

基于机器学习的儿童颅骨畸形分类研究

张顺雨, 胡 骏, 林 勇

(上海理工大学健康科学与工程学院, 上海 200093)

✉ zsy3216@163.com; hujun@163.com; yong_lynn@163.com



摘 要: 儿童颅骨畸形的判断是典型的机器学习分类问题, 文章对电子计算机断层扫描 (Computed Tomography, CT) 图像进行三维建模和数据增强, 从中提取 9 个数值特征并分析其在正常颅骨与畸形颅骨之间的区别, 表明所选特征的有效性。选用 Stacking (堆叠) 方法构建针对颅骨畸形自动判断的集成学习分类模型, 将其应用到医院提供的临床 CT 数据集进行十折交叉验证实验。实验结果表明, 集成学习模型优于单一的分类模型, 其平均预测准确度约为 0.986, F1 值 (统计学中用于衡量分类模型精确度的指标, 同时兼顾了分类模型的准确率和召回率) 约为 0.982, 各项评价指标也优于目前同类模型。

关键词: 颅骨畸形; 集成学习; 图像处理; 分类预测

中图分类号: TP311.5 **文献标志码:** A

Research on Cranial Deformities Classification of Children Based on Machine Learning

ZHANG Shunyu, HU Jun, LIN Yong

(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

✉ zsy3216@163.com; hujun@163.com; yong_lynn@163.com

Abstract: Diagnosis of children cranial deformities is a typical classification problem of machine learning. This paper proposes to conduct three-dimensional modeling and data augmentation on CT (Computed Tomography) images, from which nine numerical features are extracted, and their differences between normal and deformed cranial bones are analyzed, demonstrating the effectiveness of the selected features. The stacking method is selected to construct an ensemble learning classification model for automatic detection of cranial deformities, and it is applied to the clinical CT dataset provided by a certain hospital for ten-fold cross-validation experiments. The experimental results show that the ensemble learning model is superior to a single classification model, with an average prediction accuracy of about 0.986 and an F1 value (a statistical indicator used to measure the accuracy of classification models, while also considering the accuracy and recall of classification models) of about 0.982. All of its evaluation indicators are also better than similar models currently available.

Key words: cranial deformities; ensemble learning; image processing; classification prediction

0 引言 (Introduction)

颅骨畸形^[1]并不仅仅是外观问题, 更是影响儿童健康发育的重要因素。颅骨畸形会改变婴幼儿面部形态, 严重的畸形可能影响大脑发育, 造成发育滞后的问题。颅缝早闭^[2]是导致颅骨变形的常见原因, 全球每 10 000 名新生儿中就有 3.1~6.4 人患病。目前, 我国对儿童颅骨畸形的诊断和治疗未有统一标

准, 诊断方法包括手工测量和 CT 临床检测以及近期出现的光学三维建模技术, 手工测量和临床检测存在速度慢、过于依赖医生经验、患儿配合性差且测量精度不高等问题, 光学三维建模诊断的精度也较难保证, CT 图像具有较高精度, 但也需要依赖检验医生读片诊断^[3]。因此, 设计一套能够自动对获得的 CT 数据进行自动分类预测的模型, 用于对儿童颅骨图像的高

精度分类预测,这对通过早发现、早诊断、早手术进而实现挽救患儿生命和改善症状直至彻底治愈具有重要意义。

1 研究现状 (Research status)

