CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 38 卷第 11 期 2016 年 11 月
Vol. 38, No. 11, Nov. 2016

文章编号:1007-130X(2016)11-2303-07

机器学习在乳腺肿瘤分类检测中的应用研究*

李　喆,吕　卫,闵　行,褚晶辉
(天津大学电子信息工程学院,天津 300072)

摘　要:机器学习算法在医学检测与诊断,尤其是乳腺肿瘤分类检测与诊断中扮演愈发重要的角色。分析比较了几种经典机器学习分类器在乳腺肿瘤分类检测中的性能,并从准确率、灵敏度、特异性及执行效率等方面对各分类器的性能进行了评估比较,根据在不同数据库上的实验结果,总结了各机器学习分类器在乳腺肿瘤分类中的性能特点:线性判别分析和极限学习机两种分类器性能优良且训练效率很高;支持向量机性能较为平均且非常稳定,但训练耗时较长;而人工神经网络分类器虽然可以给出良好的特异性指标,但灵敏度指标不够理想。

关键词:乳腺肿瘤;机器学习;性能比较
中图分类号:TP391.4　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2016.11.022

Application of machine learning algorithms in breast tumor detection

LI Zhe, LÜ Wei, MIN Hang, CHU Jing-hui
(School of Electronic and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Machine learning algorithms are playing an increasingly important role in medical detection and diagnosis, especially for breast tumor classification, detection and diagnosis. We evaluate these machine learning methods based on criterions including accuracy, sensitivity, specificity and efficiency. We then summarize the characteristics of different classifiers according to the experimental results of different breast tumor databases: all of the classifiers can achieve relatively ideal performance in terms of testing efficiency. The linear discriminant analysis and the extreme learning machine have excellent classification performance and high training efficiency while the support vector machine has average classification performance and a long training time, and the artificial neural network has relatively low sensitivity but an extremely high specificity.

Key words: breast tumor; machine learning; performance comparison

1　引言

机器学习是研究如何使计算机根据经验来学习的算法[1]。目前,机器学习领域已经出现诸如线性判别分析 LDA(Linear Discriminant Analysis)、支持向量机 SVM(Support Vector Machine)、人工神经网络 ANN(Artificial Neural Network)和极限学习机 ELM(Extreme Learning Machine)等多种通用算法,并在医疗检测中得到了广泛应用。各种常见机器学习算法已经在乳腺肿瘤检测[2]、肺部肿瘤检测[3]以及肝肿瘤检测[4]等诸多医学领域展现出优良性能和巨大的潜力。随着大数据时代的来临,机器学习算法将在医学检测中扮演愈来愈重要的角色。

乳腺癌是女性中最为常见的恶性肿瘤,居全世界妇女恶性肿瘤死亡率的首位[5]。目前,及早诊断与及时治疗是应对乳腺癌最有效的措施。医学影

* **收稿日期**:2015-07-01;**修回日期**:2015-11-05
通信地址:300072 天津市南开区卫津路 92 号天津大学 26 教学楼 D435
Address: Room D435, Building 26, Tianjin University, 92 Weijin Rd, Nankai District, Tianjin 300072, P. R. China

像学方法,如X线、核磁共振、超声检测等,是目前最主要的检测和诊断乳腺癌的手段[6]。然而,在乳腺检查中产生的大量影像信息易使医生疲劳,且诊断精度受医师的职业能力、经验等主观因素影响。在此背景下,通过机器学习方法来判定肿瘤是否存在及其良恶性成为一个广泛关注的研究热点。

本文将几种经典的机器学习算法应用于乳腺肿瘤分类中,并对其分类效果进行比较与分析。首先在乳腺肿瘤图像中提取特征构建正负样本集,考虑到医学图像数据库中普遍存在的正样本数目远少于负样本数目的实际情况,本文在应用分类器进行分类之前对数据集中的负样本进行随机降采样(Random Undersampling)处理使降采样得到的子负样本集与全体正样本构成平衡的独立训练集。随后运用线性判别分析、支持向量机、人工神经网络和极限学习机四种常用分类器,分别从检测肿瘤是否存在和判别肿瘤良恶性两个方向对图像进行分类识别,并从准确率、灵敏度、特异性以及执行效率等方面对各算法的性能进行评估比较,并对比较结果进行说明。

