通过遗传算法对随机森林中的决策树实行进化搜索，以提高集成分类准确率为目标选出决策树的满意集成。实验分别使用基于GA-RF的集成分类方法、决策树和参数优化的随机森林训练预测模型预测结直肠癌患者的存活性，利用SEER数据库的结直肠癌数据集对3种方法分别进行10折交叉验证，然后用准确性、敏感性和特异性3个指标对它们进行评价。

实验结果显示，基于GA-RF的集成分类方法的预测精度最高(88.2%)，参数优化的随机森林的预测精度次之(86.4%)，但集成复杂度远高于基于GA-RF的集成分类方法，决策树的预测精度最差(74.2%)，而基于GA-RF的集成分类方法还表现出了最好的泛化性能。

该集成分类方法对随机森林进行了有效的改进，能以更高的运算效率和更好的准确性预测结直肠癌存活性，可以为结直肠癌的预后提供决策参考，弥补经验预测的不足，该方法的提出对节约医疗资源、降低医疗成本、提高患者满意度具有实际意义。

关键词：随机森林；遗传算法；集成分类；存活性预测；结直肠癌
中图分类号：TP181 **文献标识码**：A **doi**：10.3969/j.issn.1672-0334.2017.01.009
文章编号：1672-0334(2017)01-0095-12

收稿日期：2016-09-10 **修返日期**：2016-12-22
基金项目：国家自然科学基金(71533001)；中央高校基本科研业务费专项资金(DUT15QY32)

作者简介：王宇燕，大连理工大学管理与经济学部硕士研究生，研究方向为医疗健康管理、数据挖掘和机器学习等，E-mail：wyy@mail.dlut.edu.cn

王杜娟，工学博士，大连理工大学管理与经济学部副教授，研究方向为服务运作管理、数据挖掘和智能优化算法等，代表性学术成果为“恶化效应下加工时间可控的新工件到达干扰管理”，发表在2016年第5期《系统管理学报》，E-mail：wangdujuan@dlut.edu.cn

王延章，工学博士，大连理工大学管理与经济学部教授，研究方向为数据挖掘和知识管理等，代表性学术成果为“基于知识元的突发事件风险熵预测模型研究”，发表在2016年第1期《系统工程学报》，E-mail：yzwang@dlut.edu.cn

Yaochu Jin，工学博士，英国萨里大学计算机系计算智能首席教授，研究方向为计算智能、机器学习、计算生物学和计算神经科学等交叉学科的理论研究和工程应用等，代表性学术成果为“A social learning particle swarm optimization algorithm for scalable optimization”，发表在2015年第291卷《Information Sciences》，E-mail：yaochu.jin@surrey.ac.uk

引言

结直肠癌是世界范围内很常见的一种癌症,根据2012年全球癌症统计数据,结直肠癌发病率在男性和女性常见恶性肿瘤中分别排第3位和第2位,其死亡率高达49%[1]。结直肠癌作为一种高死亡率的癌症,长期面临的一个重要临床问题是发病后预测其结局(存活或者死亡),即癌症存活性预测,此处存活的定义是确诊病人在指定时间内处于活着状态。癌症存活性预测是癌症预后的一项重要工作[2],疾病结局预测准确性越高,医生做出的临床决策就会越精准,从而更好地提高治疗效率和效果[3]。因此,提高癌症存活性预测的准确性十分重要。

医院多使用传统的统计学方法进行回顾性数据分析,很少做前瞻性的预测研究。机器学习技术的发展使事前预测成为可能,一些机器学习方法在医疗领域得到了较好的应用,如决策树[4-5]、人工神经网络[6-7]、支持向量机[8-9]和随机森林(random forest, RF)[10-11]等。但是,当前应用机器学习方法对癌症的研究多集中于癌症的诊断,对癌症预后(如存活性预测)的研究较少,而且多为研究乳腺癌[12-13]、前列腺癌[14-15]等,对结直肠癌的研究较少。本研究提出一种新的集成学习方法预测结直肠癌患者的存活性,将癌症存活性预测视为一个分类问题,预测病人是否在指定时间后依然存活。叶强等[16]曾利用遗传算法构建分类器融合模型,得到了明显优于单个分类器的分类效果。本研究利用遗传算法(genetic algorithm,GA)在随机森林的决策树集合中构造更优集成,通过与几种常用算法进行实验对比,发现该方法在结直肠癌存活性预测中可以获得更高的准确性。

1 相关研究评述

本研究提出改进随机森林的方法预测结直肠癌的存活性,下面将从癌症存活性预测研究和随机森林的改进研究两个方面介绍该领域的相关工作。

1.1 癌症存活性预测研究

目前,机器学习方法已经广泛用于癌症研究,利用它们构造的预测模型使医疗决策变得更加高效和准确[2]。同时,随着该领域研究的深入,也促进了各种方法的改进和发展[3]。

生存分析是医疗预后的一项重要工作,可以利用一些方法和技术,基于病人的历史数据进行存活性预测[17]。早些年的研究多将机器学习方法与统计学方法进行对比,验证机器学习方法可以有效用于癌症存活性预测。DELEN[18]将决策树、人工神经网络、支持向量机3种流行的机器学习方法和一种最常用的统计分析方法logistic回归应用于预测前列腺癌患者的存活性,实验结果表明,支持向量机预测的准确性最高,决策树和人工神经网络次之。

随着机器学习方法使用的普及,学者们开始改进传统单一的算法,得到混合模型,应用效果通常比原始算法更好。KHAN et al.[19]分析了利用基于模糊逻辑的分类器预测癌症患者存活性的可行性,将模糊理论和决策树结合起来构造出加权模糊决策树(wFDT),并利用SEER乳腺癌数据集进行存活性预测实验,发现wFDT的预测性能要优于决策树;WANG et al.[20]提出将合成少数类过采样法(synthetic minority oversampling technique,SMOTE)与粒子群优化算法(particle swarm optimization,PSO)、logistic回归、决策树、k临近算法等分类算法中的一种结合起来,形成一种新的分类算法,并应用于乳腺癌病人存活性预测,SMOTE对原始的类别不平衡数据进行调整,使用PSO进行特征选择,然后使用分类算法进行分类实验,并用10折交叉验证测试算法,其中,SMOTE与PSO、C5决策树的结合在实验中表现出了最好的分类性能,研究表明这种混合算法可以有效提高乳腺癌病人存活性分类的准确性。

上述研究中均使用单个分类器进行预测,虽然各种方法构造的分类器在实验中获得了较好的准确率,但是单个分类器始终存在预测精度不够高、泛化性能不够好等缺点。近年来,一些学者在该领域已经不再局限于单个分类器的使用,越来越多地使用由多个分类器构成的集成分类器,集成分类器通常具有更高的分类准确性和更强的泛化性能[21],可有效用于癌症患者存活性预测。ZOLBANIN et al.[22]在考虑并发癌症的情况下预测癌症的存活性,利用逻辑回归、人工神经网络、决策树、随机森林4种方法分别进行实验,随机森林作为一种集成学习方法获得了最高的准确率。相对于单一学习方法,集成学习方法本身在很多方面具有优越性,而随机森林更是被誉为"代表集成学习技术水平的方法",随机森林简单、容易实现、计算开销小,在很多现实任务中展现出强大的性能[21]。但为了进一步提高分类精度和效率,有学者对其进行改进。

