第31卷第1期 管理科学 ISSN 1672-0334 Vol. 31 No. 1 33-47
2018年1月 Journal of Management Science January, 2018

基于GMM-RBF神经网络的前列腺癌诊断方法

崔少泽[1]，王杜娟[1]，王苏桐[1]，夏江南[1]，王延章[1]，JIN Yaochu[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：前列腺癌是近年来发病率上升速度最快的男性癌症，严重威胁着患者的身体健康，准确地判断癌症患者的患病情况对于节约医疗资源、提高患者满意度起着至关重要的作用。近年来，基于数据挖掘的癌症诊断方法逐渐成为疾病诊断领域的研究热点，在提高诊断准确性上显示出极大优势。

针对现有前列腺癌早期诊断方法准确性不高的问题，提出一种基于高斯混合模型改进径向基函数神经网络的前列腺癌诊断方法——GMM-RBF神经网络方法。该方法通过使用高斯混合模型对径向基函数神经网络中径向基函数的参数进行预训练，使模型避免陷入局部最优，之后采用改进的粒子群优化算法对神经网络进行训练。采用国家临床医学科学数据中心提供的数据进行前列腺癌诊断实验，将所提出的方法与径向基神经网络、分类回归树、支持向量机和逻辑回归等主流的机器学习算法进行对比，并使用准确性、特异性、敏感性和*AUC*值对模型的性能进行评价。

研究结果表明，与改进前的神经网络模型相比，GMM-RBF神经网络模型收敛速度更快、初始准确度更高；与其它机器学习算法相比，GMM-RBF神经网络模型在10折交叉验证中取得了较高的准确性、敏感性、特异性和*AUC*值。

GMM-RBF神经网络方法在模型预测精度上比传统的径向基函数神经网络模型有很大提升，能够得到更为可靠的前列腺癌诊断结果，为医疗工作者初步诊断前列腺癌和穿刺活检操作提供有效的辅助决策支持，该方法的提出对于减少患者痛苦、提高患者满意度和节约医疗资源具有实际意义。

关键词：前列腺癌；径向基函数神经网络；高斯混合模型；粒子群优化算法；疾病诊断
中图分类号：TP183 **文献标识码**：A **doi**：10.3969/j.issn.1672-0334.2018.01.003
文章编号：1672-0334(2018)01-0033-15

收稿日期：2017-09-19 **修返日期**：2017-12-27
基金项目：国家自然科学基金(71533001,71672019,71271039)

作者简介：崔少泽，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、机器学习和智能优化算法等，E-mail：csz2016@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
王苏桐，大连理工大学管理与经济学部硕士研究生，研究方向为机器学习和规则挖掘等，E-mail：sutongwang@mail.dlut.edu.cn
夏江南，大连理工大学管理与经济学部硕士研究生，研究方向为图像识别和文本挖掘等，E-mail：jn_xia@foxmail.com
王延章，工学博士，大连理工大学管理与经济学部教授，研究方向为电子政务和知识管理等，代表性学术成果为“模型管理的知识及其表示方法”，发表在2011年第6期《系统工程学报》，E-mail：yzwang@dlut.edu.cn
JIN Yaochu，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，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

在现代社会中,前列腺癌已经成为致死率极高的疾病。2008年世界卫生组织统计全年前列腺癌病例超过90万人,其中约有26万的男性患者最终死亡[1]。JEMAL et al.[2]在2011年做的一项全球癌症统计研究中表明,前列腺癌的发病率在男性癌症中排第2位。在中国前列腺癌的发病率也在逐年上升,从2000年的第10位升至2011年的第6位,成为上升速度最快的男性癌症类型[3]。

在临床上,前列腺癌需要经过穿刺活检才能够进行确诊,但由于穿刺活检会对患者的身体造成损伤,且通常情况下进行穿刺活检的患者有近50%左右检查结果为阴性,即该病人未患前列腺癌[4]。在实际医疗诊断中,为降低上述过程对未患癌患者造成的损伤,医生会在穿刺活检之前,通过直肠超声检查、直肠指诊和观察血液中前列腺特异抗原浓度进行初步判断,确定是否需要为患者安排穿刺活性检查。这些检查中前列腺特异抗原浓度是进行前列腺癌初步诊断的重要指标,临床上认为前列腺特异抗原浓度在4ng/ml以下为正常水平,前列腺特异抗原的浓度越高,患者患有前列腺癌的风险越大[5]。然而,由于其他前列腺疾病也可能引起前列腺特异抗原水平的升高,所以不能单纯依据前列腺特异抗原水平对患者进行确诊。

随着大数据时代的到来,数据驱动的电子健康服务管理研究成为新的热门领域[6-7]。由于之前电子健康数据不完善,样本有限,所以医院临床上多采用传统的统计学方法进行实验分析,较少使用数据挖掘方法。随着数据的增长、数据存储的规范化以及数据挖掘技术的发展,越来越多的数据挖掘方法在医疗领域中得到广泛应用,如决策树[8-9]、支持向量机[10-11]和人工神经网络[12-13]等方法均有所使用。针对前列腺癌早期诊断这一问题,本研究提出使用GMM-RBF神经网络方法对前列腺癌症患者进行诊断,该方法在使用径向基函数(radial basis function,RBF)神经网络进行前列腺癌诊断之前,使用高斯混合模型(Gaussian mixture model,GMM)对径向基函数神经网络的径向基函数初始参数进行预训练,用优化后的参数代替随机初始化参数,从而减少模型训练时陷入局部最优的可能性,并使用改进的粒子群优化(particle swarm optimization,PSO)算法对网络进行训练。通过与其他几种流行算法在实际数据集上进行实验对比,发现本研究模型在前列腺癌诊断中具有更高的准确性。

1 相关研究评述

本研究提出使用高斯混合模型对径向基函数神经网络进行预训练的方法进行前列腺癌诊断的预测,下面从前列腺癌诊断方法研究和神经网络的改进研究两个方面介绍该领域的相关工作。

1.1 前列腺癌诊断方法

前列腺癌是男性泌尿系统常见的恶性肿瘤,近年来其发病率呈现逐年上升趋势[3]。已有研究表明,直肠指检、前列腺特异抗原指标、经直肠超声、核磁共振成像(MRI)和前列腺穿刺活检等技术提高了前列腺癌的早期发现比例,但仍然有10%~15%的前列腺癌被漏诊[14]。虽然前列腺特异抗原可以作为前列腺癌特异性肿瘤标志物,但也有局限性[15-17]。有研究表明,血清总前列腺特异抗原(tPSA)和前列腺特异抗原密度(PSAD)对前列腺癌有较高的诊断价值[18],但只依据前列腺特异抗原浓度并不能对前列腺癌进行准确诊断,需要结合与诊断结果有关的多种特征来提高诊断的准确性。

在前列腺癌的初步诊断阶段,运用机器学习方法构建的诊断模型为医疗工作者对患者是否进行穿刺活检操作提供了有效的决策辅助和方法支持。机器学习技术能够结合多种数据特征,利用历史数据训练出反映前列腺癌诊断过程的模型,为提高前列腺癌诊断的准确性提供帮助。LEE et al.[19]针对前列腺癌的诊断问题,使用逻辑回归(logistic regression,LR)算法,利用病人的年龄、前列腺特异抗原、直肠指诊和超声检查这些特征对前列腺癌诊断结果进行预测;FINNE et al.[20]使用逻辑回归方法,对1 775名年龄在55岁~67岁的男性患者接受前列腺癌诊断的结果进行预测,实验结果表明,相对于采用单一特征的模型,采用多种特征的分类模型具有更高的准确性,能够减少未患病者接受穿刺活检的概率;BERMEJO et al.[21]研究前列腺癌和良性前列腺增生的诊断问题,采用决策树和逻辑回归算法,利用病人的年龄、前列腺特异抗原和直肠指诊3个指标构建诊断模型,实现了对这两种疾病的有效识别。

但是上述方法仍难以描述多个输入特征与输出结果之间复杂的非线性关系,且存在实际使用中诊断准确性较低的问题,因此需要使用准确性更高的分类方法进行前列腺癌的诊断。

1.2 神经网络的改进

在众多的机器学习分类方法中,人工神经网络(artificial neural network,ANN)能够结合多种特征,并对输入数据与输出结果之间复杂的非线性关系进行准确描述,该方法在前列腺癌诊断领域受到关注[22]。SNOW et al.[23]针对病人的穿刺活检结果预测问题,使用人工神经网络方法进行前列腺癌诊断结果的预测,实验结果显示该方法的准确率达到87%;BABAIAN et al.[24]使用人工神经网络方法对151位接受穿刺活检病人的检查数据进行训练,得到的模型在诊断准确性上高于仅使用游离前列腺特异抗原浓度进行前列腺癌诊断的方法;STEPHAN et al.[25]为预测病人被诊断为前列腺癌的风险,基于928位病人的前列腺特异抗原、游离前列腺特异抗原、年龄、前列腺体积和直肠指诊数据,使用人工神经网络训练前列腺癌诊断模型,并利用该模型对1 188位病人的诊断结果进行预测,实验结果表明人工神经网络在前列腺癌诊断问题上具有有效性。