目前,机器学习在医学图像领域的应用越来越广泛,但将机器学习用于颅骨畸形的研究并不多见,YOU 等^[4]使用迁移学习方法从 CT 图像自动分类颅骨畸形,该方法首先从 CT 图像切片中分割 3D 颅骨,通过半球投影将 3D 颅骨投影到二维空间以获得二值图像并进行数据扩充;其次在生成的数据集上对预训练的深度学习模型进行微调,最终在数据集上的预测准确度达到了 90% 以上。SABETI 等^[5]将机器学习技术应用于新生儿颅缝早闭识别研究,头部边界采用 GrabCut 算法进行分割和识别,然后计算人体测量指标,如颅骨指数(CI)、颅顶不对称指数(CVAI)、前中线宽度比(AMWR)、前后宽度比(APWR)、左右高度比(LRHR);分类器采用 KNN(K-Nearest Neighbors)、支持向量机(SVM)^[6]、随机森林和 Bagging 集成学习,预测准确率为 0.85~0.92。此研究提出了 5 个常见的数值测量特征,为了进一步丰富对头颅畸形进行量化的数值特征,LEI 等^[7]使用 SVM 和高斯径向基函数考虑了非线性超平面,采用留一法交叉验证优化软间隔参数和高斯宽度;结合 2D 和 3D 形状指数索引对颅骨畸形进行预测,准确率达到 0.958。但是,该研究只进行了单一模型的预测,其精度还有一定的提升空间。

针对上述研究的不足,本文设计了一套基于 Stacking 方法构建的异构分类器模型,并利用图像处理的方式从 CT 数据集中提取出包括 CI、CVAI、AMWR、WPWR、LRHR 等 9 个量化特征,选用 SVM、KNN、随机森林(RandomForest)、XGBoost 作为初级分类器,随机森林作为次级分类器进行集成。实验结果表明,本文构建的分类模型和量化特征在精度和分类性能上均优于之前的研究。

2 材料与方法 (Materials and methods)

2.1 数据集介绍

本研究使用的数据集来自项目合作医院提供的 463 例儿童颅骨 CT 扫描数据集,年龄分布为 0~9 岁,其中 100 例为患有颅缝早闭而导致颅骨畸形的头部 3D 计算机断层扫描(CT)数据,其余 363 例为正常的头颅 CT 数据。

2.2 CT 图像处理

为了从原始的 CT 图像数据集中提取所需的数值特征,需要进行一系列的数字图像操作,具体步骤如下。

为了完整地提取到头颅模型的颅骨特征,并对所有样本进行标准化处理,需要将原始的 CT 扫描切片进行面绘制,生成 pcd 格式的 3D 模型。由于拍摄环境存在噪声[图 1(a)],因此需要对提取到的 pcd 数据进行过滤。本文将最大联通区域提取算法^[8]应用到点云数据中,结果如图 1(b)所示。本研究的兴趣区域仅为头颅部分,有别于临床研究中的头颅划分方式,利用开源库 PCL(Point Cloud Library)进行点云运算,提取出 3D 模型中的 z 轴中点以上的点云数据并保存[图 1(c)]。考虑到原始模型中,异常数据仅为 100 例,对异常模型的 pcd 数据进行逐层抽样[图 1(d)],用此方案扩充后的颅骨畸形样本量达到 323 例,与正常颅骨样本 363 例的数据量基本平衡。

为了方便后续的机器学习和特征提取,需要将三维空间中

记录的 pcd 点云数据投影到 2D 平面中。在提取图像的数值特征时,为了避免因图像像素不同而造成的误差,将所有图像归一化为 512×512 的像素大小。采用随机翻转和旋转的方式对原始数据集进行扩充,扩充后的数据集大小为正常头颅图像样本 969 例、畸形头颅样本 1 164 例。