1.2 随机森林的改进研究

有学者认为通过对随机森林的关键参数进行优化,可以实现在分类运行效率可接受范围内的更高分类精度[23]。QIAN et al.[24]将随机森林应用于定位前列腺癌发病的位置,遍历了所有的参数组合方式,并用交叉验证的方法选择最佳的参数,以达到最高的精度,但是运算效率较低;周天宁等[25]将特征个数和决策树数量作为遗传算法要优化的变量,最终进化得到满意的参数组合,这种方法在保证计算效率的同时改善了随机森林的分类效果。虽然,通过各种方式可以得到随机森林的满意参数,改进随机森林的性能,但是,由于随机森林对参数的敏感性不是很强,所以参数优化后性能的提高幅度并不大。

作为一种集成学习方法,要想得到好的集成,还需要考虑个体学习器的特点,主要是准确性、多样性和个体学习器的数量。随机森林由决策树组成,其性能很大程度上取决于构成它的决策树。随机森林最大的特点是通过增加样本扰动和属性扰动使个体学习器之间产生差异性[26],进而使集成的性能增加。属性扰动往往导致个体学习器的性能下降,但随着个体学习器数量的增加,随机森林通常会收敛

到更小的泛化误差,然而预测效率也会跟着降低,随之产生的问题是使用少量的个体学习器能否达到更好的效果。ZHOU et al.[27]提出选择性集成的概念。选择性集成是指在已构建的个体学习器中用某种策略选出一部分构成新集成,研究表明选择性集成可以把预测性能不好的个体学习器剔除,只保留少量优质的个体学习器,从而提高集成预测性能[28]。另外,选择性集成还可以提高集成的泛化能力[29]。

HONG et al.[30]使用多样化集成遗传规划方法进行精确的癌症分类,首先通过特征选择生成多个分类规则,然后计算分类规则之间的差异性,选出差异较大的分类规则构成一个子集,最后利用该子集中的分类规则组成一个集成分类器,强调个体分类器的多样性越大,集成分类器的准确性越高。但是,如果个体分类器的性能很差,集成之后性能也不会太好,所以个体分类器的准确性和多样性都很重要。HONG et al.[30]研究中的差异性是直接通过比较规则的结构得到的,增大多样性的同时并不会使准确性降低,避免了传统方法用分类结果计算差异性会产生的"准确性 - 多样性"困境,所以在准确性较好的情况下尽量增大多样性,从而实现了集成分类器准确性的提高。但HONG et al.[30]研究中计算多样性的方法并不适用于非规则形式的分类器,因此,按照这种直接比较的方式判断随机森林中个体学习器的优劣性是不可行的。

考虑到随机森林中并非所有决策树都对提高集成分类器的准确率产生积极作用,那么,如何在随机森林的众多决策树中选出一部分以更低的复杂度构成更优的集成分类器是问题的关键。目前,选出最佳的分类器组合构造好的集成是一个NP难问题[31]。本研究提出使用遗传算法解决该问题,形成基于GA-RF的集成分类方法。利用遗传算法对随机森林中的决策树进行进化搜索,以提高集成分类准确率为目标选出决策树的满意集成。

2　基于GA-RF的集成分类方法

集成学习通过构建并结合多个学习器完成学习任务[21]。与一般的学习方法不同的是,一般的学习方法是从训练数据构造一个学习器,而集成学习方法是构造多个学习器并将它们结合起来[32],常用的结合策略是针对数值型输出的平均法和针对分类任务的投票法。个体学习器通常由一个现有的学习算法从训练数据中产生,如决策树、神经网络等。集成学习通过将多个学习器进行结合,常可获得比单一学习器显著优越的泛化性能。

分类器的集成是指将一个分类器集合里面所有个体的决策结果以某种方式(典型的是带权重或不带权重的投票)结合,并用来对新的实例进行分类[33]。

为了提高癌症存活性预测精度,本研究在随机森林的基础上提出一种新的集成分类方法,该方法利用遗传算法对随机森林中的决策树进行进化搜索,选出决策树的满意组合,这些决策树以某种策略相结合构成新的集成。基于GA-RF的集成分类器构建流程见图1,其中,N为随机森林的决策树总量,N'为新决策树集合的决策树数量,$N' < N$。该方法主要由3个部分构成,①根据训练集构建一定规模的随机森林;②利用遗传算法以提高测试集分类精度为目标对随机森林中的决策树进行进化搜索,得到新的决策树集合;③根据某种策略将得到的决策树结合为集成分类器。下面将具体介绍上述3个部分。

2.1　随机森林的生成

随机森林是决策树预测器的组合,每棵树的生成都依赖于一个独立采样的随机向量值,这些随机向量具有相同的分布,每棵树独立运算得到其分类结果,然后投票决定最终的分类结果[34]。

随机森林的主要思想是:如果单棵树是好的,那么只要树之间有足够的差异就会构成效果更好的森林。随机森林最有特点的地方是它运用两种方法从一个标准数据集创造随机。第一种方法是bagging,构建每棵树时,先对数据集进行自举重采样(bootstrap)得到一个训练集,这样训练每棵树使用的都是不同的训练集。第二种方法是限制构造决策树时的可用特征,训练决策树时,每个结点处生成一个特征的随机子集,树的分裂节点只能从该特征子集中选取,这样做不仅增加了训练每棵树的随机性,还由于遍历

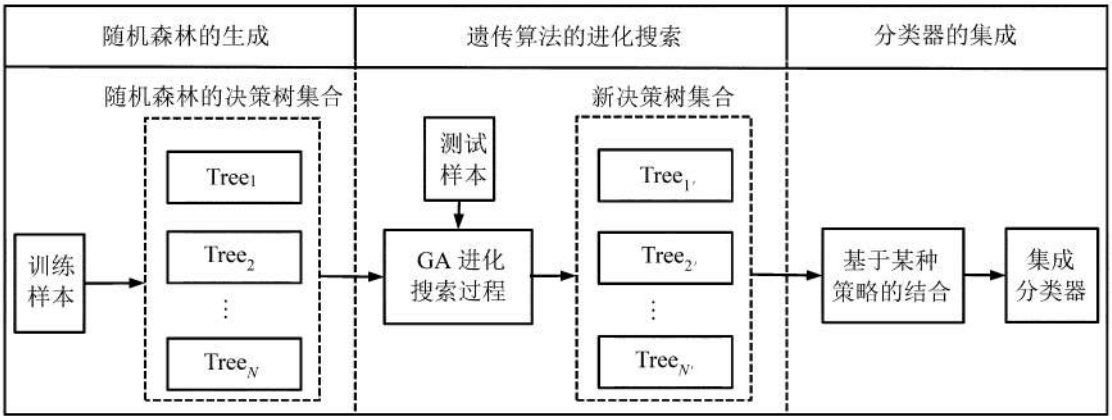


图1　基于GA-RF的集成分类器构建流程

Figure 1　Process of Constructing GA-RF Based Ensemble Classifier

更少的特征而加快了训练速度,当然,可用特征数量设置的越少,产生的决策树多样性越大。

本研究使用的随机森林训练算法[35]训练过程如下。随机森林中每棵树的构建重复以下步骤:①对训练集做一次bootstrap得到子训练集;②使用得到的子训练集开始训练一棵决策树;③在决策树的每个结点处,随机选择n个特征并通过计算它们的信息增益(或基尼不纯度)选出最优的一个作为当前结点,重复该步骤直到一棵树构造完成。

基于此训练算法可得到随机森林的决策树集。考虑到随机森林通过特征数量限制和对训练样本的bootstrap,可以保证所构造决策树的多样性,本研究将随机森林的决策树空间作为个体分类器的选择空间,从训练好的决策树中选出有益于提高集成分类性能的个体,这种对个体分类器的优选为得到比随机森林更好的集成奠定了基础。为了使个体分类器产生足够的差异性,特征数量应该尽可能地少。