传统人工神经网络使用Sigmoid函数作为隐含层

的激活函数,层与层之间的连接权重较多,所以存在训练时间长、容易陷入局部极小的问题。为克服传统人工神经网络的不足,MOODY et al.[26]在1989年提出径向基函数神经网络,径向基函数神经网络使用径向基函数作为隐含层激活函数,能以任意精度逼近任意非线性关系,解决了传统人工神经网络预测准确率不足、容易陷入局部极小的问题[27]。目前已有研究尝试将径向基函数神经网络应用于前列腺癌诊断问题,MARÍN et al.[28]针对前列腺癌的诊断问题,使用两种类型的径向基函数神经网络方法对病人进行分类,结果表明两种类型的径向基函数神经网络预测准确率均高于目前在医疗领域流行的多层感知器方法。然而WALLACE et al.[29]认为对径向基函数神经网络模型初始参数进行预训练可以提高模型的准确性,提升网络训练的收敛速度。

径向基函数神经网络通常采用反向传播算法(back propagation,BP)进行神经网络的训练[30]。然而反向传播训练神经网络存在收敛速度较慢、可能陷入局部极小的问题[31]。肖斌卿等[32]结合遗传算法和反向传播算法对神经网络进行训练,取得了优于反向传播神经网络的拟合精度。由于遗传算法存在搜索速度较慢的缺点[33],王亮等[34]使用粒子群优化算法对反向传播网络的初始权重进行优化,对比反向传播网络,PSO-BP模型取得了更高的预测精度。但是多维优化背景下,标准粒子群优化算法存在早熟、可能陷入局部最优的问题[33]。为克服这个缺陷,本研究提出一种带随机初始化策略的改进粒子群优化算法对网络进行训练。

本研究针对提高前列腺癌初步诊断准确性的问题,提出GMM-RBF神经网络方法,采用高斯混合模型对输入数据实例的特征进行训练,将训练得到的高斯函数作为径向基函数神经网络隐含层节点中的初始基函数,然后使用输入数据实例训练径向基函数神经网络模型;针对反向传播算法在径向基函数神经网络模型训练过程中存在计算复杂、收敛速度较慢的问题,采用改进的粒子群优化算法对径向基函数神经网络模型中的参数进行训练,简化计算过程,提高模型训练效率;最后使用国家临床医学科学数据中心提供的前列腺疾病检查数据进行实验,检验提出的方法在前列腺癌实际诊断中的有效性。

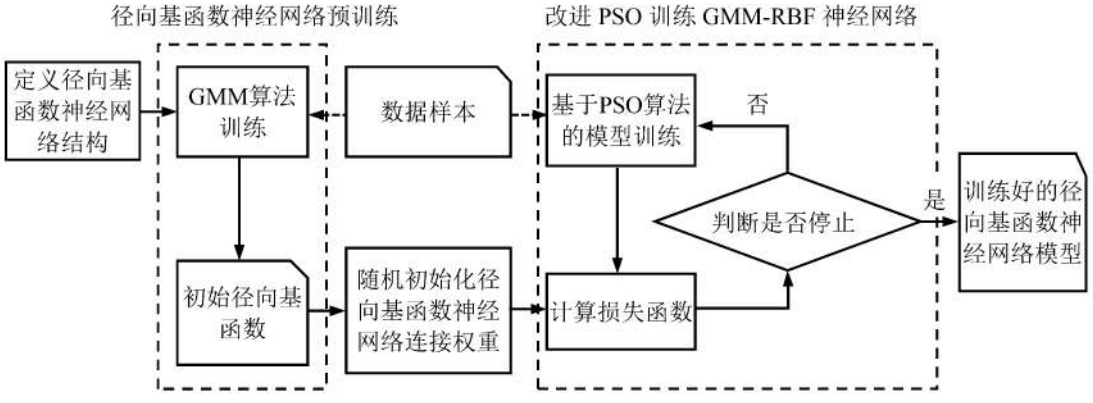
2　基于GMM-RBF神经网络的前列腺癌诊断

为了能够提高对前列腺癌患者诊断的准确性,本研究在径向基函数神经网络的基础上提出一种径向基函数神经网络的改进算法——GMM-RBF神经网络。针对径向基函数神经网络模型中存在的隐含层径向基函数初始参数设置问题,本研究提出使用高斯混合模型对输入数据进行训练,从而确定径向基函数的初始参数。径向基函数神经网络权重的训练也是关键环节,为减少计算复杂度,加快训练的收敛速度,本研究使用改进的粒子群优化算法进行权重训练,并且对隐含层径向基函数的参数进行编码寻优,整体的GMM-RBF神经网络构建流程见图1。该方法主要由2个部分构成,①定义径向基函数神经网络结构,使用高斯混合模型对径向基函数神经网络进行预训练,得到初始的径向基函数;②随机初始化GMM-RBF神经网络连接权重后,使用改进的粒子群优化算法进行参数优化,达到终止条件后输出训练好的GMM-RBF神经网络模型。下面将对以上两个部分进行详细介绍。

2.1　GMM-RBF神经网络

2.1.1　径向基函数神经网络

传统反向传播神经网络中常常出现收敛过分依赖初值和局部收敛的问题,针对此类问题,MOODY et al.[26]在20世纪80年代末提出径向基网络,它是以径向基函数作为隐含层神经元激活函数的3层前向型神经网络,具有较快的运算速度、较强的非线性映射能力和较好的预测能力。径向基函数神经网络具有3层结构,包括输入层、隐含层和输出层,网络模型的拓扑结构见图2。径向基函数神经网络包含D个输入层节点、H个隐含层节点和1个输出层节点,输入层节点x_i对应输入数据实例的D维特征,$i=1,2,\cdots,D$,在网络中起到传输信号的作用,输入层节点与隐含层节点之间可以看作连接权值u_{ih}为1的连接。隐含



注:实线箭头表示模型的构建过程,虚线箭头表示数据样本在对应步骤中进行使用。

图1　GMM-RBF神经网络构建过程

Figure 1　Construction Process for GMM-RBF Neural Network

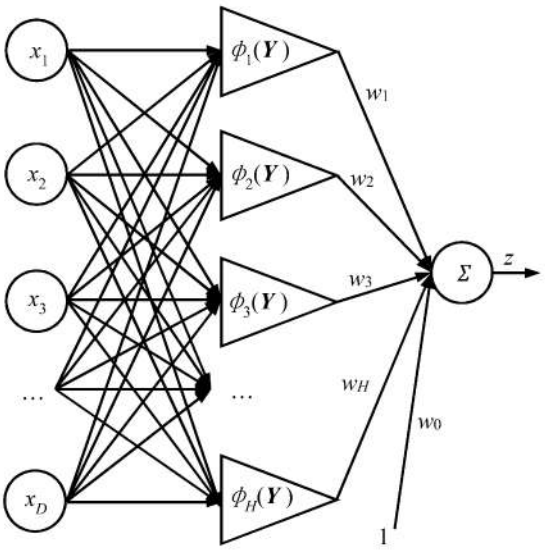


图2 径向基函数神经网络结构

Figure 2 Structure of RBF Neural Network

层节点 ϕ_h 的输入数据为向量 $\boldsymbol{Y}_h$，$\boldsymbol{Y}_h=(y_{1h},y_{2h},\cdots,y_{Dh})$，$h=1,2,\cdots,H$，$y_{ih}$ 通过(1)式计算得到，即

$$y_{ih}=u_{ih}x_i \tag{1}$$

隐含层的每一个节点都使用非线性函数 $\varphi(\cdot)$ 作为激活基函数，对输入数据进行非线性变换。在众多可选的径向基函数中，高斯函数作为径向基函数神经网络的激活函数常表现出很好的效果[35]，高斯函数的表达式为

$$\varphi_h(\boldsymbol{Y})=\exp\left(-\frac{||\boldsymbol{Y}_h-\boldsymbol{\mu}_h||^2}{\sigma_h}\right) \tag{2}$$

其中，φ_h 为隐含层节点 ϕ_h 中的激活基函数，φ_h 的输入向量为 $\boldsymbol{Y}_h$，$\boldsymbol{\mu}_h$ 和 σ_h 为高斯函数 φ_h 中的参数，$||\cdot||$ 为两个向量的欧氏距离。径向基函数神经网络的输出层对隐含层的输出进行加权汇总，作为输出层节点的输入值，即

$$z=\sum_{h=1}^{H}\varphi_h(\boldsymbol{Y})w_h+w_0 \tag{3}$$

其中，z 为输出层节点 Σ 汇总结果，w_h 为隐含层节点 ϕ_h 与输出层节点间的权重，w_0 为偏倚权重。对于分类问题，由于输出为离散型数值，输出节点通常采用 Sigmoid 函数作为激活函数将输出值映射到(0,1)取值区间内，使输出值代表取值为1的概率。因此径向基函数神经网络输出层节点的输出结果由 z 转换为 Out，即

$$Out=\frac{1}{1+e^{-z}} \tag{4}$$

由于隐含层节点激活函数的参数对径向基函数神经网络模型的预测准确性有很大影响，因此在训练之前得到较好的初始参数取值十分重要。为解决其初始参数的问题并提高模型的准确性，本研究采用高斯混合模型算法对输入数据实例进行预训练，以得到经过高斯混合模型优化的径向基函数的初始参数取值。

2.1.2 高斯混合模型

在使用径向基函数神经网络模型对输入数据实例进行分类之前，可以利用高斯混合模型对神经网络模型中的参数进行预训练，即将高斯混合模型的训练结果作为径向基函数的初始参数。使用经过训练的初始参数代替随机选择的初始参数，可以减少初始参数选择对最终训练结果的影响，使得到的径向基函数神经网络模型准确性更高[36]。根据REYNOLDS et al.[37]的研究，高斯混合模型假设训练样本在空间中存在簇结构，即由几个不同类的数据组成，每个类 j 服从已知的概率分布 $\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)$，$j=1,2,\cdots,K$，$\theta$ 为概率密度函数的参数集合。设样本空间存在 K 个类，$\boldsymbol{X}^{(l)}$ 为样本，$l=1,2,\cdots,N$，N 为数据集的总样本数。样本在空间中出现的概率可通过(5)式进行估计，即

$$p(\boldsymbol{X}^{(l)}|\theta)=\sum_{j=1}^{K}\alpha_j\varphi_j(\boldsymbol{X}^{(l)}|\theta_j) \tag{5}$$