2 算法介绍

如上文所述,本文所用数据集中,正样本数目远远少于负样本数目。这种类数据不平衡现象会导致正样本(少数样本)的检测灵敏度较低[7]。为保证分类的准确度和有效性,本文采用简单的随机降采样算法使数据集变为平衡数据集后再进行训练。

根据分类原理,机器学习分类器可大体分为两类。一类的基本思想是寻找最优超平面,使得分类后各类保证低类间耦合度和高类内耦合度,这种思想的代表算法有线性判别分析和支持向量机;而另一类受生物学习过程启发,通过神经传播与迭代的方式尽可能逼近目标函数来完成训练过程并用于分类,这种思想的代表算法则有人工神经网络和极限学习机[8]。下面将对这些分类器进行基本介绍。

2.1 寻找最佳超平面的分类器

2.1.1 线性判别分析

线性判别分析LDA是在1996年由Belhumeur引入模式识别和人工智能领域的,其基本思想是将高维的模式样本投影到最佳鉴别矢量空间,以达到抽取分类信息和压缩特征空间维数的效果,投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离,即模式在该空间中有最佳的可分离性[9]。其推导过程如下:

假设给定一组 n 个 d 维的样本 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,其中有 n_1 个样本属于类别 χ_1,而剩余的 n_2 个样本属于类别 χ_2。取各自的类内均值为LDA的目标即为确定最佳的直线方向 $\boldsymbol{w}$,使分类效果最好。两类中,各自可得到类内均值 $\boldsymbol{m}_i=\frac{1}{n_i}\sum_{x\in\chi_i}\boldsymbol{x}$。LDA算法的目标即为找到使式(1)结果最大(max($J(\boldsymbol{w})$))的 $\boldsymbol{w}$:

$$J(\boldsymbol{w})=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w}} \tag{1}$$

其中,$\boldsymbol{S}_{\mathrm{B}}=(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}}$ 为类间散布矩阵(Between-Class Scatter Matrix),用于表征两类样本之间的离散程度;$\boldsymbol{S}_{\mathrm{W}}=\sum_{i=1}^{2}\sum_{x\in\chi_i}(\boldsymbol{x}-\boldsymbol{m}_i)(\boldsymbol{x}-\boldsymbol{m}_i)^{\mathrm{T}}$ 为总类内散布矩阵(Within-Class Scatter Matrix),表示两类样本内各样本点之间的离散程度的总和。当 $J(\boldsymbol{w})$ 取得最大时,即可搜索到一个方向 $\boldsymbol{w}$ 并由此得到一个判决边界,使各样本点的投影在此方向上满足类间离散度最大的同时两类各自的类内方差之和最小,即取得最高的类内耦合度和最低的类间耦合度。

对于一个新输入的样本 $\boldsymbol{x}$,在对其进行分类时,首先计算判别函数 $\boldsymbol{y}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$,将其与判决边界进行比较,从而完成分类。

由于该种算法的目标是将高维模式样本投影以将所有样本分成两类,因此具有非常良好的降维效果,可大大提升分类效率。LDA更多依赖数据分布的均值信息,在两类均值具备投影可分的性质时可表现出良好的分类性能[10]。

2.1.2 支持向量机

支持向量机SVM是Cortes等人于1995年首先提出的[11]。它建立在统计学习理论的VC维理论和结构风险最小原理基础之上,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以求获得最好的推广能力[12]。支持向量机将向量映射到一个更高维的空间里使其线性可分,并在该空间中找到一个可将样本分为两类的超平面,随后通过求解一个二次优化问题以尽可能最大化分类间隔。其推导过程如下:

假定有线性可分的数据集 $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$,其中 $\boldsymbol{x}$ 为实数空间中的 d 维向量;y