2.2 遗传算法的设计

遗传算法是解决复杂优化问题使用最广泛的元启发式方法[36-38]。遗传算法模拟生物遗传进化的过程[39],首先初始化种群,其中每条染色体代表一个解,由适应度函数值衡量解的好坏并确定出下一代的父母,然后通过交叉和变异生成下一代种群,如此不断循环得到满意解或达到设置的代数时结束进化。

本研究的遗传算法流程见图2。当进化到设定代数时,会得到该参数条件下的最优决策树组合,该决策树组合集成后会得到比原随机森林更高的准确性。下面分别介绍遗传算法设计的几个关键点,即染色体编码、适应度评估、选择操作和交叉变异操作的设置。图2中两个虚线箭头分别表示本研究选择操作和交叉操作选用的方法。

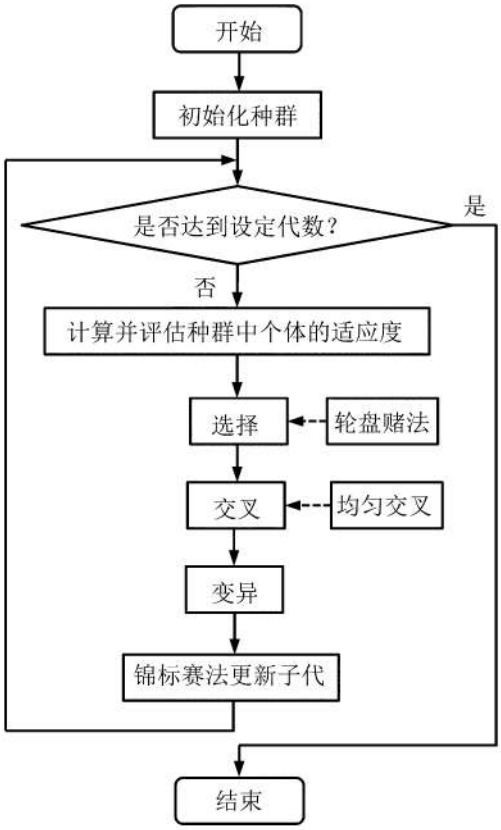


图2 遗传算法流程图

Figure 2 Flow Chart of Genetic Algorithm

(1)染色体编码

遗传算法的选择、交叉、变异不是对问题的决策变量直接实施的操作,而是对可行解编码后的个体进行的运算,所以要先进行染色体编码[40]。编码是问题表示的过程,是指将待求解问题的变量转化成遗传算法能直接处理的染色体形式,使每条染色体对应问题的一个解。常用的染色体编码方式有二进制编码、实数编码、有序串编码和结构式编码等,而最常用的是二进制编码和实数编码。

随机森林是利用已构建的所有决策树进行分类计算,而本研究旨在从随机森林的决策树集合中选出一部分,排除掉对提高集成分类准确性无益的那部分,从而使构成的集成分类器更准确。所以对每棵决策树而言,需要确定是否应该选用。因而编码方式采用二进制编码,染色体的长度即为随机森林中决策树的数量,染色体上每一位对应一棵决策树,每一位上的值为1或0,1为选用该决策树,0为不选用该决策树,假设随机森林有10棵决策树,则染色体的一种编码结果见图3,图中编号为2、4、6、8、10的决策树为根据这种编码得到的决策树组合。

决策树编号	1	2	3	4	5	6	7	8	9	10
二进制编码	0	1	0	1	0	1	0	1	0	1

图3 染色体编码举例

Figure 3 An Example of Chromosome Coding

(2)适应度评估

在遗传算法中,适应度评估决定了种群进化的方向。适应度评估需要设计适应度函数并计算适应度函数值,适应度函数也叫评价函数,适应度值的大小可以评估种群中个体的优劣程度。使用适应度函数计算种群中每一个个体(即每一个解)对应的适应度值,使该适应度值成为选择交配个体的依据,从而使种群向有利于求得最优解的方向演化。

从前面介绍的编码方式可知,一条染色体对应的是基于随机森林决策树集的一种决策树组合方式,而哪一种组合方式形成的集成更好还需要适应度函数来评估。本研究旨在得到分类正确率高的集成分类器,因此适应度函数被设计为计算集成分类器的准确率,以此评估决策树组合的优劣。随机森林的决策树集建好之后,可以得到每棵决策树对所有测试样本的分类结果,然后通过投票法就可以得到每种决策树组合对所有测试样本的分类结果,进而将该分类结果与测试样本类别相比较得到每种决策树组合的集成分类正确率,该值即作为对应染色

体的适应度。

对M个测试样本的二分类问题,适应度函数定义为

$$F = 1 - \frac{1}{M}\sum_{k=1}^{M}[P(\boldsymbol{x}_k) - y_k]^2 \quad k = 1,\cdots,M \tag{1}$$

其中,F为集成分类器的准确率,$\boldsymbol{x}_k$为第k个测试样本,$P(\boldsymbol{x}_k)$为集成分类器对第k个测试样本的分类结果,y_k为第k个测试样本的类标0或1。

(3)选择操作

适应度值的大小表征了染色体对应的个体的优劣,适应度值越大表示个体越优。按照优胜劣汰的自然选择原则,好的个体的基因保留下来的概率更大,因而适应度值高的个体被选作下一代的父母的概率更高,得到下一代的父母是由选择操作实现的。本研究用轮盘赌法进行选择操作,该方法使个体被选择的概率与其适应度值成正比,个体α被选择的概率p^{α}可表示为[35]

$$p^{\alpha} = \frac{F^{\alpha}}{\sum_{\alpha'} F^{\alpha'}} \tag{2}$$

其中,F^{α}为个体α的适应度值,$F^{\alpha'}$为个体α'的适应度值。通过轮盘赌法对个体的选择,使每一代中优秀的决策树组合以更高的概率保留下来,并有机会在此基础上产生更优的组合。

(4)交叉和变异操作

生成子代的两个基本操作是交叉和变异。交叉操作是对父代的两个染色体实施的,从而得到由父代的第1个染色体的一部分和第2个染色体的一部分构成的新染色体。最常见的交叉方式是单点交叉,先在染色体上随机选择一个点,在该交叉点前面的部分从父代中第1个染色体获得,后面的部分则由父代中第2个染色体获得。按照这种方式可以扩展为多点交叉。还有一种方式,独立地从父代两个染色体中随机选择基因从而构成新染色体,叫均匀交叉。以二进制编码方式为例,3种交叉类型见图4。图4中,单点交叉的交叉位点位于第3个基因位与第4个基因位之间,其子染色体由第1条染色体的前3个基因和第2条染色体的后8个基因构成;多点交叉有两个交叉位点,分别位于第3个基因位与第4个基因位之间和第9个基因位与第10个基因位之间,其子染色体由第1条染色体的前3个基因、第2条染色体的中间6个基因和第1条染色体的后两个基因构成;均匀交叉的子染色体的基因构成取决于随机交叉位点,交叉位点为0的基因位采用染色体0的基因,交叉位点为1的基因位采用染色体1的基因,由此构成子染色体。为了增大产生多样个体的可能性,本研究选用均匀交叉法进行交叉操作。

实施交叉后,利用变异操作产生解空间的随机性。二进制编码中,常用的变异方式为染色体某一位的翻转,即1变为0或0变为1,这种变化对于随机森林的决策树集而言,意味着某棵决策树从被选用状态到不被选用状态的转变或者从不被选用状态到被选用状态的转变。变异可以在一定程度上增加决策树组合的随机性,避免陷入局部最优[41]。

经过交叉、变异操作得到子代之后,为了保留上一代中的优秀个体,本研究采用锦标赛法对该子代进行更新。锦标赛法是将父代和经交叉、变异后得到的子代放在一起,按适应度进行排序,再依据种群大小选出优秀个体作为新子代。这样,每一代进化得到的个体是两代种群中最优秀的,可以加快进化速度。