其中，α_j 为混合系数，代表由分布 φ_j 产生样本 $\boldsymbol{X}^{(l)}$ 的概率，$\sum_{j=1}^{K}\alpha_j=1$；$\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)$ 通常为均值为 $\boldsymbol{\mu}_j$、标准差为 σ_j 的高斯分布，其概率密度函数的表达式为

$$\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)=\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(-\frac{||\boldsymbol{X}^{(l)}-\boldsymbol{\mu}_j||^2}{2\sigma_j^2}\right) \tag{6}$$

高斯混合模型的参数通常采用期望最大化(expectation maximization，EM)算法进行训练[38]。使用高斯混合模型时，要找到一组参数 θ_j，使生成已有数据点的概率最大，这一概率可以表示为似然函数，即

$$\prod_{i=1}^{N}p(\boldsymbol{X}^{(l)}|\theta) \tag{7}$$

通常单个点概率很小，相乘之后易造成浮点数下溢，因此取对数，得

$$\sum_{i=1}^{N}\log\left[\sum_{j=1}^{K}\alpha_j\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)\right] \tag{8}$$

(8)式为高斯混合模型对数似然函数表达式。为取得(8)式的最大值，本研究使用期望最大化算法寻找式中最佳的模型参数 α_j、$\boldsymbol{\mu}_j$ 和 σ_j。使用期望最大化算法进行高斯混合模型参数估计的详细步骤如下。

步骤1 使用K-means算法确定初始 K 个类的类中心所在位置，即均值 $\boldsymbol{\mu}_j$。

步骤2 计算混合系数 α_j，α_j 为样本出现在 K 个聚类中第 j 类的概率，$\alpha_j=\frac{N_j}{N}$，N_j 为属于第 j 类的样本总数。

步骤3 计算每一类的标准差 σ_j，$\sigma_j=\frac{1}{N_j}\sum(\boldsymbol{x}-\boldsymbol{\mu}_j)(\boldsymbol{x}-\boldsymbol{\mu}_j)^T$，$\boldsymbol{x}$ 为属于第 j 类的样本。

步骤4 计算第 l 个样本 $\boldsymbol{X}^{(l)}$ 由第 j 类产生的概率 γ_{lj}，$\gamma_{lj}=\frac{\alpha_j\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)}{\sum_{j=1}^{K}\alpha_j\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)}$。

步骤5　更新聚类中心所在位置$\boldsymbol{\mu}_j$，$\boldsymbol{\mu}_j=\dfrac{\sum_{l=1}^{N}\gamma_{lj}\boldsymbol{X}^{(l)}}{\sum_{l=1}^{N}\gamma_{lj}}$。

步骤6　重复步骤2~步骤5，达到高斯混合模型最大迭代次数T_g停止。

2.1.3　GMM-RBF神经网络训练过程

使用高斯混合模型算法确定径向基函数神经网络初始径向基函数的参数后，本研究采用改进的粒子群优化算法进行模型训练，提升模型训练效率。下面对GMM-RBF神经网络的训练过程和关键环节进行详细叙述。

步骤1　定义径向基函数神经网络结构Θ，包括径向基函数神经网络的输入层节点数量D、隐含层节点数量H和输出层节点数量O，即$\Theta=\{D,H,O\}$。输入层节点数量D与前列腺癌诊断的输入特征相同；模型输出的是诊断结果，采用0和1表示，0为未患前列腺癌，1为患前列腺癌，因此输出层节点数量$O=1$；隐含层节点数量H的取值可通过多次试验进行确定。

步骤2　高斯混合模型算法训练。使用高斯混合模型算法对输入的数据样本进行聚类，由于高斯混合模型算法的结果将作为径向基函数的初始参数，聚类中心数K要与神经网络隐含层节点数H相同，因此设置$K=H$。通过高斯混合模型算法的训练过程，利用期望最大化算法对模型中的参数α_j和θ_j进行极大似然估计[39]，训练后的模型可以表示为K个高斯分布φ_j，这些高斯分布的均值$\boldsymbol{\mu}_j$和标准差σ_j将作为径向基函数的初始参数。

步骤3　随机初始化径向基函数神经网络连接权重。步骤1中已经确定径向基函数神经网络的结构，在进行径向基函数神经网络训练之前，将径向基函数神经网络的输入层与隐含层之间连接权重u_{ih}置为1，并随机初始化隐含层节点与输出层节点之间连接权重w_h。

步骤4　计算损失函数。通过损失函数可以评价当前迭代次数t的径向基函数神经网络模型W_t的预测结果与真实情况的偏差，即模型损失。这里采用基于准确率的模型损失度量方式，设训练样本$\boldsymbol{X}^{(l)}$的实际目标值为$o^{(l)}$，使用模型W_t的输出值为$\delta_t^{(l)}$，则径向基函数神经网络模型W_t在整个训练数据集上的损失函数为F_t，即

$$F_t=1-\frac{TP+TN}{P_{num}+N_{num}} \tag{9}$$

其中，TP为实际为患者且预测为患病的人数，TN为实际为未患病且预测也为未患病的人数，P_{num}为实际患病的人数，N_{num}为实际未患病的人数。当样本$\boldsymbol{X}^{(l)}$的$\delta_t^{(l)}$大于等于0.5且$o^{(l)}$为1时，TP增加1；当样本$\boldsymbol{X}^{(l)}$的$\delta_t^{(l)}$小于0.5且$o^{(l)}$为0时，TN增加1。由于本研究采取目标最小化的方式进行优化，$\dfrac{TP+TN}{P_{num}+N_{num}}$的取值区间为[0,1]，所以使用$1-\dfrac{TP+TN}{P_{num}+N_{num}}$将最大化问题转换成最小化问题。(9)式在PSO算法中作为适应度函数计算每个粒子的适应度。

步骤5　基于改进粒子群优化算法的模型训练。根据径向基函数神经网络模型W_t的损失函数F_t，通过迭代的方式对W_t中的参数进行训练。需要训练的参数包括神经网络节点间权重w_h和隐含层节点中的径向基函数$\varphi_j(\boldsymbol{X}^{(l)}|\theta_j)$。本研究提出改进粒子群优化算法对上述参数进行迭代学习，直至达到最大的迭代次数。

下面对使用改进粒子群优化算法训练GMM-RBF神经网络模型的过程进行详细叙述。

2.2　改进粒子群优化算法训练GMM-RBF神经网络

为解决采用反向传播算法在GMM-RBF神经网络模型训练过程中计算复杂、收敛较慢的问题，本研究提出使用改进粒子群优化算法对GMM-RBF神经网络模型的参数进行训练。粒子群优化算法将种群中的每个粒子的位置作为优化问题的一个候选解，每个粒子都对应一个由目标函数决定的适应度值[40]。在算法迭代过程中，粒子根据自身和其他粒子的位置，调整自身的速度和位置，逐步接近自身的最优位置。而算法整体迭代中会不断搜索种群中位置最优的粒子，直到找到满足条件的最优解，每次迭代中，粒子根据(10)式和(11)式更新自身的速度和位置，即

$$\hat{\boldsymbol{v}}_m^{t+1}=\omega\hat{\boldsymbol{v}}_m^t+c_1r_1(\vec{\boldsymbol{p}}_m^t-\hat{\boldsymbol{s}}_m^t)+c_2r_2(\vec{\boldsymbol{g}}^t-\hat{\boldsymbol{s}}_m^t) \tag{10}$$

$$\hat{\boldsymbol{s}}_m^{t+1}=\hat{\boldsymbol{s}}_m^t+\hat{\boldsymbol{v}}_m^{t+1} \tag{11}$$

$$m=1,2,\cdots,N_p$$

其中，$\hat{\boldsymbol{v}}_m^t$为第m个粒子在第t次迭代中的速度，$\hat{\boldsymbol{s}}_m^t$为第m个粒子在第t次迭代中的位置，$\vec{\boldsymbol{p}}_m^t$为第m个粒子在第t次迭代中的历史最优位置，$\vec{\boldsymbol{g}}^t$为所有粒子第t次迭代中的全局最优位置，N_p为种群规模；ω为惯性权重；c_1和c_2为加速因子，为非负常数；r_1和r_2为[0,1]之间的随机数。对于粒子群优化过程中粒子向自身和全局的历史最佳位置聚集容易造成结果陷入局部最优的情况，可以引入变异策略[41]，即对粒子以一定的概率重新初始化，以提高算法找到全局最优值的概率。

在使用改进粒子群优化算法训练GMM-RBF神经网络模型时，首先将模型中的隐含层节点与输出层节点间连接权重w_h以及径向基函数φ_j中对应的均值$\boldsymbol{\mu}_j$和标准差σ_j作为位置编码包含在粒子群优化算法的每一个粒子中，粒子的位置编码代表一个候选的神经网络模型。粒子群优化粒子的位置编码见图3。

粒子群优化算法训练GMM-RBF神经网络的流程见图4，具体步骤如下。

步骤1　粒子群优化算法参数初始化。具体参数包括种群规模N_p、最大迭代次数T、惯性因子ω以及粒子的位置取值区间$[S_{\min},S_{\max}]$和速度取值区间$[V_{\min},V_{\max}]$。