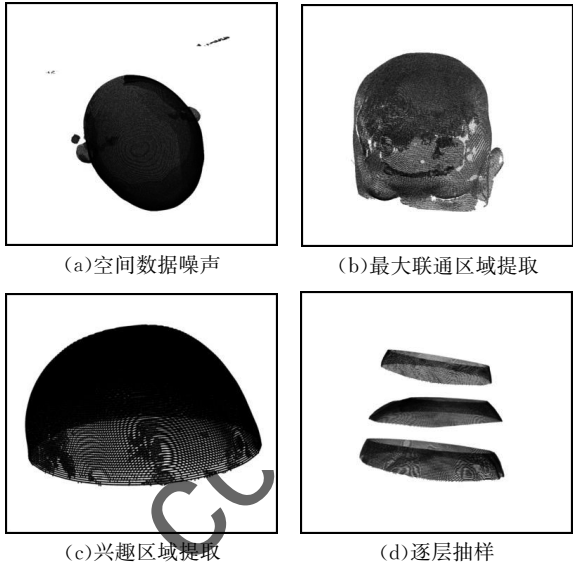


图 1 CT 图像处理结果图

Fig. 1 Diagram of CT image processing results

2.3 特征提取

为了区别由颅缝早闭造成的颅骨畸形和正常颅骨,研究人员参考了目前针对颅骨畸形的研究文献,提出了 9 个数值特征,包括描述正常颅骨和变形颅骨之间突出区域的判别形状特征的 4 个指数 CR、CPS、CES、NCES,以及 5 个颅骨测量指数 CI、CVAI、AMWR、APWR、LRHR;9 个数值特征名称及其描述如表 1 所示。

表 1 9 个数值特征名称及其描述

Tab.1 Names of nine numerical features and their descriptions

特征	描述
CR	颅骨半径指数,计算头颅平面定向边界点与质心距离的傅立叶系数,用于描述头骨中突出区域的位置和程度
CPS	颅骨部分斜率指数,计算头颅平面质心距离之间差值的总和
CES	颅骨极点指数,计算头颅平面边界点距离分布的极值点方差和极值点区域之间的方差,用来描述头骨中突出区域的位置和分布
NCES	近颅极点指数,计算头颅平面边界点之间距离分布的近极值点的方差,用来描述头骨中突出区域的位置和分布
CI	颅骨指数,头颅的最大宽度线和最大长度(AP)线的比值
CVAI	颅穹窿不对称指数,计算颅骨平面两条对角线差值与对角线中最大值之间的比值,用于颅骨畸形评估和分类
AMWR	前中宽度比指数,计算颅骨平面 1/4 分位线和中位线的比值
APWR	前中宽度比指数,计算颅骨平面 3/4 分位线和中位线的比值
LRHR	左右高度比指数,计算颅骨平面 AP 线左侧与 AP 线右侧两条线段的最小值与最大值之比

2.4 特征预处理与分析

提取特征之后,为了保证数据应用于模型的适用性,并减少数据噪声和异常值对模型精度和可靠性的影响,特进行了数据清洗、特征差异分析和正态分布纠正等操作。

本文采用 Tukey 方法^[9]去除异常值,计算数据上下四分位数(Q1, Q3)和四分位距(IQR),根据 Tukey 的规则将位于 1.5 倍 IQR 之外的数据点视为极端值,将其从数据中删除。将去除异常值后的数据绘制成箱线图,来进行正常和异常样本的特征差异分析,数据清洗后样本箱线图如图 2 所示,从图 2 中可见,多数特征具有较高的区分度,表明所选特征对于区分正常样本和异常样本是有效的。