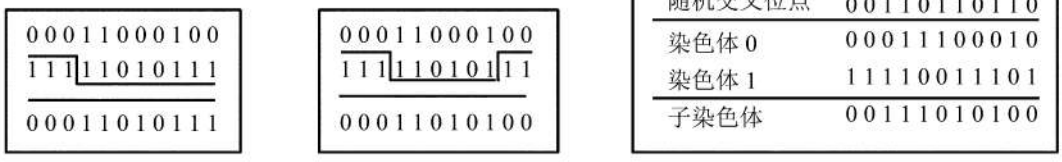
2.3 分类器的集成

通过遗传算法的进化搜索得到随机森林决策树空间中个体分类器的满意组合,由最终进化得到的染色体中值为1处对应的决策树构成,接下来的任务是以合适的方法将个体分类器结合起来。分类器集成问题的重点在于找到一种策略来综合个体分类器的分类结果,从而得到集成的分类结果,结合策略会对集成的性能产生一定影响。

分类任务中最常用的结合策略是投票法。假设有一个由T个分类器构成的分类器集合$\{h_1,\cdots,h_T\}$,需要将这些分类器结合起来从类标$\{c_1,\cdots,c_l\}$中预测出样本$\boldsymbol{x}$的类别,分类器h_i的预测结果用l维向量表示为$[h_i^1(\boldsymbol{x}),\cdots,h_i^l(\boldsymbol{x})]^T$,$h_i^j(\boldsymbol{x})$为分类器$h_i$关于类标$c_j$的输出,$h_i^j(\boldsymbol{x})$有两种不同形式的取值,$h_i^j(\boldsymbol{x}) \in \{0,1\}$或$h_i^j(\boldsymbol{x}) \in [0,1]$。$i = 1,2,\cdots,T$, $j = 1,2,\cdots,q,\cdots,l$。对于第1种形式,当$h_i$的预测结果为$c_j$时,$h_i^j(\boldsymbol{x})$取值为1,否则取值为0;对于第2种形式,$h_i^j(\boldsymbol{x})$的值为$h_i$的预测结果是$c_j$的概率。

投票法也有不同的机制,最常见的是绝对多数投票法和相对多数投票法。绝对多数投票法中每个分类器只能向一个类标投票,得到一半以上票数的类标为最终的类别,若无类标得到一半以上票数则无分类结果。设$H(\boldsymbol{x})$为集成输出的类标,基于绝对多数投票法的集成输出类标可定义[32]为

$$H(\boldsymbol{x}) = \begin{cases} c_j & \text{若} \sum_{i=1}^{T} h_i^j(\boldsymbol{x}) > \frac{1}{2}\sum_{q=1}^{l}\sum_{i=1}^{T} h_i^q(\boldsymbol{x}) \\ \text{无类别} & \text{其他} \end{cases} \tag{3}$$



(a)单点交叉　　(b)多点交叉　　(c)均匀交叉

图4 单点交叉、多点交叉、均匀交叉示例

Figure 4 Examples of Single Point Crossover, Multi-point Crossover and Uniform Crossover

相对多数投票法则将票数最多的类标作为最终类标,不会出现无分类结果的情况,若出现多个类标获得最高票数,则随机选择一个作为最终类标。基于相对多数投票法的集成输出类标可定义[32]为

$$H(\boldsymbol{x}) = c_{\arg\max_j \sum_{i=1}^{T} h_i^j(\boldsymbol{x})} \tag{4}$$

本研究探讨一个二分类问题,预测结直肠癌患者是否会存活,分类器h_i将从类标{0,1}中预测出样本的类别。由于该预测必须得到类标,明确癌症患者的结局是生存还是死亡,故选择相对多数投票法作为分类器集成的方法,将得到的决策树组合进行结合。若有一半以上的分类器预测结果为0,则集成分类结果为0;若有一半以上的分类器预测结果为1,则集成分类结果为1;若预测结果为0和1的分类器各占一半,则随机选择一个作为集成分类结果。至此,集成分类器就生成了。

3 结直肠癌存活性预测实验

3.1 数据准备

如果一开始的输入数据集质量不高,那么任何结果和发现都会受到质疑[42]。所以,在数据挖掘过程中,实验数据处理需要投入大量的时间和精力。下面介绍本次实验关于数据的准备工作。

3.1.1 数据获取

本次实验使用的数据是SEER的结直肠癌数据,SEER计划(The Surveillance, Epidemiology, and End Results Program)是美国国家癌症协会的一个项目,从全美国各个地区和机构收集癌症病例的数据,在美国是一个权威的癌症数据源,也被看作是全世界癌症注册机构的质量标准[17]。SEER癌症数据量较大,数据的结构文档完整,可供研究人员免费获取和使用,从SEER的官网(http://www.seer.cancer.gov)可以知道数据获取的详细信息。SEER数据库已被广泛应用于各种分析研究项目,美国医学的数据库(PUBMED)在2014年的搜索显示,超过540个出版物使用该数据集作为主要数据源或做癌症统计报告[22]。

本研究采用的数据包含1973年至2013年的癌症数据,根据癌症类型共有9个TXT文档,其中COLRECT存储的是结肠、直肠癌(合称结直肠癌)的数据,本研究利用该数据进行结直肠癌患者存活性预测实验。

3.1.2 数据预处理

数据准备过程包括理解数据的含义、探知变量的统计和分布情况,进行适当的转换、处理缺失值、缩减数据量等,因此数据准备会花掉数据挖掘开始部分工作时间的一半甚至80%[43]。

结直肠癌的原始数据集包含134个变量和超过110 000条记录,但并非所有的变量和记录均用于此研究。为了更好地理解并选择有意义的变量,研读数据的说明文档和癌症的编码、分期手册,了解字段的名称、含义、数字编码方式和癌症各个属性的统一编码标准。公开版本的数据中的变量大致可分为7类,具体见表1。

表1 变量分类

Table 1 Classification of Variables

类别编号	变量类别
1	记录标识
2	人口统计信息
3	肿瘤描述
4	数据来源
5	第一个疗程的情况
6	跟进信息
7	重新编码的变量

关于数据的变量做出如下处理。重新编码使原有列重复出现,因而去掉由于癌症编码标准更新而重新编码的字段;一些字段与病人的死因或关键的状态直接相关,不能作为模型的输入,故删除。经过删减最终剩余56个字段。此时,数据已比较完整,个别字段只存在个别缺失值。根据数据的说明文档,存在缺失值的这些字段,说明文档对缺失值均有特定的数字编码,于是直接向每列的缺失值处填充即可。

得到完整的数据后,还需要找到目标变量。结直肠癌数据中一个名称为STAT_REC的字段是表示癌症患者在SEER随访研究期间是否死亡的状态,4为死亡,1为存活,本研究将该字段作为数据的目标变量进行分类实验。本实验旨在预测癌症存活性,因此还需要删除死因为非癌症的数据。由于结直肠癌数据总量较大,删除非癌症死因的记录之后依然较多,且年份较早的数据缺失值较多,为方便实验,基于正负类样本数量平衡的原则,从诊断年份为2013年的数据中随机选取1 000条作为实验数据。

3.2 实验设置

本次实验主要针对两种方法进行,分别是改进参数的随机森林[25]和基于GA-RF的集成分类方法,并利用几个评价分类性能的指标对实验方法进行评估。下面先说明实验参数设置情况,然后介绍本研究使用的分类性能评价方法。