为类标签,此处仅可能取－1 和＋1 两个值,即只有两类样本。本文中称类标签为＋1 的样本为正例,类标签为－1 的样本为反例。若要对这两类样本进行分类,则目标即应为根据训练样本确定最大分类间隔的分割超平面,设最优超平面方程为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=0$,此时样本与最佳超平面之间的距离应为 $\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}|}{\|\boldsymbol{w}\|}$。对超平面进行规范化,选择使得距超平面最近的样本 $\boldsymbol{x}_k$ 满足 $|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}|=1$ 的 $\boldsymbol{w}$ 和 $\boldsymbol{b}$,即得到规范化超平面。此时从最近样本到边缘的距离为:

$$\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}|}{\|\boldsymbol{w}\|}=\frac{1}{\|\boldsymbol{w}\|}$$

且分类间隔为:

$$m=\frac{2}{\|\boldsymbol{w}\|} \tag{2}$$

由上可知,最终的目标应该是寻找使得式(2)中 m 值最大化的法向量 $\boldsymbol{w}$,之后将 $\boldsymbol{w}$ 代入关系式 $|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_k+\boldsymbol{b}|=1$,即可得到 $\boldsymbol{b}$ 值,从而得到判决超平面。

使 m 最大化就等价于使向量 $\boldsymbol{w}$ 的模最小,即使下式中的结果最小:

$$J(\boldsymbol{w})=\frac{1}{2}\|\boldsymbol{w}\|^2$$

且由于距离超平面最近的点 $\boldsymbol{x}_k$ 满足 $|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}|=1$,因此其他样本点距离超平面的距离应大于该点,因此还应满足约束条件:

$$|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}|\geqslant 1 \tag{3}$$

在式(3)约束下利用拉格朗日乘子法可将式(2)中的条件极值问题转化为一个对偶问题,随后利用二次规划可以解出最优的拉格朗日乘数 $\boldsymbol{\alpha}^*$,并由此解出最优的 $\boldsymbol{w}$ 和 $\boldsymbol{b}$,分别记为 $\boldsymbol{w}^*$ 和 $\boldsymbol{b}^*$,最终得到分类函数如下所示:

$$h(\boldsymbol{x})=\mathrm{sgn}((\boldsymbol{w}^*\boldsymbol{x})+\boldsymbol{b}^*)=$$

$$\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i^* y_i(\boldsymbol{x}_i.\boldsymbol{x})+\boldsymbol{b}^*)$$

其中,向量 $\boldsymbol{x}$ 为待分类的测试样本,向量 $\boldsymbol{x}_i(i=1,2,\cdots,N)$ 为训练集中的全体样本。

而对于非线性问题,可以通过式(3)的约束条件中引入松弛变量放宽约束条件进行求解,也可以运用核函数,通过非线性变换将其转化为某个高维空间中的线性问题,在变换空间求得最佳分类超平面。

SVM 算法理论上可得到全局最优解,避免陷入局部最优[13],可在样本集较小或线性不可分的情形下得到较好的分类结果,并能够推广应用到函数拟合、回归分析等其他机器学习问题中。

2.2　模拟生物学习的分类算法

2.2.1　人工神经网络

人工神经网络 ANN 是通过对人脑或生物神经网络若干基本特性的抽象和模拟构建的数学模型,通过模拟大脑神经系统的信息传播机制,对输入信息进行处理。其模型是由输入层、隐含层和输出层以节点的形式按照一定规律相互连接而成的,通过节点间相互作用的动态过程来调节每一个神经元对最终结果产生影响的幅度权值,最终完成分类[14]。

根据神经元之间相互连接的拓扑结构的不同,人工神经网络主要可以分为前馈网络(Feedforward Neural Network)和反馈网络(Recurrent Neural Network)两种[15]。在前向网络中,神经元的输入来自于上一级神经元的输出,信号只向前传输,将相同的输入经简单非线性函数的多次复合以获取精度的提高。这种网络结构较为简单,其典型代表有反向传播网络 BP(Back Propagation)和径向基函数网络(Radial Basis Function)。而在反馈神经网络中,网络内部的神经元之间存在反馈,信号可能会反向传导。Hopfield 网络和玻尔兹曼机(Boltzmann Machine)均属于该种网络。