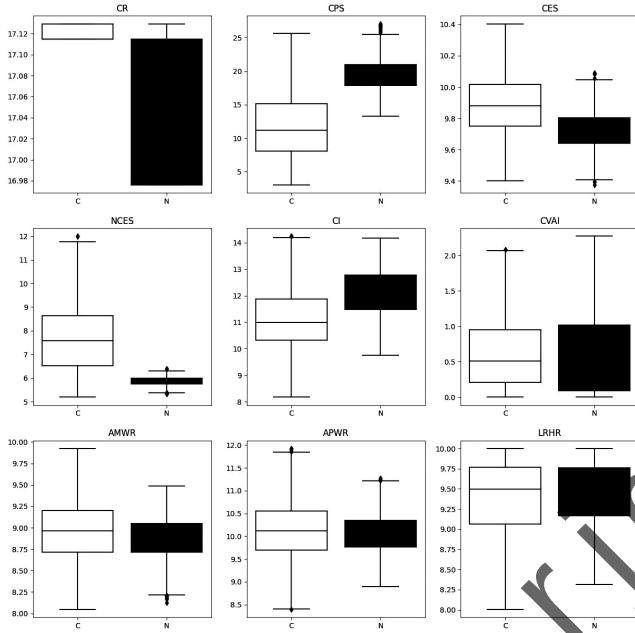


图 2 数据清洗后样本箱线图

Fig. 2 Sample boxplot after data cleaning

颅骨畸形描述特征属于随机变量,从统计学的角度看,随机变量的分布应该满足正态分布,通常用偏度和峰度作为衡量数据符合正态分布形态的指标。由于样本数据量有限和人为进行的数据增强,造成样本分布很难满足正态分布的要求,导致直接针对原始的数据集进行建模效果不佳,需要对数据进行正态分布纠正,以进一步提高模型的准确性和实用性。

分位数归一化转换法^[10]可以将数据转换为具有类似正态分布的形状而不考虑特定的概率分布函数。该方法使用数据经验分布的分位数(数据集中的百分比位置上的数值),具体来说:该转换法首先计算输入数据的经验分布函数,并将其映射到均匀分布。然后应用逆变换函数将均匀分布映射到目标分布。将数据转换为正态分布时,转换的目标分布是正态分布。正态纠正后样本分布统计表如表 2 所示。其中,偏度越接近 0,则越接近正态分布,峰度为正表示数据分布更加陡峭或集中,而峰度为负表示分布更加平坦或散开。由表 2 可知,大部分特征经过纠正后基本能符合正态分布。

表 2 正态纠正后样本分布统计表

Tab.2 Normal corrected sample distribution statistical table

特征	正常样本		畸形样本	
	偏度	峰度	偏度	峰度
CR	0.423	-1.067	0.789	-1.107
CPS	1.317	0.993	-0.001	0.840
CES	-0.001	1.865	0.001	1.585
NCES	-0.098	1.446	0.001	0.846
CI	-0.001	1.871	-0.079	1.223
CVAI	-1.528	5.438	-0.397	2.575
AMWR	-1.468	0.993	0.009	0.863
APWR	-7.099	0.993	0.002	0.837
LRHR	0.895	4.382	1.152	4.967

2.5 畸形分类模型构建

集成学习通过合并多个学习器进一步提升预测准确性。本文采用 Stacking^[11]集成的方式,首先从原始训练集训练出初级学习器,其次将初级学习器的预测结果作为新的数据集训练次级学习器。在新的数据集中,初级学习器的输入作为输入特征,而样本标记与原始数据集中的样本标记一致,本文使用的初级学习器使用不同的学习算法产生,即初级集成是异构的。在训练阶段,次级训练集是用初级学习器生成的,若直接用全部的原始数据集训练次级训练集,就会增大过拟合的风险,本实验采用 k 折交叉验证的方式生成次级训练集,具体过程如下。

为了使每次划分的训练集足够大,本文将初始训练集 D 随机划分成 k 个大小相同的子集 $D_1, D_2, \dots, D_k$ 。 D_j 表示第 j 折的测试集, $\bar{D}_j$ 表示第 j 折的训练集,其表示如下:

$$\bar{D}_j = \frac{D}{D_j}, j \in [1, 5] \quad (1)$$

本文给定 T 个初级学习算法,初级分类器 $l_t^{(j)}$ 为在训练集 $\bar{D}_j$ 上使用第 t 个分类算法所得。其中, D_j 中的每个样本 x_i 为计算出的 9 维特征向量。初级分类器的预测结果为 p_{it} , 表示如下:

$$p_{it} = l_t^{(j)}(x_i), t \in [1, 10] \quad (2)$$

由 x_i 产生的次级训练样本集示例为 $p_i = (p_{i1}, p_{i2}, p_{i3}, \dots, p_{iT})$, 标记部分 y_i 为样本的类别,其中正常样本标记为 0, 异常样本标记为 1。于是,在整个交叉验证过程结束后,从 T 个分类器产生的次级训练数据集如下:

$$D' = \{(l_i, y_i)\}_{i=1}^m \quad (3)$$

将 D' 用于训练次级分类器 ST_j , 本实验预选的次级分类算法的集合如表 3 所示,将选择预测精度最高的预选次级分类算法训练次级分类器,最终得到本文基于 Stacking 的异构分类器。