3.2.1 实验参数设置

首先介绍本研究提出的基于GA-RF的集成分类方法的参数设置。该方法利用遗传算法在随机森林中寻找满意的集成,首先需要构建随机森林,然后才是遗传算法的整个流程,所以参数设置也涉及到这两个方面。①随机森林涉及到的两个主要参数是特征数量和决策树数量。为了使随机森林的决策树多样性足够大,特征数量设为1;随机森林的决策树集是新集成的个体分类器选择空间,又考虑计算效率的因素,按照经验将决策树数量设为100。②遗传算

法主要有种群大小、变异率和进化代数3个参数需要设置,综合考虑运算效率和效果,将种群大小设为10,变异率设为0.05,进化代数设为50,其他参数取默认值。

周天宁等[25]提出利用遗传算法对随机森林参数进行优化,优化的参数为特征数量和决策树数量,采用实数编码,染色体长度为2,两个基因位的值即为两个参数的取值。特征数量的最小取值为1,最大值为实验数据的总特征数;决策树数量的最小取值为1,最大值可以是+∞,但是取值范围过大会影响运算效率,故根据经验人为选定一个较大的数来代替最大值,以减少计算量。参数优化的目的是通过找到满意的参数组合获得更高的分类精度,因此适应度函数设置为计算随机森林的分类正确率。本次实验中,特征数量取值范围为[1,55],决策树数量的取值范围设为[1,500],种群初始化以及后续的交叉、变异操作都在该取值范围内进行。考虑到比较的公平性,同样将种群数量设为10,变异率设为0.05,并引入锦标赛法来选择下一代个体,当代数达到50代时算法停止,遗传算法其他参数取默认值。

3.2.2　分类性能评价方法

预测模型的好坏需要适用的衡量手段来评估。本研究选用医疗领域诊断预测中最常用的评价指标:敏感性、特异性和准确性,并使用交叉验证进行测试,下面分别做出介绍。

(1)分类性能评价指标

本研究使用的分类性能评价指标为敏感性、特异性和准确性。这3个指标在医疗领域的诊断预测中被广泛使用,主要用于衡量某项预测的效果和可靠性[44],DELEN et al.[17-18]、KHAN et al.[19]和ZOLBANIN et al.[22]在评价预测癌症存活性的模型时均使用了它们。敏感性用来评价对实为阳性者的检测效果,如患病者有多大可能被检查出有病;特异性评价对实为阴性者的检测效果,如无病者有多大可能被正确地排除;准确性由敏感性和特异性决定,从整体上判断一项预测的准确性。

$$\text{敏感性} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{特异性} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{准确性} = \frac{TN + TP}{TN + TP + FN + FP} \tag{7}$$

其中,TP为真正类,TN为真负类,FP为假正类,FN为假负类。敏感性表示正类样本被正确分类的比例,特异性表示负类样本被正确分类的比例,准确性表示所有测试样本被正确分类的比例。上述3个指标均为越大越好。

(2)K折交叉验证

为了测试分类方法独立于训练样本的泛化能力,本研究使用K折交叉验证对分类方法的精度进行估计。K折交叉验证[45]是指将完整的数据集分成大致相等的K个互斥子集,对模型进行K次训练、测试,每次轮流使用不同的$(K-1)$个子集训练模型,剩余一个子集测试模型。然后将K次运算得到的正确率的均值作为该方法的精度估计值。K折交叉验证的正确率$\bar{A}$为

$$\bar{A} = \frac{1}{K}\sum_{m=1}^{K} A_m$$
$$m = 1,2,\cdots,K \tag{8}$$

其中,A_m为第m折运算得到的正确率,K为折数。通过比较K折交叉验证得到的正确率,可以知道分类方法的整体性能,再通过计算其标准差可比较不同方法的稳定性。

3.3　实验结果和分析

本研究选用的分类性能评价方法在前面已经进行详细介绍,下面先从指标值比较的角度分析交叉验证得到的实验结果,然后对两种集成分类方法的集成复杂度进行比较分析。

(1)指标值比较分析

为了验证集成学习方法比构成它的单一学习方法具有更好的性能且本研究提出的方法具有更高的预测精度,本研究将基于GA-RF的集成分类方法与改进参数的随机森林以及与构成它们的单一学习方法决策树进行对比,3种方法分别进行10折交叉验证实验。表2为对结直肠癌数据分类的实验结果,包含了由3种方法构建的分类器在结直肠癌数据集上10折交叉验证得到的各项指标值以及它们的均值和标准差,其中参数优化的随机森林的指标值使用遗传算法进化完成获得的参数值得到,基于GA-RF的集成分类方法的指标值由该方法构建的集成分类器得到。

首先,对每一折实验中3种方法的3个指标值进行对比,图5以折线图的形式展示了3种方法在10折交叉验证中得到的敏感性值,图6和图7分别为特异性值和准确性值。由图5可以看出,随着折数的变化,即训练集和测试集的更改,3种方法得到的敏感性值变化趋势相似且波动较大,该波动是正类样本在每折验证中分布不均衡引起的,从每折实验看,基于GA-RF的集成分类方法和参数优化的随机森林在敏感性上的优劣很难分辨,且二者整体上优于决策树,但也存在决策树获得较高敏感性值的情况。由图6可以看出,基于GA-RF的集成分类方法几乎每折的特异性值都优于参数优化的随机森林,且二者均优于决策树。而由图7可以看出,3种实验方法在每折实验的准确性上的排序为:基于GA-RF的集成分类方法优于参数优化的随机森林,参数优化的随机森林优于决策树。

然后从整体上对交叉验证的结果进行分析。由表2可知,决策树方法、参数优化的随机森林和基于GA-RF的集成分类方法这3种方法的敏感性的均值分别为0.482、0.583、0.588,特异性的均值分别为0.823、0.950、0.972,准确性均值分别为0.742、0.864、0.882,3个指标的均值都呈现出从小到大的顺序;3种方法敏感性的标准差分别为0.132、0.121、0.118,特异性的标准差分别为0.065、0.020、0.013,准确性的标

表 2 分类实验结果

Table 2 Experiment Results of Classification

折数	决策树			参数优化的随机森林			基于 GA-RF 的集成分类方法		
	敏感性	特异性	准确性	敏感性	特异性	准确性	敏感性	特异性	准确性
1	0.667	0.810	0.780	0.619	0.949	0.880	0.762	0.975	0.930
2	0.360	0.720	0.630	0.440	0.920	0.800	0.440	0.960	0.830
3	0.316	0.765	0.680	0.684	0.938	0.890	0.526	0.975	0.890
4	0.370	0.822	0.700	0.630	0.945	0.860	0.556	0.986	0.870
5	0.429	0.754	0.640	0.600	0.954	0.830	0.657	0.969	0.860
6	0.536	0.917	0.810	0.607	0.986	0.880	0.607	0.986	0.880
7	0.520	0.907	0.810	0.600	0.947	0.860	0.600	0.960	0.870
8	0.692	0.878	0.830	0.731	0.960	0.900	0.692	0.973	0.900
9	0.375	0.845	0.770	0.313	0.976	0.870	0.375	0.988	0.890
10	0.556	0.817	0.770	0.611	0.927	0.870	0.667	0.951	0.900
均值	0.482	0.823	0.742	0.583	0.950	0.864	**0.588**	**0.972**	**0.882**
标准差	0.132	0.065	0.073	0.121	0.020	0.030	**0.118**	**0.013**	**0.027**