人工神经网络可以通过学习训练集标签提取出精度较高的求解规则,适合求解诸如模式识别等难以得到准确求解规则的问题[16]。

2.2.2　极限学习机

极限学习机 ELM 最早由 Huang Guang-bin 等人[17]于 2006 年提出,是一种用于求解单隐层前馈神经网络的学习算法。它是一种新型的神经网络模型,与传统的人工神经网络模型不同,该算法中内部神经元的各项参数随机选取,无需进行任何迭代运算,输出权值是利用代数方法将损失函数最小化后得到的最小二乘解[18]。

设单隐层神经网络的隐层神经元数目为 L,数据集 $\{(\boldsymbol{x}_1,t_1),(\boldsymbol{x}_2,t_2),\cdots,(\boldsymbol{x}_n,t_n)\}$,其中 $\boldsymbol{x}_i$ 为实数空间中多维向量,t_i 为类标签,G 为神经网络中的激励函数。存在随机选取的 a_i、b_i(隐层节点参数)和经训练得到的连接第 j 个隐层和网络输出之间的外权 β_i,使第 j 个输入样本的输出函数满足下式:

$$f_L(\boldsymbol{x}_j)=\sum_{i=1}^{L}\beta_i G(a_i,b_i,\boldsymbol{x}_j)=t_j,j=1,2,\cdots,N$$

可简记为:

$$\boldsymbol{H\beta} = \boldsymbol{T} \tag{4}$$

其中,$\boldsymbol{H}$ 为神经网络隐层的输出矩阵,$\boldsymbol{\beta}$ 为各节点的外权矩阵,$\boldsymbol{T}$ 为类标签矩阵。由式(4)易知,最优的外权向量可直接用 Moore-Penrose 广义逆求解,记为 $\boldsymbol{\beta}^*$。

与传统的 ANN 相比,ELM 可以有效避免陷入局部最优,同时与 SVM 相比基本省去了参数调节所需的全部时间。实例和研究结果表明,该分类器的准确率与 SVM 相当,在某些应用领域甚至更高,且具有计算时间短的绝对优势[19]。

3 实验结果与分析

3.1 数据集概述

为保证实验结果的可靠性与准确性,本文采用了两组不同的乳腺肿瘤 X 线图像数据集,并分别提取不同特征在 Matlab 2014a 上搭建平台进行实验。数据集 1 提取自数字乳腺 X 线图像数据库(Digital Database for Screening Mammography)[20],感兴趣区域共 1 950 个。该数据集用于考察分类器对于肿瘤是否存在的分类准确度,数据集中正负样本(即有肿瘤与无肿瘤样本)数目不平衡,共有正样本 401 个,负样本 1 549 个,比例大约为1∶4。对每个疑似区域,本文依据文献[21,22]提取形态及纹理特征用于分类,分别是圆度、径向长度的平均值和标准差、灰度熵、灰度均值、灰度标准差、肿块面积、平均分形维数、分形维数标准差、光度惯性动力、各向异性、轮廓梯度熵、平滑度、偏度和峰度,共 34 维,在进行归一化处理后完成训练与分类。

数据集 2 是由威斯康星大学医学院经多年搜集与整理建立的乳腺肿瘤病灶组织细胞核显微图像数据库[23],该数据库用于考察各分类器对于肿瘤良恶性的判定能力,包含恶性肿瘤样本 212 个,良性肿瘤样本 357 个,共 569 个样本,正负样本的数量比约为 1∶2。该数据库提取了细胞核半径、质地、周长、面积、光滑性、紧密度、凹陷度、凹陷点数、对称度和断裂度共 10 个与肿瘤性质密切相关的量化特征共 30 维,在进行归一化处理后完成训练与分类。