步骤2　初始化每个粒子的速度和位置。在粒子的初始位置编码$\hat{\boldsymbol{s}}_m^0$中，径向基函数的均值$\boldsymbol{\mu}_j$和标准差σ_j已经由高斯混合模型算法确定，此处仅需要设置粒子的初始速度$\hat{\boldsymbol{v}}_m^0$和位置$\hat{\boldsymbol{s}}_m^0$编码中代表隐含层节

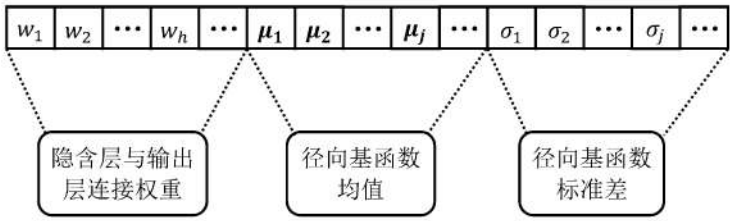


图3 粒子群优化粒子位置编码

Figure 3 Encoding for PSO Particle Location

点与输出层节点之间连接权重 w_h 的部分为取值区间[$V_{\min}$,$V_{\max}$]和[$S_{\min}$,$S_{\max}$]内随机值。

步骤3 计算每个粒子的适应度。利用粒子位置 $\hat{s}_m$ 对GMM-RBF神经网络的系数赋值,使用该系数取值下的GMM-RBF神经网络对每个训练样本进行估计,并根据(9)式计算每个粒子的适应度 fit_m。

步骤4 寻找个体最优位置和全局最优位置。根据本次迭代中每个粒子位置 $\hat{s}_m$ 的适应度 fit_m,确定个体极值和群体极值,并分别与个体最优位置 $\vec{p}_m$ 和全局最优位置 $\vec{g}$ 进行比较,从而确定新的个体最优位置和全局最优位置。

步骤5 更新粒子的速度和位置。分别根据(10)式和(11)式更新粒子的速度 $\hat{v}_m$ 和位置 $\hat{s}_m$。

步骤6 随机初始化粒子速度和位置。在粒子每次更新后以概率 $Prob$ 重新初始化其速度 $\hat{v}_m$ 和位置 $\hat{s}_m$。

步骤7 判断是否停止。设置达到最大迭代次数 T 作为算法的停止准则。若满足停止准则,算法终止,输出种群中全局最优位置 $\vec{g}$,即经过粒子群优化算法优化的GMM-RBF神经网络模型;否则返回步骤3重复执行步骤3~步骤6。

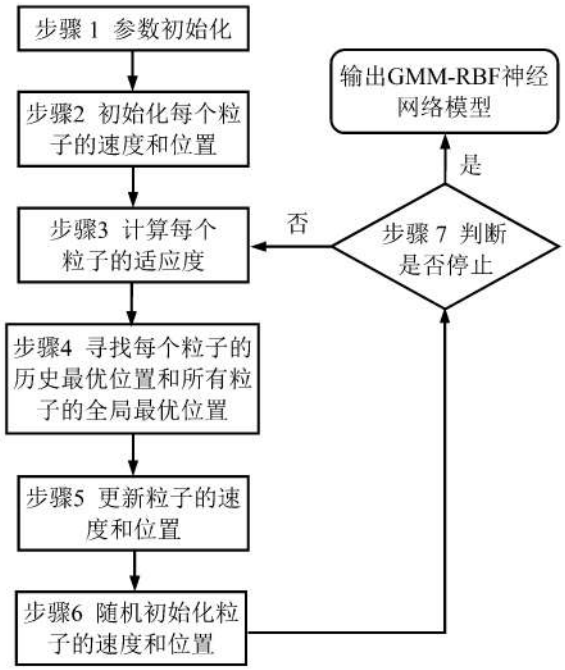


图4 粒子群优化算法训练GMM-RBF神经网络流程

Figure 4 Flow Chart for PSO-based GMM-RBF Neural Network

3 前列腺癌诊断预测实验

3.1 数据准备

为验证本研究提出的GMM-RBF神经网络模型在前列腺癌实际诊断中的有效性,本研究使用国家临床医学科学数据中心(http://101.201.55.39/index?u=25#/)提供的数据进行仿真实验。国家临床医学科学数据中心由中国医学科学院北京协和医院和中国人民解放军总医院共同承担,是国家科技基础条件平台科学数据共享工程的重大项目。该数据中心提供的数据真实、可靠、可用性强,已成为中国医疗数据分析和挖掘研究的主要数据获取来源。本研究的实验数据是由中国人民解放军总医院提供的前列腺肿瘤数据集,该数据集包含2007年至2013年中国人民解放军总医院的前列腺癌患者与诊断有关的记录,其中包括生化检查、血常规、前列腺特异抗原、导尿、放疗信息、核医学、检查、尿常规、膀胱镜、手术情况、药物、诊断等相关信息的表格,信息全部以Excel格式存储。

在分析数据和构建模型之前,需要对数据的质量进行控制。在构建机器学习模型的过程中,数据预处理最为耗时,约占整个数据挖掘工作时间的一半,甚至80%[42],但数据处理结果会对准确性产生很大的影响。

对获得的前列腺癌原始数据集进行表间连接、去重,并排除不完整的数据记录,最终得到1 482条患者样本实例和43个变量,变量的具体信息见附表1。为了更好地理解并选择具有意义的变量,本研究对前列腺癌方面的医学资料进行收集和整理,了解每个变量的具体含义和取值类型等,数据集中涉及到的43个变量大致可以分为6大类,变量分类见表1。

针对前列腺癌诊断问题,根据诊断结果将样本分为两组。诊断结果为前列腺癌的病例,将其标记为1;诊断结果为前列腺炎等其他疾病的病例,将其标记为0。数据集样本类别分布情况见表2。

3.2 前列腺癌诊断变量选择

变量选择是数据预处理工作中非常重要的一部分,从众多的变量中筛选出对分类有意义、具有重要影响的变量是非常有必要的。对于变量的选择,一方面,可以减少模型的计算复杂度,加快模型训练和预测的速度;另一方面,剔除无关变量之后有可能会进一步提升模型的表现[43]。

表 1　变量分类

Table 1　Classification of Variable

类别编号	变量类别
1	病人基本信息
2	物理检查
3	前列腺特异抗原检查
4	尿常规检查
5	血常规检查
6	病人类别

表 2　样本类别分布

Table 2　Distribution of Sample Category

分类标签	诊断结果	数量/人
0	诊断为其他前列腺疾病	768
1	诊断为前列腺癌	714

相关系数法是在进行变量选择时采用的一种简单有效的方法[43],通过计算变量的相关系数,可以知道各变量之间的相关关系。本研究将患者的类别标识作为目标变量,通过计算 Pearson 相关系数得到各项指标与类别之间的相关程度,并且以数据集样本类别将样本分为两组,进行两个独立样本 t 检验,得到各项指标不同类别之间差异的显著性程度,计算结果见表3。

由表3可以清晰地了解各个变量与类别之间的相关程度,与前列腺癌诊断相关程度最强的变量为年龄、前列腺体积、内腺前后径、前列腺特异抗原总浓度和游离前列腺特异抗原浓度。但由于数据分布原因,相关系数法可能会存在偏差,因此本研究在实验中使用目前数据挖掘中广泛使用的随机森林算法进行变量重要性评价,有研究证明使用随机森林法识别关键变量是可行并且有效的[44]。本研究使用R语言的 random forest 程序包进行前列腺癌的特征选择,形成随机森林特征选择图,见图5。图5左侧是按照平均准确度降低量指标进行变量排序,右侧是按照节点基尼不纯度降低量指标进行变量排序,排序方式均是重要性从大到小。由于剩余22项变量两项指标值接近于0,因此在图5中不予列示。由图5可知,两种排序方式均认为内腺前后径、年龄、前列腺体积、游离前列腺特异抗原浓度和前列腺特异抗原总浓度5个变量最为重要,该结果与相关系数方法筛选出的结果一致。

表 3　变量相关系数及两个独立样本t检验结果

Table 3　Results for Variable Correlation Coefficient and Two Independent Sample t-test

变量	相关系数	变量	相关系数	变量	相关系数
年龄	−0.097**	前列腺体积	−0.302***	内腺前后径	−0.377***
前列腺特异抗原总浓度	0.153***	游离前列腺特异抗原浓度	0.143***	浓度比值	−0.020
尿 70% 红细胞前向散射光所在位置	−0.055	尿白细胞	0.063	尿白细胞检查	−0.019
尿比重测定	−0.015	尿胆红素定性试验	0.016	尿胆原定性试验	0.001
尿蛋白定性试验	−0.003	尿红细胞	0.021	尿红细胞检查	0.036
尿红细胞前向散射光分布宽度	−0.016	尿酵母细胞	0.017	尿上皮细胞	−0.050
尿上皮细胞检查镜检	0.094**	尿糖定性试验	0.002	尿酮体试验	−0.006
尿液导电率	0.063	尿液管型	−0.048	尿液结晶	−0.038
尿液酸碱度测定	0.004	尿液亚硝酸盐试验	0.028	尿液颜色	−0.047
白细胞计数	0.003	单核细胞	−0.085*	红细胞比积测定	−0.032
红细胞计数	−0.031	红细胞体积分布宽度测定 CV	0.012	淋巴细胞	0.004
平均红细胞体积	0.005	平均红细胞血红蛋白量	0.026	平均红细胞血红蛋白浓度	0.029
嗜碱性粒细胞	0.052	嗜酸性粒细胞直接计数	−0.030	嗜酸性粒细胞	−0.008
血红蛋白测定	−0.024	血小板计数	−0.039	中性粒细胞	0.011