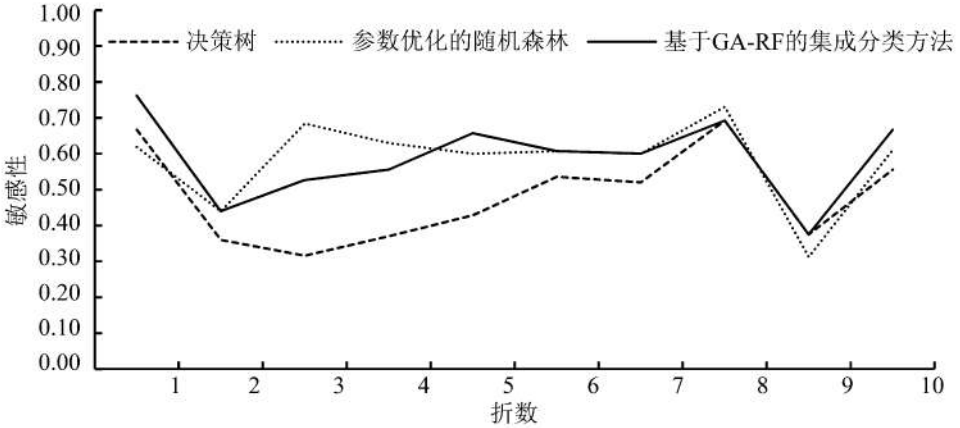


图 5 10 折交叉验证得到的敏感性对比图

Figure 5 Comparison of Sensitivity from 10-fold Cross-validation

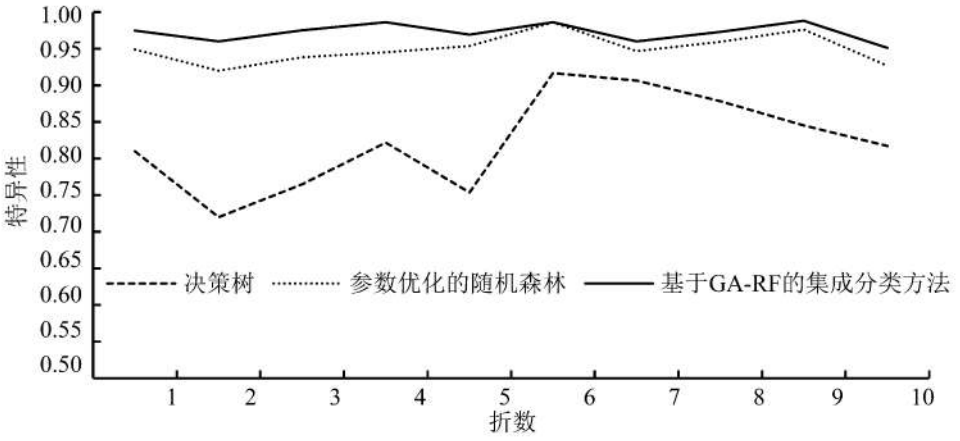


图 6 10 折交叉验证得到的特异性对比图

Figure 6 Comparison of Specificity from 10-fold Cross-validation

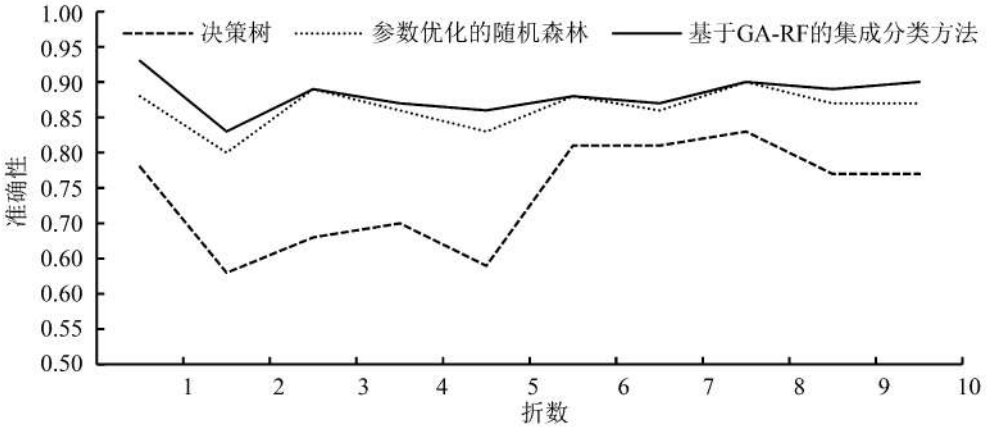


图7　10折交叉验证得到的准确性对比图

Figure 7　Comparison of Accuracy from 10-fold Cross-validation

准差分别为0.073、0.030、0.027,3个指标的标准差都呈现出由大到小的顺序。将两种集成学习方法与决策树方法比较发现,两种集成学习方法的3个指标值的均值都大于决策树方法,且有明显的差距,标准差也都小于决策树,这验证了集成学习方法通常比单一学习方法具有更高预测精度和更好泛化性能的预测。再比较两种集成学习方法,基于GA-RF的集成分类方法的3个指标的均值都更高,说明本研究所提方法的预测准确性更好;而且本研究所提方法得到的3个指标的标准差都更小,证明该方法独立于样本的泛化性能更强。

本研究的方法不仅预测精度最高,而且表现出了最好的泛化性能和稳定性,可以有效辅助医生做出治疗决策。敏感性值表征的是对可以存活的病人的分类正确率。如果分类错误,即将那些可以存活的病人预测为将会死亡,医生会及时对患者再次做出诊断,通过病理分析对癌症的转移或复发情况进行判断,排除复发或转移的疑虑或给出进一步的诊疗方案,因而低敏感性值可能导致多余的诊疗过程,而且会占用医院资源、增大治疗成本,但对癌症患者的健康是无害的。特异性值表征的是对不会存活的病人的分类正确率。如果分类错误,即将不会存活的病人预测为会存活,患者可能因没有得到及时的复诊而错过最佳的治疗时机,导致其存活时间缩短,这类错误关乎到患者的存活时长,所以需要尽可能地避免。本研究所提方法不仅在整体准确性上表现为最优,而且在敏感性和特异性方面也获得了更高的均值,相对于其他两种方法可以更大程度地避免上述两类错误的发生。特异性与敏感性相比,医院更看重特异性值的高低。本研究提出的方法的特异性平均值为0.972,达到了较高的水平,敏感性值仅有0.588,依然偏低。特异性已经达到比较理想的值,虽然敏感性值不高,但是由特异性和敏感性共同决定的准确性获得了较高的值,所以本研究的方法整体上可以达到较好的预测效果。

机器学习技术的发展及其在医学领域中的应用,使人们可以通过对历史数据的有效分析发现有趣的模式。医院常使用传统统计学方法做回顾性的数据分析,而机器学习技术可以辅助医生做前瞻性预测,当然准确和完整的医疗数据是必不可少的。SEER癌症数据受到美国国家癌症协会的严格机制保证,具有较好的准确性和完整性,因而基于该数据得到的预测模型具有一定的可靠性。

综合上述分析,与另外两种方法相比,本研究提出的集成分类方法在当前数据集下预测结直肠癌存活性的性能最优,无论医生更看重准确性、特异性、敏感性中的哪个指标,相对于决策树和随机森林,基于GA-RF的集成分类方法都是更好的选择。

(2)集成复杂度对比

很多集成学习方法都是将构造的所有个体学习器结合起来,随着集成复杂度的增加,预测精度得到提高,但运算速度却明显降低。表3给出10折交叉验

表3　决策树数量对比

Table 3　Comparison of Numbers of Decision Trees

折数	参数优化的随机森林	基于GA-RF的集成分类方法
1	39	41
2	191	46
3	30	47
4	115	42
5	83	39
6	471	41
7	38	44
8	32	43
9	339	43
10	277	40
均值	162	43