3.2 评价标准

传统的分类学习中,最常用的分类评价指标是分类精度(Accuracy),但在二项分类中,灵敏度(Sensitivity)与特异性(Specificity)两个指标也经常用于表征分类性能。两个指标的定义如下:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TP}{TP + FP} \tag{6}$$

除单独考察上述各项指标外,将其中几项指标综合考量也是一种非常重要的评价方法。最常见的综合评价指标有 F 值和 G 值。

F 值(F-Measure)是总体分类精度和灵敏度的加权调和平均,其定义如下:

$$F = \frac{(1+\beta) * Accuracy * Sensitivity}{Accuracy + Sensitivity} \tag{7}$$

其中,β 是一个可调节参数,表示总体分类精度和灵敏度的重要性之比。当将总体分类精度和灵敏度视为同等重要,即:

$$F_1 = \frac{2 * Accuracy * Sensitivity}{Accuracy + Sensitivity} \tag{8}$$

F_1 是最为常用的 F 值,本文所采用的 F 值即为 F_1 值给出的结果。

G 值(G-measure)表示总体分类精度和灵敏度的几何平均,也是非常重要的分类器评价标准,其表达式为:

$$G = \sqrt{Accuracy * Sensitivity} \tag{9}$$

为尽可能准确地得到上述指标,本文对每种算法运用十折交叉验证取平均的方法对数据集进行测试。

在实际应用中,执行效率也是衡量算法性能的非常重要的指标,执行效率越高的算法在实际应用中越有可能得到广泛应用,本文也将对各算法的执行效率进行比较与讨论。考虑到分类器的训练过程通常都在线下完成,与实际应用关系不大,因此应将训练时间与测试时间分开考虑。训练时间定义为每种算法寻找参数与建立模型所需要的时间,而测试时间则定义为分类器进行十折交叉验证测试所需要的时间。耗时越长的算法执行效率越低,反之则执行效率越高。

3.3 实验环境与参数设置

3.3.1 针对数据不平衡现象的参数设置

由于一般情况下肿瘤数据库中的正样本数目要远少于负样本的数目,若数据非线性可分,则这种不平衡现象在分类过程中会导致正样本的判决边界被负样本所侵占,最终导致分类性能受到严重影响。本文对数据集中的多数样本即负样本进行随机降采样以避免由于数据不平衡所导致的分类性能下降的问题。

在数据集1中,正负样本比例接近1∶4,对该数据集的负样本进行随机降采样,使训练集中包含用于训练的全部正样本和经随机采样得到的与正样本数目相近的负样本。该平衡训练集中包含正负样本各361个。

而在数据集2中,正负样本比例接近1∶2,因此对数据集中的负样本进行随机降采样得到平衡的训练集,其中包含正负样本各190个。

3.3.2　针对各分类器的参数设置

本文中涉及到的算法均在Matlab 2014a下实现,训练与分类过程也都在上述环境中完成。

LDA算法无需专门设置参数,只需求得类间离散度和类内离散度并求出二者最大比值即可确定分类超平面。

而在随机降采样SVM中,为解决样本线性不可分问题,所有并联的SVM分类器均使用径向基函数作为核函数,损失参数C及核参数$gamma$通过十折交叉验证网格搜索在$[2^{-5},2^{5}]$寻找最佳,训练及分类均采用libsvm软件包[24]。

ANN分类器中,本文选择BP和学习矢量量化LVQ(Learning Vector Quantification)两种常用神经网络,两种分类器均需要设置网络的层数、隐层节点数、学习步长以及迭代次数等多项参数。对于BP网络,本文选择单隐层反向传播网络,设置隐层节点为100个,输入层到隐层和隐层到输出层的传递函数分别为$tansig$函数和$logsig$函数。为获得较快的收敛速度,训练函数采用$traingdx$函数。训练时,设置步长为10^{-4},最小梯度为10^{-10},最大训练次数为1 000;对于LVQ网络,本文设置隐层节点为1 000个,训练时,设置步长为10^{-4},最小梯度为10^{-10},最大训练次数为1 000。训练及分类过程均使用Matlab自带的神经网络工具箱完成。