注:*** 为 $p < 0.010$, ** 为 $p < 0.050$, * 为 $p < 0.100$, 下同。

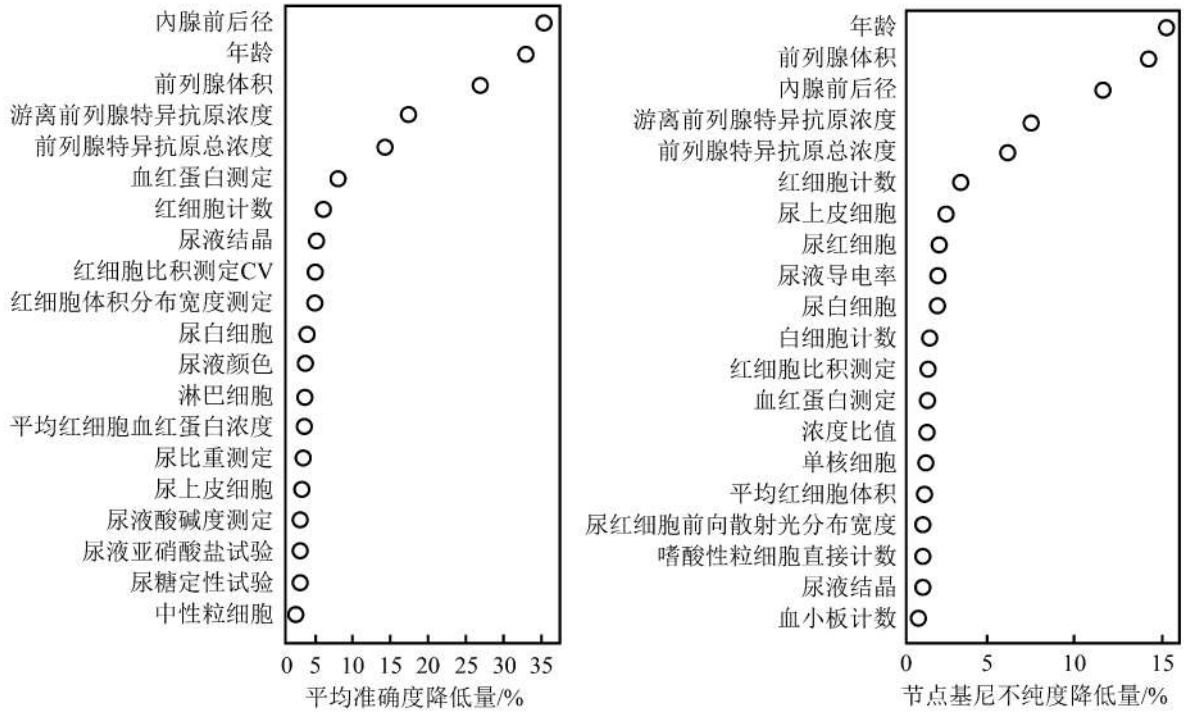


图5 随机森林变量选择

Figure 5 Variable Selection Using Random Forests

为保证使用变量选择方法筛选出的变量在前列腺癌诊断过程中的科学性和合理性,本研究的工作人员到大连市某三甲医院的泌尿外科对前列腺癌的诊断过程进行实地调研,经过与医院主治医师进行确认验证,并结合已有研究[45]中对前列腺癌诊断相关指标的设置,本研究最终确定从42个诊断指标变量中选取出内腺前后径、年龄、前列腺体积、游离前列腺特异性抗原浓度和前列腺特异抗原总浓度5个变量作为前列腺癌诊断建模过程使用的指标。

3.3 实验设置

实验中首先对GMM-RBF神经网络模型与传统的径向基函数神经网络和人工神经网络在前列腺癌诊断上的准确性进行对比分析。并在此基础上,对比采用改进粒子群优化算法训练的GMM-RBF神经网络和采用反向传播算法训练的GMM-RBF神经网络在模型训练过程中训练误差收敛情况,检验采用改进粒子群优化算法在模型训练上相对于反向传播算法的提升效果。最后,将本研究所提方法与当前几种流行的机器学习方法进行比较,验证GMM-RBF神经网络在前列腺癌诊断问题上的有效性。

3.3.1 实验参数设置

根据上文对GMM-RBF神经网络模型训练过程和改进粒子群优化算法的介绍,对涉及的参数通过多次实验进行调优,经过调优后的GMM-RBF神经网络和改进粒子群优化算法的参数设置见表4。为保证前列腺癌预测实验结果的可靠性和科学性,本研究最终确定采用数据科学领域流行的10折交叉验证方法将数据实例划分为独立的训练集和测试集[42]。

表4 GMM-RBF神经网络算法参数

Table 4 Algorithm Parameters for GMM-RBF Neural Network

参数	参数名称	参数取值
K	高斯混合模型算法聚类簇数	4
T_g	高斯混合模型算法最大迭代次数	100
T	粒子群优化算法中最大迭代次数	300
H	径向基函数神经网络隐含层节点数	4
N_p	粒子群优化算法中种群规模	100
$[V_{\min}, V_{\max}]$	粒子群优化算法中粒子速度取值范围	$[-1,1]$
$[S_{\min}, S_{\max}]$	粒子群优化算法中粒子位置取值范围	$[-1,1]$
ω	粒子群优化算法中的惯性系数	0.400
c_1	粒子群优化算法中的加速因子	1.494
c_2	粒子群优化算法中的加速因子	1.494
$Prob$	粒子群优化算法中粒子随机初始化概率	0.150

3.3.2 实验结果评价方法

在分类预测性能评价时,本研究采用医学领域诊断预测中使用的准确率、特异性、敏感性和 *AUC*

(area under roc curve,AUC)值4项评价指标。

在分类器表现评价时,使用混淆矩阵[46]这一工具。混淆矩阵是一种可视化的工具,它将分类结果和实际值放在一个矩阵中,可以清楚地了解到模型预测值与真实情况之间的差异。混淆矩阵的具体表现形式见表5。

表5　分类结果混淆矩阵

Table 5　Classification Results for Confusion Matrix

真实结果	预测结果	
	正例	反例
正例	*TP*	*FN*
反例	*FP*	*TN*

注:*TP* 为混淆矩阵中真正例的数据实例数量,*FP* 为混淆矩阵中假正例的数据实例数量,*TN* 为混淆矩阵中真反例的数据实例数量,*FN* 为混淆矩阵中假反例的数据实例数量;对于前列腺癌诊断问题,正例是被诊断为前列腺癌的实例,反例是未患前列腺癌的实例。

准确率表示分类器总体的分类精度,计算公式为

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{12}$$

其中,*Accuracy* 为准确率。

特异性表示分类器正确识别未患前列腺癌病人的能力,计算公式为

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

其中,*Specificity* 为特异性。

敏感性表示分类器正确识别患前列腺癌病人的能力,计算公式为

$$Sensitivity = \frac{TP}{TP + FN} \tag{14}$$

其中,*Sensitivity* 为敏感性。

受试者工作特征曲线下面积即 *AUC* 值[47],是测量模型在数据集上预测准确性的一种有效指标,与 *Accuracy* 相比,*AUC* 值能更好地反映在数据类别不均衡分布情况下模型的表现,因此被广泛地应用。*AUC* 的计算公式为

$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_{ins_i} - \frac{N_{positive}(N_{positive}+1)}{2}}{N_{positive} \cdot N_{negtive}} \tag{15}$$

其中,*rank* 为根据模型预测实例属于正例概率大小进行排序,$N_{positive}$ 为实际为正例的个数,$N_{negtive}$ 为实际为负例的个数。当 $N_{positive}$ 个正例均排在 $N_{negtive}$ 个负例之前时,*AUC* 取值为1;当 $N_{positive}$ 个正例均排在 $N_{negtive}$ 个负例之后时,*AUC* 取值为0。

3.4　实验结果和分析

除上述对样本和变量的处理之外,为消除不同变量取值范围不同对实验结果的影响,需要对输入数据进行归一化处理,将不同变量的数据统一到相同的取值范围之内[48]。本研究采用z-score标准化方法,这种方法对原始数据的均值 μ 和标准差 σ 进行数据的标准化,经过处理的数据符合标准正态分布,转换公式为

$$X^* = \frac{X_i^{(l)} - \mu_i}{\sigma_i} \tag{16}$$

其中,$X_i^{(l)}$ 为第 l 个样本第 i 维的数据,μ_i 为所有样本第 i 维数据的均值,σ_i 为所有样本第 i 维数据的标准差。

本研究基于参数设置和评价指标进行前列腺癌的诊断实验,实验对比采用改进粒子群优化算法训练的GMM-RBF神经网络方法与其他方法在模型训练过程中训练误差的收敛情况,以验证本研究提出的算法在模型训练过程中相对于其他算法的改进情况。参与比较的方法包括采用改进粒子群优化算法训练的GMM-RBF神经网络(PSO-GMM-RBFNN)、采用反向传播算法训练的GMM-RBF神经网络(BP-GMM-RBFNN)、采用反向传播算法训练的径向基函数神经网络(BP-RBFNN)和采用反向传播算法训练的人工神经网络(BP-ANN)。针对一次典型的训练过程,上述算法的训练误差变化曲线见图6。