2.6 评估指标

为验证不同分类模型的有效性,本文采用准确率(Accuracy)、查准率(Precision)、召回率(Recall)和 F1 度量对不同的模型进行评估。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP}$$

(5)

$$Recall = \frac{TP}{TP + FN}$$

(6)

$$F1 = \frac{2 \times Precision \times Recall}{Recall + Precision}$$

(7)

对于二分类问题, TP 为真阳性, 表示头颅畸形的样本预测结果也是畸形的; TN 为真阴性, 表示头颅正常的样本预测结果也是正常的; FP 为假阳性, 表示将正常的样本预测为畸形; FN 为假阴性, 表示将畸形的样本预测为正常。

3 实验结果与分析 (Experimental result and analysis)

对本文使用的初级分类器进行十折交叉验证, 计算准确率、查准率、召回率和 $F1$ 度量在 10 次验证过程中的平均得分以及标准差, 初级分类性能评价指标得分均值与标准差 ($\bar{x} \pm s$) 如表 3 所示。

表3 初级分类性能评价指标得分均值与标准差 ($\bar{x} \pm s$)
Tab.3 The mean and standard deviation of scores for primary classification performance evaluation indexes ($\bar{x} \pm s$)

初级分类器	Accuracy	Precision	Recall	F1 值
SVM	0.822±0.019	0.957±0.018	0.971±0.011	0.966±0.009
KNN	0.955±0.011	0.978±0.012	0.961±0.014	0.968±0.015
LR	0.712±0.045	0.715±0.049	0.716±0.049	0.716±0.049
GaussianNB	0.786±0.066	0.786±0.066	0.786±0.066	0.786±0.066
BernoulliNB	0.545±0.001	0.621±0.040	0.588±0.035	0.571±0.046
SGD	0.813±0.053	0.815±0.049	0.796±0.069	0.680±0.059
RandomForest	0.859±0.045	0.881±0.045	0.873±0.044	0.855±0.045
EXTree	0.900±0.041	0.912±0.035	0.907±0.051	0.904±0.038
XGBoost	0.912±0.038	0.913±0.039	0.912±0.038	0.912±0.038

由表 3 可知, 排名前 4 的初级分类器分别为 SVM、KNN、RandomForest、XGBoost, 其精度分别约为 0.822、0.955、0.859、0.912, 性能优于其余的候选分类器。用排名最靠前的 4 个异构的初级分类器对原始训练集进行预测, 将预测的结果作为新的训练集用于训练次级分类器, 并选出性能最优的次级分类器。次级分类器性能评价指标得分及标准差如表 4 所示。由表 4 中的数据可以看出, 当用 RandomForest 作为次级分类器时, 集成学习分类器的准确率约为 0.984, 相比其他次级分类器 SVM (准确率约为 0.981)、LR (准确率约为 0.982)、Tree (准确率约为 0.981), 分别提升了 0.003、0.002、0.003。

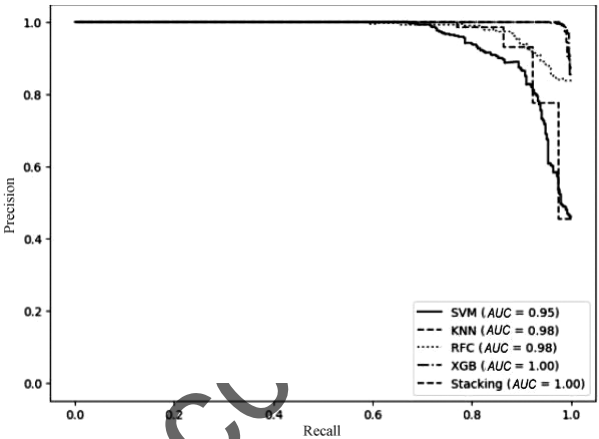
表4 次级分类器性能评价指标得分与标准差 ($\bar{x} \pm s$)