ELM仅需要设置隐层节点数。在[10,100]以10为步进并用十折交叉验证搜索最佳节点数。考虑到函数的可微性和对数据样本的表示能力,激励函数选择$Sigmoid$函数。

3.4　实验结果

将各分类算法分别在平衡的数据集1和数据集2上进行十折交叉验证,根据上文所述的公式(5)~公式(9),评价两数据集上的分类效果分别如表1和表2所示。

在分类准确度方面,由表1和表2可以看出,在不同的数据集上,各算法均可达到较高的总体精准度且没有显著差距。但是,考虑到肿瘤检测对正负样本各自检出率更为关注,因此应把灵敏度和特异性等指标作为更重要的评价标准。又考虑到两数据集中的正样本为医学上更需要关注的部分,将正样本误判为负样本的代价要远大于将负样本判为正样本的代价,因此灵敏度、F值和G值往往是受到更多关注的评判标准。

Table 1　Performance of classifiers on dataset 1

表1　各分类算法在数据集1上的分类效果

方法	准确率/%	灵敏度/%	特异性/%	训练时间/s	测试时间/s	F值/%	G值/%
LDA	86.15	87.45	85.82	0.05	0.024	86.8	86.8
SVM	87.13	85.31	87.60	80.7	0.03	86.21	86.22
ANN-BP	89.73	71.35	94.49	5.41	0.043	79.49	80.01
ANN-LVQ	82.14	82.35	82.09	78.2	0.072	82.24	82.25
ELM	86.61	84.94	87.04	0.11	0.09	85.77	85.77

Table 2　Performance of classifiers on dataset 2

表2　各分类算法在数据集2上的分类效果

方法	准确率/%	灵敏度/%	特异性/%	训练时间/s	测试时间/s	F值/%	G值/%
LDA	96.68	94.37	98.05	0.023	0.015	95.51	95.52
SVM	96.32	97.64	95.53	59.67	0.021	96.98	96.98
ANN-BP	95.08	88.94	98.09	3.07	0.039	91.9	91.96
ANN-LVQ	93.31	92.86	97.22	50.81	0.061	94.99	95.01
ELM	95.67	93.61	96.92	0.12	0.08	94.63	94.78

由实验结果可知,两种人工神经网络分类器得到的灵敏度在两个数据集上均低于其他三种算法,而其特异性指标均较高,而其他三种算法的特异性指标相差不大,但明显低于人工神经网络。

考察两种神经网络算法,可以看到,LVQ的灵敏度性能明显好于BP但与其他几种分类器之间仍然存在较大差距。除人工神经网络之外,在数据集1中,极限学习机的性能较为平均,而线性判别分析拥有最高的灵敏度、F值和G值,支持向量机的特异性指标最好;在数据集2中,除人工神经网络之外的各算法特异性指标相差不大,而支持向量机的灵敏度、F值和G值均远高于线性判别分析和极限学习机。

在执行效率方面,考察各种算法的耗时。在训练耗时方面,LDA拥有最好的执行效率,ELM速度也相对较快,人工神经网络相对较慢但效率仍然明显高于SVM;而在测试耗时方面,各分类器均有较快的速度,执行效率均可满足需求。

3.5　实验分析

本文对几种经典机器学习算法在两种乳腺肿

瘤数据上的分类表现进行了测试和评价。

线性判别分析拥有极高的执行效率,而在其他几项指标中也均表现出了良好的性能。在训练所用特征选择得当的情况下,测试性能十分理想;但该分类器对特征选择较为敏感,若特征选择不当可能导致投影后两类样本不可分,此时 LDA 的性能将出现明显下降。

支持向量机在整体准确率、灵敏度、特异性这几项指标中均表现出较为理想的性能,虽然由于优化参数而耗时较长、训练效率远低于其他三种算法,但其实际测试效率依然很高,总体性能可以满足分类需求。