由图6可知,在对同样的训练数据集进行300次迭代训练之后,不同的算法取得不同的训练误差。本研究从训练误差和收敛速度两个方面比较分析4种算法。

(1)训练误差。在4种算法中,采用PSO-GMM-RBFNN算法的训练误差最低,且与其他算法相比优势明显,而采用BP-ANN算法的训练误差最高。比较BP-ANN算法与BP-RBFNN算法可以看出,径向基函数神经网络比人工神经网络有着更好的准确性;比较BP-GMM-RBFNN算法与BP-RBFNN算法,融合高斯混合模型算法后神经网络的初始训练误差更低,且训练误差也更小;比较PSO-GMM-RBFNN算法与BP-GMM-RBFNN算法,粒子群优化算法在网络训练方面明显优于反向传播算法。

(2)收敛速度。比较BP-ANN算法与BP-RBFNN算法,径向基函数神经网络比人工神经网络具有很快的收敛速度;比较BP-GMM-RBFNN算法与BP-RBFNN算法可以看出,高斯混合模型算法的使用也加快了网络训练的速度;比较PSO-GMM-RBFNN算法与BP-GMM-RBFNN算法,证明粒子群优化算法比反向传播算法有着更快的收敛速度。

本研究进一步验证采用粒子群优化算法训练的GMM-RBF神经网络方法在前列腺癌诊断问题上的准确性。实验中将本研究方法与神经网络方法及当前流行的支持向量机、逻辑回归和分类回归树等机器学习方法进行比较,使用准确率、特异性、敏感性和 *AUC* 值4个评价指标,实验结果见表6。

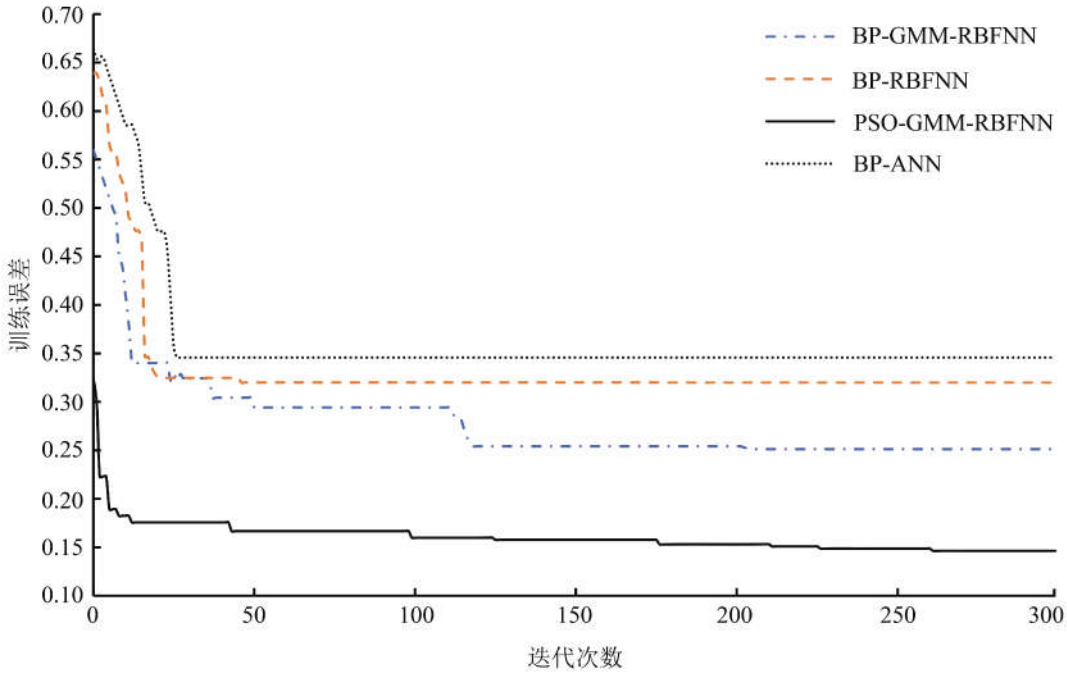


图 6 不同算法训练误差变化曲线

Figure 6 Variation Curve of Training Error between Different Algorithms

表 6 分类实验结果对比

Table 6 Classification Results for Comparison

模型	准确率	特异性	敏感性	*AUC*
支持向量机	0.710	0.761	0.613	0.752
逻辑回归	0.706	0.803	0.540	0.803
分类回归树	0.771	0.831	0.698	0.794
BP-ANN	0.653	0.612	0.610	0.749
BP-RBFNN	0.699	0.630	0.630	0.746
BP-GMM-RBFNN	0.730	0.700	0.674	0.751
PSO-GMM-RBFNN	**0.815**	**0.866**	**0.726**	**0.821**

由表 6 的实验结果可知,在使用10折交叉验证方法的情况下,与其他几种算法相比,本研究提出的使用高斯混合模型对径向基函数神经网络进行预训练并使用改进粒子群优化算法进行权重调整的方法效果最好,得到了0.815的准确率、0.866的特异性、0.726的敏感性和0.821的*AUC*。实验结果表明本研究方法准确率更高,并且可以更好地识别出真正患有前列腺癌的病人,能够为前列腺癌诊断提供更可信的结果,为前列腺穿刺活检确诊过程提供有效的决策支持。

然而,只是依照各项模型评价指标对模型的表现进行评价还缺少类似于统计学检验所具有的科学性和客观性。基于这一考虑,本研究使用配对样本t检验对实验10折结果进行统计学上的检验[49],检验模型之间的差异性情况,检验结果见表7。

由表 7 的配对样本t检验结果可知,本研究提出的PSO-GMM-RBF神经网络方法在10折交叉验证的每一折测试集中,在准确性这一指标上非常显著,p值均小于0.050,而在其他3项指标上则显示出了不同程度的优势。这个结果也验证了本研究提出方法的有效性和优越性。

综合上述实验分析,针对前列腺癌诊断问题,采用改进的粒子群优化训练的GMM-RBF神经网络与传统径向基函数神经网络方法和人工神经网络方法相比,在模型训练上进行了两方面的改进,①在模型初始参数方面,采用高斯混合模型算法对初始径向基函数进行训练,减少模型陷入局部最优的可能,提高了模型的准确性;②在模型的训练过程中,采用改进的粒子群优化算法进行模型参数训练,有效减少计算量,并实现训练误差的快速收敛。经过改进使训练后的径向基函数神经网络模型在前列腺癌诊断上表现更加稳定,能够为前列腺癌诊断提供更加准确、可靠的结果,并为医生是否要对前列腺病人进行穿刺活检提供一定的决策支持。

4 结论

前列腺癌初步诊断的准确性对病患来说至关重要,提高诊断的准确性可以为医疗工作者对病患是否进行穿刺活检操作提供有效的辅助决策和方法支持。针对使用径向基函数神经网络进行前列腺癌诊断时模型的诊断准确性易受初始参数选择影响从

表 7　模型预测性能的配对样本 t 检验结果

Table 7　Results for Paired t-test of Prediction Performance of Each Model

评估指标	模型	逻辑回归	分类回归树	BP-ANN	BP-RBFNN	BP-GMM-RBFNN	PSO-GMM-RBFNN
准确率	支持向量机	0.401	0.006***	0.003***	0.380	0.092*	0.001***
	逻辑回归		0.003***	0.021**	0.400	0.137	0.001***
	分类回归树			0.001***	0.012**	0.046**	0.010**
	BP-ANN				0.094*	0.001***	0.001***
	BP-RBFNN					0.189	0.001***
	BP-GMM-RBFNN						0.004***
敏感性	支持向量机	0.056*	0.159	0.001***	0.010**	0.126	0.073*
	逻辑回归		0.043**	0.001***	0.004***	0.071*	0.022**
	分类回归树			0.001***	0.068*	0.275	0.375
	BP-ANN				0.172	0.077*	0.001***
	BP-RBFNN					0.280	0.124
	BP-GMM-RBFNN						0.268
特异性	支持向量机	0.095*	0.004***	0.001***	0.077*	0.354	0.002***
	逻辑回归		0.174	0.001***	0.049**	0.258	0.027**
	分类回归树			0.001***	0.043**	0.197	0.059*
	BP-ANN				0.007***	0.001***	0.001***
	BP-RBFNN					0.222	0.030**
	BP-GMM-RBFNN						0.130
AUC	支持向量机	0.019**	0.018**	0.465	0.408	0.486	0.034**
	逻辑回归		0.342	0.044**	0.032**	0.006***	0.318
	分类回归树			0.117	0.082*	0.049**	0.245
	BP-ANN				0.468	0.463	0.059*
	BP-RBFNN					0.405	0.001***
	BP-GMM-RBFNN						0.026**

而导致模型准确性偏低的问题,本研究提出改进的GMM-RBF神经网络前列腺癌诊断方法。

经过繁琐的数据清洗工作后,本研究使用相关系数、两独立样本t检验和随机森林变量选择方法对原有的42维变量进行重要性评价,最终挑选出对前列腺癌诊断预测具有临床意义的5个重要变量,即内腺前后径、年龄、前列腺体积、游离前列腺特异抗原浓度和前列腺特异抗原总浓度。在前列腺癌诊断实验中,通过对原有的径向基函数神经网络进行逐步改进对比,本研究提出的使用高斯混合模型对径向基神经网络进行预训练能够取得更好的初始解,使用改进后的粒子群优化算法进一步提升了预测模型的收敛速度和准确性。本研究将提出的方法与现阶段流行的支持向量机、逻辑回归和分类回归树方法进行对比实验,结果表明本研究提出的方法在前列腺癌诊断预测上的优越性。

本研究使用国家临床医学科学数据中心提供的前列腺癌数据进行诊断实验,可以将该方法推广到

实际前列腺癌穿刺活检前的初步诊断决策中,为穿刺活检确诊工作提供指导,避免未患病者接受对身体有不良影响的穿刺活检;同时弥补了传统诊断方式准确率上的不足,使前列腺癌患者能够尽早确诊并得到及时治疗,节约医疗资源,降低医疗成本,提高患者满意度。

考虑到在前列腺癌的诊断过程中仍存在对诊断结果有影响的其他特征,本研究在未来工作中将进一步融合和挖掘其他来源的医疗诊断数据,利用更多的特征进行前列腺癌诊断工作,利用更丰富的数据训练,得到更加准确而稳定的诊断模型,更好地对医疗人员的诊断工作提供支持。针对本研究方法中高斯混合模型聚类数目需要人为确定调整的情况,下一步考虑将这一参数设计成为可自适应动态调整的参数,进一步提升模型的诊断能力。此外,针对乳腺癌和肺癌等其他疾病患病情况的诊断需求而对本研究提出的方法进行改进,也将成为后续的研究工作,以进一步提升本研究提出方法在疾病诊断中的普适性和实用价值。

参考文献:

[1] REGNIER-COUDERT O, MCCALL J, LOTHIAN R, et al. Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. ***Artificial Intelligence in Medicine***, 2012, 55(1): 25-35.