证中每一折参数优化的随机森林得到的决策树数量和基于GA-RF的集成分类器的决策树数量,从均值看,参数优化的随机森林的集成复杂度是基于GA-RF的集成分类方法的近4倍,说明本研究提出的方法不仅性能优于随机森林,而且集成复杂度更低,运算效率更高。

4 结论

癌症存活性预测的准确性对病患至关重要,针对结直肠癌存活性预测问题,为提高预测的准确性,本研究利用遗传算法改进随机森林,提出基于GA-RF的集成分类方法。在结直肠癌存活性预测中,通过与传统的决策树方法、参数优化的随机森林方法比较,得到3种方法在性能上从高到低的排序为:基于GA-RF的集成分类方法、参数优化的随机森林、决策树。一方面,该实验结果验证了个体分类器通过有效的集成可以得到比原个体分类器更好的分类准确率和泛化性能;另一方面,表明本研究对随机森林的改进十分有效,改进得到的方法在结直肠癌存活性预测中具有更好的性能。随机森林方法本身就有准确率高且泛化能力强的特点,本研究提出的方法在随机森林的基础上又提高了运算精度和泛化能力,而且由于集成复杂性的降低,运算速度也得到明显加快。

本研究使用实际的癌症数据进行实验,且所提出的方法预测性能较好,因而可以将该方法推广到医疗预后中辅助医生做出更准确的后续诊疗决策,弥补传统经验预测的不足,进而增加患者满意度、节约医疗资源、降低医疗成本。如医生可以在该方法的辅助下对癌症复发或癌症转移概率做出更准确的判断,及时给出精准的诊疗方案或减少不必要的后续诊疗。一方面,本研究结果为构造更优的集成提供了一种新的方法;另一方面,癌症存活性预测准确性的提高对于癌症的预测、治疗有很大的现实意义。但是本研究依然有不足之处,可以通过丰富实验数据使问题更接近实际,还可以通过分析并引入其他集成分类方法的优势以进一步提高方法的预测性能。

未来的研究中,考虑将本研究提出的癌症存活性预测方法扩展到如肺癌、前列腺癌、乳腺癌等其他癌症的预后中,以便辅助各种癌症做出更准确的诊疗决策,降低医疗成本。另外,在后续的研究中,可以将从存活性预测具体到存活期预测,并进一步地研究癌症转移、癌症复发的预测,以更好地辅助医疗决策。

参考文献:

[1] TORRE L A, BRAY F, SIEGEL R L, et al. Global cancer statistics, 2012. ***CA: A Cancer Journal for Clinicians***, 2015,65(2):87-108.

[2] KOUROU K, EXARCHOS T P, EXARCHOS K P, et al. Machine learning applications in cancer prognosis and prediction. ***Computational and Structural Biotechnology Journal***, 2015,13:8-17.

[3] PARK K, ALI A, KIM D, et al. Robust predictive model for evaluating breast cancer survivability. ***Engineering Applications of Artificial Intelligence***, 2013,26(9):2194-2205.

[4] AZAR A T, EI-METWALLY S M. Decision tree classifiers for automated medical diagnosis. ***Neural Computing and Applications***, 2013,23(7/8):2387-2403.

[5] CHEN K H, WANG K J, WANG K M, et al. Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. ***Applied Soft Computing***, 2014,24:773-780.

[6] DEVI M A, RAVI S, VAISHNAVI J, et al. Classification of cervical cancer using artificial neural networks. ***Procedia Computer Science***, 2016,89:465-472.

[7] LIN D, VASILAKOS A V, TANG Y, et al. Neural networks for computer-aided diagnosis in medicine: a review. ***Neurocomputing***, 2016,216:700-708.

[8] ZIęBA M, TOMCZAK J M, LUBICZ M, et al. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. ***Applied Soft Computing***, 2014,14(Part A): 99-108.

[9] AZAR A T, EI-SAID S A. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. ***Neural Computing and Applications***, 2014,24(5): 1163-1177.

[10] CHEN H, LIN Z, WU H, et al. Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. ***Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy***, 2015,135:185-191.

[11] AZAR A T, ELSHAZLY H I, HASSANIEN A E, et al. A random forest classifier for lymph diseases. ***Computer Methods and Programs in Biomedicine***, 2014,113(2):465-473.

[12] 彭勇,陈俞强. 基于概率神经网络的乳腺癌诊断系统. **合肥工业大学学报:自然科学版**, 2013,36(6):684-687.
PENG Yong, CHEN Yuqiang. Diagnosis system of breast cancer based on probabilistic neural network. ***Journal of Hefei University of Technology: Natural Science***, 2013,36(6):684-687. (in Chinese)

[13] SHEIKHPOUR R, SARRAM M A, SHEIKHPOUR R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. ***Applied Soft Computing***, 2016,40: 113-131.

[14] 李梅,张伟,李永忠,等. 支持向量机神经网络在判别前列腺癌中的应用研究. **四川大学学报:医学版**, 2013,44(4):666-668.
LI Mei, ZHANG Wei, LI Yongzhong, et al. Application of support vector machine neural network in distinguishing prostate cancer. ***Journal of Sichuan University: Medical Science Edition***, 2013,44(4):666-668. (in Chinese)

[15] GE P, GAO F, CHEN G. Predictive models for prostate cancer based on logistic regression and artificial neural network // ***Proceedings of IEEE International Conference on Mecha-***

tronics and Automation. Beijing, 2015:1472-1477.

[16] 叶强,张洁. 基于遗传算法的多分类器融合模型在信用评估中的应用. **哈尔滨工业大学学报**, 2006, 38(9): 1504-1505, 1536.
YE Qiang, ZHANG Jie. Application of multiple classifiers syncretizing model in credit evaluation. ***Journal of Harbin Institute of Technology***, 2006, 38(9): 1504-1505, 1536. (in Chinese)

[17] DELEN D, WALKER G, KADAM A. Predicting breast cancer survivability: a comparison of three data mining methods. ***Artificial Intelligence in Medicine***, 2005, 34(2): 113-127.

[18] DELEN D. Analysis of cancer data: a data mining approach. ***Expert Systems***, 2009, 26(1): 100-112.

[19] KHAN U, SHIN H, CHOI J P, et al. wFDT-weighted fuzzy decision trees for prognosis of breast cancer survivability // ***Proceedings of the Seventh Australasian Data Mining Conference***. Glenelg/ Adelaide, SA, 2008: 141-152.

[20] WANG K J, MAKOND B, CHEN K H, et al. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. ***Applied Soft Computing***, 2014, 20: 15-24.

[21] 周志华. **机器学习**. 北京:清华大学出版社, 2016: 178-180.
ZHOU Zhihua. ***Machine learning***. Beijing: Tsinghua University Press, 2016: 178-180. (in Chinese)

[22] ZOLBANIN H M, DELEN D, ZADEH A H. Predicting overall survivability in comorbidity of cancers: a data mining approach. ***Decision Support Systems***, 2015, 74: 150-161.

[23] RODRIGUEZ-GALIANO V F, GHIMIRE B, ROGAN J, et al. An assessment of the effectiveness of a random forest classifier for land-cover classification. ***ISPRS Journal of Photogrammetry and Remote Sensing***, 2012, 67: 93-104.

[24] QIAN C, WANG L, GAO Y, et al. In vivo MRI based prostate cancer localization with random forests and auto-context model. ***Computerized Medical Imaging and Graphics***, 2016, 52: 44-57.

[25] 周天宁,明冬萍,赵睿. **参数优化随机森林算法的土地覆盖分类**. 北京:中国石油天然气集团, 2016.
ZHOU Tianning, MING Dongping, ZHAO Rui. ***Science of surveying and mapping, parameter optimization algorithm of random forest land***. Beijing: China National Petroleum Corporation, 2016. (in Chinese)

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

[27] ZHOU Z H, WU J, TANG W. Ensembling neural networks: many could be better than all. ***Artificial Intelligence***, 2002, 137(1/2): 239-263.