人工神经网络算法虽然可以得到较高的整体准确率,但其灵敏度指标明显低于其他算法,即经人工神经网络算法训练和测试后的结果中假阳性比例极高,其更高的整体准确率是由于特异性极高所导致的,这一现象使得人工神经网络算法得到的整体准确率较高。但是,过低的灵敏度会对肿瘤检测结果的可信度产生较大影响,因此该种分类器得到的整体准确率并不具有足够的代表性,虽然拥有最好的整体准确率,但综合考虑整体准确率和灵敏度之后得到的 F 值和 G 值均不够理想,无法准确诊断出恶性肿瘤。而对于 BP 和 LVQ 两种经典人工神经网络分类器,作者在实验中发现,相比于 LVQ,BP 神经网络的单次迭代时长不够稳定,这表明 BP 网络在训练中更容易陷入局部最优,影响其整体性能。

极限学习机拥有较快的执行效率和非常平均的分类性能。与人工神经网络相比,极限学习机结构更加简单,参数调节过程耗时更短,计算复杂度较低,且在许多场合其分类性能甚至更好,是一种具有良好应用前景的分类器。

4 结束语

本文将机器学习领域常见的各种算法应用于乳腺肿瘤分类并进行比较与研究。在提取各种特征的基础上,运用极限学习机、支持向量机、线性判别分析和人工神经网络等主要机器学习分类器对肿瘤数据进行分类识别并从整体准确率、灵敏度、特异性、F 值、G 值以及执行效率等方面对各算法的性能进行评估比较。结果表明,极限学习机和线性判别分析两种算法在各项指标中均表现出较为理想的性能,而人工神经网络在特异性方面表现出了显著优势;支持向量机虽然训练耗时较长,但测试耗时较短,且在其他各项指标下均有较为理想的性能,表现出了较好的稳定性,依然是乳腺肿瘤分类中可行性较强的分类器。

参考文献:

[1] Mitchell T. Machine learning[M]. Beijing: China Machine Press,2003.

[2] Ganesan K,Acharya U R,Chua C K,et al. Computer-aided breast cancer detection using mammograms: A review[J]. IEEE Reviews in Biomedical Engineering,2013,6(77):98.

[3] Orozco H M,Villegas O V,Maynez L O,et al. Lung nodule classification in frequency domain using support vector machines[C]// Proc of 2012 11th International Conference on Information Science,Signal Processing and their Applications (ISSPA),2012:870-875.

[4] Pinheiro F M R,Kuo M H. Poster: Applying data mining algorithms to early detection of liver cancer[C]// Proc of 2012 IEEE 2nd International Conference on Computational Advances in Bio and Medical Sciences (ICCABS),2012:1.

[5] Pei Cheng-dan,Xu Sheng-zhou. Segmentation of mammography images based on the label controlling watershed algorithm[J]. Science Technology and Engineering,2013,13(5):1210-1214. (in Chinese)

[6] Wang Zhi-qiong,Kang Yan,Yu Ge,et al. Breast tumor detection algorithm based on feature selection ELM[J]. Journal of Northeastern University(Natural Science),2013,34(6):792-796. (in Chinese)

[7] Japkowicz N. Learning from imbalanced data sets: A comparison of various strategies[C]// Proc of AAAI'2000 Workshop on Learning from Imbalanced Data Sets,2000:10-15.

[8] Tan P N,Steinbach M,Kumar V. Introduction to data mining [M]. Beijing: Posts & Telecom Press,2006:426.

[9] Duda R O. Pattern recognition[M]. 2nd edition. Beijing: China Machine Press,2004.

[10] Liu X M,Tang J S. Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method[J]. IEEE Systems Journal,2014,8(3):910-920.

[11] Li Zhen-xiang,Wang Wen-jian,Guo Hu-sheng, et al. SVM classification algorithm for solving multi-class imbalance data[J]. Computer Engineering and Design,2014,35(7):2499-2503. (in Chinese)

[12] Zhang Zheng,Wang Yan-ping,Xue Gui-xiang. Digital image processing and machine vision: Implemented by Visual C++ and Matlab[M]. Beijing: Posts & Telecom Press,2010. (in Chinese)

[13] Harrington P. Machine learning in action[M]. Beijing: Posts and Telecom Press,2013.