[2] JEMAL A, BRAY F, CENTER M M, et al. Global cancer statistics. ***CA: A Cancer Journal for Clinicians***, 2011, 61(2): 69-90.

[3] CHEN W, ZHENG R, BAADE P D, et al. Cancer statistics in China, 2015. ***CA: A Cancer Journal for Clinicians***, 2016, 66(2): 115-132.

[4] KIM S Y, MOON S K, JUNG D C, et al. Pre-operative prediction of advanced prostatic cancer using clinical decision support systems: accuracy comparison between support vector machine and artificial neural network. ***Korean Journal of Radiology***, 2011, 12(5): 588-594.

[5] CATALONA W J, SMITH D S, RATLIFF T L, et al. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. ***New England Journal of Medicine***, 1991, 324(17): 1156-1161.

[6] 郭熙铜,张晓飞,刘笑笑,等. 数据驱动的电子健康服务管理研究:挑战与展望. **管理科学**, 2017, 30(1): 3-14.
GUO Xitong, ZHANG Xiaofei, LIU Xiaoxiao, et al. eHealth service management research in the big data era: challenges and future directions. ***Journal of Management Science***, 2017, 30(1): 3-14. (in Chinese)

[7] YANG H L, GUO X T, WU T S, et al. Exploring the effects of patient-generated and system-generated information on patients' online search, evaluation and decision. ***Electronic Commerce Research and Applications***, 2015, 14(3): 192-203.

[8] 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.

[9] AZAR A T, EI-METWALLY S M, et al. Decision tree classifiers for automated medical diagnosis. ***Neural Computing & Applications***, 2013, 23: 2387-2403.

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

[11] 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.

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

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

[14] 李庆,谢江凌,杨敏敏. 经直肠超声弹性成像诊断前列腺癌的价值. **临床超声医学杂志**, 2014, 16(7): 497-498.
LI Qing, XIE Jiangling, YANG Minmin. Value of transrectal elastography in diagnosis of prostatic carcinoma. ***Journal of Clinical Ultrasound in Medicine***, 2014, 16(7): 497-498. (in Chinese)

[15] 朱林,黄君,詹洁群,等. 血清 PSA 指标与经直肠超声造影对前列腺癌的诊断. **暨南大学学报(自然科学与医学版)**, 2015, 36(6): 515-519.
ZHU Lin, HUANG Jun, ZHAN Jiequn, et al. The diagnostic values of serum prostate specific antigen and transrectal contrast-enhanced ultrasonography of prostate cancer. ***Journal of Jinan University*** (***Natural Science & Medicine Edition***), 2015, 36(6): 515-519. (in Chinese)

[16] LEE S E, CHUNG J S, HAN B K, et al. Relationship of prostate-specific antigen and prostate volume in Korean men with biopsy-proven benign prostatic hyperplasia. ***Urology***, 2008, 71(3): 395-398.

[17] LOEB S, CATALONA W J. Prostate-specific antigen in clinical practice. ***Cancer Letters***, 2007, 249(1): 30-39.

[18] 王金萍,徐浩. 血清 TPSA、FPSA/TPSA 及 PSAD 对前列腺癌的诊断价值. **暨南大学学报(自然科学与医学版)**, 2007, 28(2): 172-175.
WANG Jinping, XU Hao. Clinical significance of serum TPSA, FPSA/TPSA and PSAD in diagnosis of prostate carcinoma. ***Journal of Jinan University*** (***Natural Science & Medicine Edition***), 2007, 28(2): 172-175. (in Chinese)

[19] LEE H J, HWANG S I, HAN S M, et al. Image-based clinical decision support for transrectal ultrasound in the diagnosis of prostate cancer: comparison of multiple logistic regression, artificial neural network, and support vector machine. ***European Radiology***, 2010, 20(6): 1476-1484.

[20] FINNE P, FINNE R, BANGMA C, et al. Algorithms based on prostate-specific antigen (PSA), free PSA, digital rectal examination and prostate volume reduce false-postitive PSA results in prostate cancer screening. ***International Journal of Cancer***, 2004, 111(2): 310-315.

[21] BERMEJO P, VIVO A, TÁRRAGA P J, et al. Development

of interpretable predictive models for BPH and prostate cancer. ***Clinical Medicine Insights : Oncology***, 2015, 9: 15-24.

[22] HU X, CAMMANN H, MEYER H A, et al. Artificial neural networks and prostate cancer tools for diagnosis and management. ***Nature Reviews Urology***, 2013, 10(3): 174-182.

[23] SNOW P B, SMITH D S, CATALONA W J. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. ***The Journal of Urology***, 1994, 152(5): 1923-1926.

[24] BABAIAN R J, FRITSCHE H, AYALA A, et al. Performance of a neural network in detecting prostate cancer in the prostate-specific antigen reflex range of 2.5 to 4.0 ng/mL. ***Urology***, 2000, 56(6): 1000-1006.

[25] STEPHAN C, JUNG K, CAMMANN H, et al. An artificial neural network considerably improves the diagnostic power of percent free prostate-specific antigen in prostate cancer diagnosis: results of a 5-year investigation. ***International Journal of Cancer***, 2002, 99(3): 466-473.

[26] MOODY J, DARKEN C J. Fast learning in networks of locally-tuned processing units. ***Neural Computation***, 1989, 1(2): 281-294.

[27] 黄星,孙明. 基于 RBF 神经网络的震伤人员快速评估模型. **系统工程**, 2016, 34(8): 129-135.
HUANG Xing, SUN Ming. The rapid assessment of wounded personnel based on RBF neural network model under the background earthquake disaster. ***Systems Engineering***, 2016, 34(8): 129-135. (in Chinese)

[28] MARÍN O, RUIZ D, PÉREZ I, et al. Use of radial basis functions in computer-aided diagnosis of prostate cancer // **2011** ***Annual International Conference of the IEEE Engineering-in-Medicine-and-Biology-Society*** (***EMBS***). Boston, MA, 2011: 6422-6425.

[29] WALLACE M, TSAPATSOULIS N, KOLLIAS S. Intelligent initialization of resource allocating RBF networks. ***Neural Networks***, 2005, 18(2): 117-122.

[30] 杨淑娥,黄礼. 基于 BP 神经网络的上市公司财务预警模型. **系统工程理论与实践**, 2005, 25(1): 12-18, 26.
YANG Shue, HUANG Li. Financial crisis warning model based on BP neural network. ***Systems Engineering - Theory & Practice***, 2005, 25(1): 12-18, 26. (in Chinese)

[31] 孙佰清,冯英浚,潘启树,等. 急性心肌梗塞诊断的智能决策支持系统. **系统工程理论与实践**, 2006, 26(10): 141-144.
SUN Baiqing, FENG Yingjun, PAN Qishu, et al. Intelligent decision support system for the diagnosis of acute myocardial infarction. ***Systems Engineering - Theory & Practice***, 2006, 26(10): 141-144. (in Chinese)

[32] 肖斌卿,杨旸,李心丹,等. 基于 GA-ANN 的中国金融安全预警系统设计及实证分析. **系统工程理论与实践**, 2015, 35(8): 1928-1937.
XIAO Binqing, YANG Yang, LI Xindan, et al. Design of China's financial security early warning system based on GA-ANN. ***Systems Engineering - Theory & Practice***, 2015, 35(8): 1928-1937. (in Chinese)

[33] 郭海湘,诸克军,李四福,等. 煤矿首采面开工进度计划的智能优化. **系统工程理论与实践**, 2009, 29(11): 135-144.
GUO Haixiang, ZHU Kejun, LI Sifu, et al. Intelligent optimization for the first coal face project scheduling in coal mining. ***Systems Engineering - Theory & Practice***, 2009, 29(11): 135-144. (in Chinese)

[34] 王亮,张宏伟,岳琳,等. PSO-BP 模型在城市用水量短期预测中的应用. **系统工程理论与实践**, 2007, 27(9): 165-170.
WANG Liang, ZHANG Hongwei, YUE Lin, et al. Application of PSO-BP model in short-term prediction of urban water consumption. ***Systems Engineering - Theory & Practice***, 2007, 27(9): 165-170. (in Chinese)

[35] 卫敏,余乐安. 具有最优学习率的 RBF 神经网络及其应用. **管理科学学报**, 2012, 15(4): 50-57.
WEI Min, YU Lean. A RBF neural network with optimum learning rate and its application. ***Journal of Management Sciences in China***, 2012, 15(4): 50-57. (in Chinese)

[36] XIANG Z Y, XIAO Z, WANG D, et al. A Gaussian mixture framework for incremental nonparametric regression with topology learning neural networks. ***Neurocomputing***, 2016, 194: 34-44.