Tab.4 The score and standard deviation of performance evaluation for indexes for secondary classifiers ($\bar{x} \pm s$)

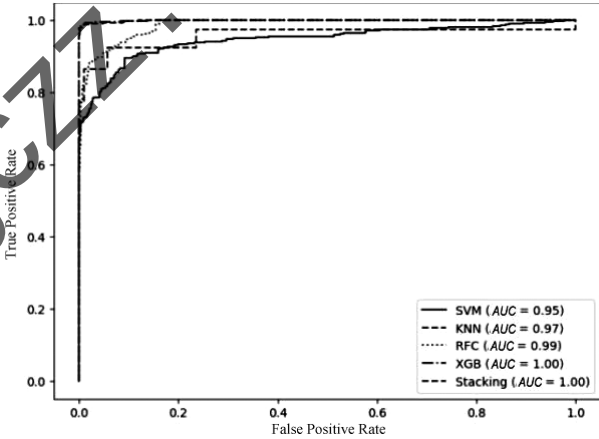
次级分类器	Accuracy	Precision	Recall	F1 值
SVM	0.981±0.016	0.988±0.021	0.970±0.026	0.979±0.018
LR	0.982±0.016	0.988±0.021	0.972±0.026	0.980±0.018
Tree	0.981±0.009	0.984±0.020	0.974±0.017	0.979±0.010
RandomForest	0.984±0.012	0.986±0.020	0.979±0.015	0.982±0.082

相比仅使用单一分类的模型中准确率最高, 本文所提模型的精度则至少提升了 0.1 左右。综合以上数据, 本文提出的将 SVM、KNN、RandomForest、XGBoost 作为初级分类器, 将 RandomForest 作为次级分类器的 Stacking 方法所集成的异构分类器的性能最优, 精度约为 0.984, 准确率约为 0.986, 召回

率约为 0.979, $F1$ 得分约为 0.982。为了更清楚地比较上述次级分类器和集成分类器的性能, 特绘制出各分类器的 ROC (受试者工作特征) 曲线以及 PR (查准率-查全率) 曲线 (图 3)。ROC 曲线边界越靠近左上角边界, AUC (曲线下面积) 值越大, 就表示分类器性能越好, PR 曲线直观地展示了学习器在总体样本上的准确率和召回率。由图 3 可见 Stacking 异构分类器的 PR 曲线完全包住了其余分类器, 可以判断 Stacking 分类器的性能最优。



(a) PR 曲线



(b) ROC 曲线

图3 ROC 曲线与 PR 曲线

Fig. 3 PR curve and ROC curve

4 结论 (Conclusion)

本研究实现了一种用于自动判断颅骨畸形的集成学习方法, 该方法首先对原始的 CT 图像数据集进行格式转化, 然后进行 3D 模型清洗、兴趣区域提取、类别平衡、2D 投影及数据扩充等一系列图像处理工作, 利用数字图像处理的方式提取出用于训练畸形判别分类模型的 9 个数值特征, 并构建训练集和测试集, 最后对比了 9 种常见的初级分类器在训练集上进行十折交叉验证的评价得分, 选出了 SVM、KNN、RandomForest、XGBoost 四种性能最优的初级分类器, 用于训练次级训练器所需的新的训练集, 并在新的训练集上选出最佳的次级分类器, 至此构建了初级分类器为 SVM、KNN、RandomForest、XGBoost, 次级分类器为 RandomForest 的 Stacking 异构集成学习分类模型。实验结果表明, 本文实现的基于 Stacking 的异构集成学习分类模型能够很好地地区别头颅畸形样本和正常样本, 并且与单一分类模型相比精度更高、性能更好。

(下转第 54 页)