[14] Wang Zhi-hui. Research on BP neural networks and ELM algorithms[D]. Hangzhou: China Jiliang University,2012. (in Chinese)

[15] Haykin S. Neural networks and learning machines[M]. third

Edition. Beijing:China Machine Press,2009.

[16] Yang Xiao-fan,Chen Ting-huai. Advantages and disadvantages of artificial neural networks[J]. Computer Science,1994,21(2):23-26. (in Chinese)

[17] Huang G B,Zhou H M,Ding X J,et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems,Man,and Cybernetics,Part B:Cybernetics,2012,42(2):513-529.

[18] Huang G B,Zhu Q Y,Siew C K. Extreme learning machine:A new learning scheme of feedforward neural networks[C]// Proc of 2004 IEEE International Joint Conference on Neural Networks,2004:25-29.

[19] Menaka K,Karpagavalli S. Mammogram classification using extreme learning machine and genetic programming[C]// Proc of 2014 International Conference on Computer Communication and Informatics (ICCCI),2014:1-7.

[20] Heath M,Bowyer K,Kopans D,et al. The digital database for screening mammography[C]// Proc of the 5th International Workshop on Digital Mammography,2000:212-218.

[21] Cascio D,Fauci F,Magro R,et al. Mammogram segmentation by contour searching and mass lesions classification with neural network[J]. IEEE Transactions on Nuclear Science,2006,53(5):2827-2833.

[22] Pradeep N,Girisha H,Sreepathi B,et al. Feature extraction of mammograms[J]. International Journal of Bioinformatics Research,2012,4(1):241-244.

[23] Lichman M. UCI machine learning repository [DB/OL]. [2015-07-01]. http://archive.ics.uci.edu/ml. Irvine.

[24] Chang C C,Lin C J. LIBSVM:A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(23):1-27.

附中文参考文献:

[5] 裴承丹,徐胜舟.基于标记控制分水岭算法的乳腺X线摄片分割[J].科学技术与工程,2013,13(5):1210-1214.

[6] 王之琼,康雁,于戈,等.基于特征选择ELM的乳腺肿块检测算法[J].东北大学学报(自然科学版),2013,34(6):792-796.

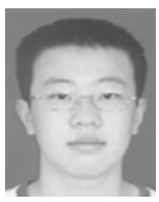
[11] 李珍香,王文剑,郭虎升,等.处理多类不平衡数据的SVM分类算法[J].计算机工程与设计,2014,35(7):2499-2503.

[12] 张铮,王艳平,薛桂香.数字图像处理与机器视觉:Visual C++与Matlab处理[M].北京:人民邮电出版社,2010.

[14] 王智慧.BP神经网络和ELM算法研究[D].杭州:中国计量学院,2012.

[16] 杨晓帆,陈廷槐.人工神经网络固有的优点和缺点[J].计算机科学,1994,21(2):23-26.

作者简介:



李喆(1992-),男,山西太原人,硕士生,研究方向为模式识别、数据挖掘和图像处理。**E-mail**:tywzlizhe29121@126.com

LI Zhe, born in 1992, MS candidate, his research interests include pattern recognition, data mining, and image processing.

吕卫(1976-),男,江苏常熟人,博士,副教授,研究方向为数字视频技术、嵌入式系统设计和模式识别。**E-mail**:luwei@tju.edu.cn

LÜ Wei, born in 1976, PhD, associate professor, his research interests include digital video technology, embedded system design, and pattern recognition.

闵行(1990-),女,辽宁营口人,硕士,研究方向为图像处理和模式识别。**E-mail**:minhang@tju.edu.cn

MIN Hang, born in 1990, MS, her research interests include digital image processing, and pattern recognition.

褚晶辉(1969-),女,天津人,博士,副教授,研究方向为数字视频技术和模式识别。**E-mail**:cjh@tju.edu.cn

CHU Jing-hui, born in 1969, PhD, associate professor, her research interests include digital video technology, and pattern recognition.