[28] ZHOU Z H, TANG W. Selective ensemble of decision trees // ***RSFDGrC'03 Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing***. Heidelberg: Springer-Berlin, 2003: 476-483.

[29] 张妤,王文剑,康向平. 一种回归 SVM 选择性集成方法. **计算机科学**, 2008, 35(4): 178-180.
ZHANG Yu, WANG Wenjian, KANG Xiangping. A regression SVM selection ensemble approach. ***Computer Science***, 2008, 35(4): 178-180. (in Chinese)

[30] HONG J H, CHO S B. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. ***Artificial Intelligence in Medicine***, 2006, 36(1): 43-58.

[31] 赵强利,蒋艳凰,徐明. 选择性集成算法分类与比较. **计算机工程与科学**, 2012, 34(2): 134-138.
ZHAO Qiangli, JIANG Yanhuang, XU Ming. Categorization and comparison of the ensemble pruning algorithms. ***Computer Engineering & Science***, 2012, 34(2): 134-138. (in Chinese)

[32] ZHOU Z H. ***Ensemble methods: foundations and algorithms***. Boca Raton: CRC Press, 2012: 72-73.

[33] DIETTERICH T G. Ensemble methods in machine learning // ***Multiple Classifier Systems***. Cagliari, Italy: Springer-Verlag Berlin Heidelberg, 2000: 1-15.

[34] BREIMAN L. Random forests. ***Machine Learning***, 2001, 45(1): 5-32.

[35] MARSLAND S. ***Machine learning: an algorithmic perspective***. Boca Raton, Florida: CRC Press, 2009: 33.

[36] ELYAN E, GABER M M. ***A genetic algorithm approach to optimising random forests applied to class engineered data***. Aberdeen: Robert Gordon University, 2016.

[37] BOUSSAÏD I, LEPAGNOT J, SIARRY P. A survey on optimization metaheuristics. ***Information Sciences***, 2013, 237: 82-117.

[38] WHITLEY D. A genetic algorithm tutorial. ***Statistics and Computing***, 1994, 4(2): 65-85.

[39] MITCHELL M. ***An introduction to genetic algorithms***. Cambridge, MA: MIT Press, 1996: 7-8.

[40] 陈李钢,叶强,李一军. 基于遗传算法的银行客户信用评估模型研究. **计算机工程**, 2007, 33(3): 70-72.
CHEN Ligang, YE Qiang, LI Yijun. Research on GA-based bank customer's credit evaluation. ***Computer Engineering***, 2007, 33(3): 70-72. (in Chinese)

[41] LOZANO M, LAGUNA M, MARTÍ R, et al. A genetic algorithm for the minimum generating set problem. ***Applied Soft Computing***, 2016, 48: 254-264.

[42] DELEN D, OZTEKIN A, TOMAK L. An analytic approach to better understanding and management of coronary surgeries. ***Decision Support Systems***, 2012, 52(3): 698-705.

[43] PIRAMUTHU S. On learning to predict web traffic. ***Decision Support Systems***, 2003, 35(2): 213-229.

[44] ZHU W, ZENG N, WANG N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations // ***Northeast SAS Users Group Proceedings: Health Care and Life Sciences***. Baltimore, Maryland, 2010: 1-9.

[45] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection // ***IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence***. San Francisco, CA: Morgan Kaufmann Publishers, 1995: 1137-1143.

Predicting Survivability of Colorectal Cancer by an Ensemble Classification Method Improved on Random Forest

WANG Yuyan[1], WANG Dujuan[1], WANG Yanzhang[1], Yaochu Jin[1,2]

1 Faculty of Management and Economics, Dalian University of Technology, Dalian 116023, China

2 Department of Computing, University of Surrey, Surrey GU2 7XH, United Kingdom

Abstract: Cancer is one of the major causes of death for human and accounts for a large proportion of the costs of healthcare in many countries. The prediction of cancer survivability is an important task for cancer prognosis and has been a challenging research problem for many researchers, which can help doctors to make more accurate diagnostic and treatment decisions and lower treatment costs. In recent years, data-driven methods for cancer survivability prediction have been gradually put into application, yet improving the accuracy of cancer survivability prediction methods has always been an active area of research as the accuracy of prediction is the main index to evaluate the performance of prediction methods.

This paper focuses on colorectal cancer which has both high incidence and high mortality. In order to make survivability prediction of colorectal cancer more accuracy, an ensemble classification method based on GA-RF is proposed. This method is the outcome of using genetic algorithm(GA for short) to make improvements to the random forest(RF for short). Genetic algorithm is used to search for parts of the decision trees in random forest aiming at getting better accuracy of ensemble classification. The method proposed along with decision tree method and the random forest method after parameter optimization are used to develop prediction models to predict the survivability of patients with colorectal cancer. Using the colorectal cancer data set of the SEER database, experiments are carried out with three methods which are tested by 10-fold cross-validation for performance comparison purposes, and then accuracy, sensitivity and specificity are used to evaluate the three methods.

The experimental results indicated that the ensemble classification method based on GA-RF had the prediction accuracy of 88.2%, higher than that of the random forest after parameter optimization and decision tree. And random forest which came out to be the second also had a high accuracy of 86.4%, but the complexity of ensemble was much more than that of the ensemble classification method based on GA-RF, and decision tree came out to be the worst of the three with 74.2% accuracy. Besides, the ensemble classification method based on GA-RF showed the best generalization ability.

The ensemble classification method proposed makes an effective improvement on random forest, which can predict survivability of colorectal cancer with higher efficiency and accuracy, provide reference for decision-making of colorectal cancer prognosis, make up for the shortage of survivability prediction based on experience, and has practical significance to saving medical resources, reducing medical costs and improving patient satisfaction.

Keywords: random forest; genetic algorithm; ensemble classification; survivability prediction; colorectal cancer

Received Date: September 10th, 2016 **Accepted Date**: December 22nd, 2016

Funded Project: Supported by the National Natural Science Foundation of China(71533001) and the Fundamental Research Funds for the Central Universities(DUT15QY32)

Biography: WANG Yuyan is a master degree in the Faculty of Management and Economics at Dalian University of Technology. Her research interests focus on medical health management, data mining and machine learning. E-mail: wyy@mail.dlut.edu.cn

WANG Dujuan, doctor in engineering, is an associate professor in the Faculty of Management and Economics at Dalian University of Technology. Her research interests include service operation management, data mining and intelligent optimization algorithm. Her representative paper titled "Disruption management for new jobs arrivals with deteriorating effect and controllable processing times" was published in the ***Journal of Systems & Management***(Issue 5, 2016). E-mail: wangdujuan@dlut.edu.cn

WANG Yanzhang, doctor in engineering, is a professor in the Faculty of Management and Economics at Dalian University of Technology. His research interests include data mining and knowledge management. His representative paper titled "Emergency risk entropy forecasting model based on knowledge element" was published in the ***Journal of Systems Engineering***(Issue 1, 2016). E-mail: yzwang@dlut.edu.cn

Yaochu Jin, doctor in engineering, is a professor and a chair of computational intelligence in the Department of Computing at University of Surrey. His research interests include computational intelligence, machine learning, computational biology and computational neuroscience and other interdisciplinary research and engineering applications. His representative paper titled "A social learning particle swarm optimization algorithm for scalable optimization" was published in the ***Information Sciences***(Volume 291, 2015). E-mail: yaochu.jin@surrey.ac.uk □