[37] REYNOLDS D A, ROSE R C. Robust text-independent speaker identification using Gaussian mixture speaker models. ***IEEE Transactions on Speech and Audio Processing***, 1995, 3(1): 72-83.

[38] KUO R J, HUANG M H, CHENG W C, et al. Application of a two-stage fuzzy neural network to a prostate cancer prognosis system. ***Artificial Intelligence in Medicine***, 2015, 63(2): 119-133.

[39] 李壮阔,薛有添. 基于粒子群算法的模糊层次分析法改进及其应用研究. **运筹与管理**, 2013, 22(4): 139-143.
LI Zhuangkuo, XUE Youtian. Improvement of fuzzy analytic hierarchy process based on particle swarm optimization and its application research. ***Operations Research and Management Science***, 2013, 22(4): 139-143. (in Chinese)

[40] 李松,刘力军,翟曼. 改进粒子群算法优化 BP 神经网络的短时交通流预测. **系统工程理论与实践**, 2012, 32(9): 2045-2049.
LI Song, LIU Lijun, ZHAI Man. Prediction for short-term traffic flow based on modified PSO optimized BP neural network. ***Systems Engineering - Theory & Practice***, 2012, 32(9): 2045-2049. (in Chinese)

[41] ALIČKOVIĆ E, SUBASI A. Breast cancer diagnosis using GA feature selection and rotation forest. ***Neural Computing & Applications***, 2017, 28(4): 753-763.

[42] 王宇燕,王杜娟,王延章,等. 改进随机森林的集成分类方法预测结直肠癌存活性. **管理科学**, 2017, 30(1): 95-106.
WANG Yuyan, WANG Dujuan, WANG Yanzhang, et al. Predicting survivability of colorectal cancer by an ensemble classification method improved on random forest. ***Journal of Management Science***, 2017, 30(1): 95-106. (in Chinese)

[43] TURGEMAN L, MAY J H. A mixed-ensemble model for hospital readmission. ***Artificial Intelligence in Medicine***, 2016, 72: 72-82.

[44] HAPFELMEIER A, ULM K. A new variable selection approach using random forests. ***Computational Statistics & Data Analysis***, 2013, 60: 50-69.

[45] 徐勇, 张志宏. **前列腺癌**. 北京: 科学技术文献出版社, 2009: 341-416.
XU Yong, ZHANG Zhihong. ***Prostate cancer***. Beijing: Scientific and Technical Documentation Press, 2009: 341-416. (in Chinese)

[46] BACH M, WERNER A, ŻYWIEC J, et al. The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. ***Information Sciences***, 2017, 384: 174-190.

[47] GORUNESCU F, BELCIUG S. Boosting backpropagation algorithm by stimulus-sampling: application in computer-aided medical diagnosis. ***Journal of Biomedical Informatics***, 2016, 63: 74-81.

[48] LI Z, XU W, ZHANG L K, et al. An ontology-based Web mining method for unemployment rate prediction. ***Decision Support Systems***, 2014, 66: 114-122.

[49] 林宇, 黄迅, 淳伟德, 等. 基于ODR-ADASYN-SVM的极端金融风险预警研究. **管理科学学报**, 2016, 19(5): 87-101.
LIN Yu, HUANG Xun, CHUN Weide, et al. Early warning for extremely financial risks based on ODR-ADASYN-SVM. ***Journal of Management Sciences in China***, 2016, 19(5): 87-101. (in Chinese)

附表 1 变量详细信息

Appendix 1 Variable Details

变量类别	变量标记	变量释义
病人	label	类别标记
基础信息	age	年龄
物理检查	volume	前列腺体积
	fbdia	内腺前后径
血常规检查	x_white_cell_count	白细胞计数
	x_mononuclear	单核细胞
	x_erythrocyte_ratio	红细胞比积测定
	_red_cell_count	红细胞计数
	x_red_cell_volume	红细胞体积分布宽度测定 CV
	x_lymphocytes	淋巴细胞
	x_average_erythrocyre_volume	平均红细胞体积
	x_average_erythrocyte_hemoglobin	平均红细胞血红蛋白量
	x_average_hemoglobin_concentration	平均红细胞血红蛋白浓度
	x_basophils	嗜碱性粒细胞
	x_eosinophils_count	嗜酸性粒细胞直接计数
	x_eosinophils	嗜酸性粒细胞
	x_hemoglobin	血红蛋白测定
	x_platelet_count	血小板计数
	x_neutrophils	中性粒细胞
前列腺特异抗原检查	tpsa	前列腺特异抗原总浓度
	fpsa	游离前列腺特异抗原浓度
	fpsa/tpsa	浓度比值
尿常规检查	n_70% location	尿 70% 红细胞前向散射光所在位置
	n_leukocyte	尿白细胞
	n_leukocyte_check	尿白细胞检查
	n_proportion	尿比重测定
	n_bilirubin	尿胆红素定性试验
	n_bile	尿胆原定性试验
	n_protein	尿蛋白定性试验
	n_red_blood_cells	尿红细胞
	n_red_blood_cells_check	尿红细胞检查
	n_distribution_width	尿红细胞前向散射光分布宽度
	n_yeast_cells	尿酵母细胞
	n_epithelial_cells	尿上皮细胞
	n_epithelial_cells_check	尿上皮细胞检查镜检
	n_sugar	尿糖定性试验
	n_ketone	尿酮体试验
	n_conductivity	尿液导电率
	n_tube_type	尿液管型
	n_crystallization	尿液结晶
	n_PH	尿液酸碱度测定
	n_nitrite	尿液亚硝酸盐试验
	n_colour	尿液颜色

Prostate Cancer Diagnosis Method Based on GMM-RBF Neural Network

CUI Shaoze[1], WANG Dujuan[1], WANG Sutong[1], XIA Jiangnan[1], WANG Yanzhang[1], JIN Yaochu[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, UK

Abstract: Prostate cancer is the fastest rising incidence of male cancer in recent years, which is a serious health threat to the patients. How to diagnose the condition of cancer patients more accurately is very important for the timely treatment and reduction of the mortality of pros-tate cancer. In recent years, cancer diagnosis based on data mining has gradually become a research focus in the field of disease diagnosis, and it has shown great advantages in improving the accuracy of diagnosis.

In order to solve the problem that the low accuracy of the existing methods for early diagnosis of prostate cancer, this paper presents a new diagnosis method called GMM-RBF neural network based on improved RBF neural network with GMM. In this method, the parameters of radial basis function in radial basis function neural network are pre-trained by using Gaussian mixture model to avoid the model getting into local optimum. Then, the improved PSO algorithm is used to train the neural network. In the experiment, the data provided by the National Clinical Medical Science Data Center is used to compare the proposed method with the other popular machine learning methods such as RBF neural network, classification and regression tree, support vector machine and logistic regression. The performance of the model is evaluated using accuracy, specificity, sensitivity, and *AUC*.

The experimental results show that the GMM-RBF neural network model has faster convergence rate and higher initial accuracy than the pre-improved neural network model. Compared with other machine learning algorithms, the GMM-RBF neural network model achieves a higher accuracy, sensitivity, specificity and *AUC* during ten-fold cross-validation.

In this paper, the proposed GMM-RBF neural network method has a great improvement on the model prediction accuracy compared with the traditional RBF neural network model, which can provide more reliable results for the diagnosis of prostate cancer. It provides effective auxiliary decision-making support for the preliminary diagnosis of prostate cancer for medical workers and has practical significance to reduce the pain of patients, improve patient satisfaction and save medical resources.

Keywords: prostate cancer; RBF neural network; Gaussian mixture model; particle swarm optimization; disease diagnosis

Received Date: September 19th, 2017　　**Accepted Date**: December 27th, 2017

Funded Project: Supported by the National Natural Science Foundation of China (71533001, 71672019, 71271039)

Biography: CUI Shaoze is a master degree candidate in the Faculty of Management and Economics at Dalian University of Technology. His research interests cover medical health management, machine learning and intelligence optimization algorithms. E-mail: csz2016@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 cover service operation management, data mining and intelligent optimization algorithm. Her representative paper titled "Disruption management for new jobs arrivals with deteriorating effect and controllable pro-cessing times" was published in the ***Journal of Systems & Management*** (Issue 5, 2016). E-mail: wangdujuan@dlut.edu.cn

WANG Sutong is a master degree candidate in the Faculty of Management and Economics at Dalian University of Technology. His research interests include machine learning and rule mining. E-mail: sutongwang@mail.dlut.edu.cn

XIA Jiangnan is a master degree candidate in the Faculty of Management and Economics at Dalian University of Technology. His research interests include image identification and text mining. E-mail: jn_xia@foxmail.com

WANG Yanzhang, doctor in engineering, is a professor in the Faculty of Management and Economics at Dalian University of Technology. His research interests include e-government and knowledge management. His representative paper titled "Knowledge and representation of model management" was published in the ***Journal of Systems Engineering*** (Issue 6, 2011). E-mail: yzwang@dlut.edu.cn

JIN Yaochu, doctor in engineering, is a professor and a chair of computational intelligence in the Department of Computing at University of Surrey, UK. His research interests cover computational intelligence, machine learning, computational biology and computational neuroscience. 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 □