

基于机器学习心肌梗死患者的心磁信号诊断

赵永鹏, 朱俊杰

(河南理工大学电气工程与自动化学院, 河南 焦作 454000)

✉2959415512@qq.com; junjiezhu@hpu.edu.cn



摘 要: 使用机器学习方法对心磁数据样本有无疾病进行诊断分类。首先从心磁数据中提取除极阶段(TT间隔)的数据构建磁场图, 然后求解电流密度图, 从电流密度图中提取相关的磁场特征。针对非平衡数据分类问题, 分别使用样本加权的SVM、LR、KNN、Adaboost和XGBoost五种学习模型进行训练, 在此基础上设计了使用加权的LR和KNN为初级学习器、SVM为次级学习器的结合学习模型对样本数据进行训练。采用结合学习模型对73名非患者和47名心肌梗死患者的36通道心磁数据进行实验, 结果显示该模型对样本不平衡的心磁数据分类有较好的效果。

关键词: 心磁数据; 电流密度图; 机器学习; 心肌梗死; 结合学习

中图分类号: TP391 **文献标识码:** A

Magnetocardiograph Signal Diagnosis of Patients with Myocardial Infarction based on Machine Learning

ZHAO Yongpeng, ZHU Junjie

(School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China)

✉2959415512@qq.com; junjiezhu@hpu.edu.cn

Abstract: This paper proposes to classify the presence or absence of diseases in magnetocardiograph data samples using machine learning method. First, magnetic field map, constructed by depolarization phase (TT interval) data, is extracted from the cardiac magnetic data. Then current density map is solved. From the current density map, the relevant magnetic field characteristics are extracted. Aiming at the problem of unbalanced data classification, five learning models of sample-weighted SVM (Support Vector Machine), LR (Logistic Regression), KNN (K-Nearest Neighbors), Adaboost and XGBoost are used for training. On this basis, weighted LR and KNN are designed as the primary learners, and weighted SVM as the secondary. The stacking model of the learners trains the sample data. The stacking learning model is used to conduct experiments on the 36-channel magnetocardiograph data of 73 non-patients and 47 myocardial infarction patients. The results show that the model has a good effect on the classification of unbalanced samples of the cardiogram data.

Keywords: magnetocardiograph data; current density map; machine learning; myocardial infarction; stacking learning

1 引言(Introduction)

目前, 缺血性心脏病是导致人类死亡的主要原因。缺血性心脏病的发生是由于冠状动脉狭窄, 流向心脏的血流受到限制, 从而引起心肌损伤, 严重的可能导致心肌猝死。所以, 早期诊断缺血性心脏病对降低死亡率至关重要。

在早期检测缺血性心脏病中, 心磁图成为一种很有前景

的无创诊断工具。心磁图携带心电活动的信息, 能够在一定程度上反映心脏功能的异常。与需要接触体表测量电位的心电图相比, 无接触测量的心磁图能够提供更高的时空分辨率的心电活动定位, 并且已经在临床试验中表现出较好的效果^[1-7]。尽管心磁图具有优越的信号质量, 但需要对它进行解释的工作量很大, 并且强烈依赖于专业医生的经验, 限制了临床的使用

度。因此,一种能够检测诊断早期缺血性心脏病的方法对临床医生来说是非常有帮助的。

近年来,机器学习的进展显示了其在自动检测缺血性心脏病方面的优势。KANGWANARIYAKUL等^[8]在缺血性心脏病的检测中比较了不同的机器学习方法,使用贝叶斯神经网络(BNN)实现了敏感度为96.65%,特异度为86.36%。TAN等^[9]从心磁信号的ST段中提取了12个形态特征,使用SVM分类器,他们获得的敏感度为95.20%,特异度为93.29%。

二维电流密度图是一种广泛应用于检测缺血性心脏病的直观方法。使用不同的方法从电流密度图中提取相关的特征参数,对早期缺血性心脏病进行诊断检测,已取得较好的诊断结果^[10-12]。我们从二维电流密度图中提取最大电流密度,并把除极阶段(TT间隔)的所有最大电流密度进行平均。从平均之后的最大电流密度中提取四个特征参数。针对非平衡数据,分别使用样本加权的SVM、LR、KNN、Adaboost和XGBoost五种学习模型对提取的特征进行训练学习。在此基础上,提出了使用加权的LR和KNN为初级学习器、SVM为次级学习器的结合学习模型,并在实测的心磁数据上验证了该分类模型的有效性。

2 数据与特征(Data and features)

该实验的数据是由120个人的心磁数据组成,其中包含73名非患者,47名心肌缺血患者。心磁数据是通过超量子干涉仪(SQUID)在人体胸部上方的36个相邻位置,以1,000 Hz的频率连续测量而获取的。根据心磁数据构建出心磁图(MCG),如图1所示。在心磁图上定位出R峰(R-peak)、T波开始(T-onset)、T波结束(T-end)及T波的峰值(T-peak)。由于缺血性心脏病与T波密切相关,因此我们只分析TT间隔(TT-interval)的数据,即T波开始的时刻和T波结束的时刻之间的心磁数据。

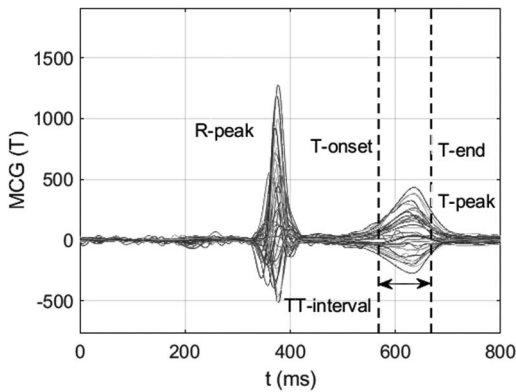


图1 36通道单周期心磁信号的蝶形图

Fig.1 Butterfly diagram of 36-channel single cycle magnetocardiograph signal

二维电流密度图中的矢量大小和方向是从心脏磁场 B_z 分量测量得到的,测量平面与 z 分量垂直。因此,可由公式(1)得出二维电流密度图^[13-14]。

$$\vec{v}(x, y, k) = \left(\frac{\partial B_z(x, y, k)}{\partial y}, -\frac{\partial B_z(x, y, k)}{\partial x} \right) \quad (1)$$

其中, k 为采样时间, (x, y) 为测量平面上的任意位置。

把单一时刻的心磁数据经过三次样条插值的处理,进而得到心磁数据的等磁场图。在等磁场图的基础上求解二维电流密度图,并提取电流密度图中最大电流密度,如图2所示,最大电流密度用粗箭头表示。

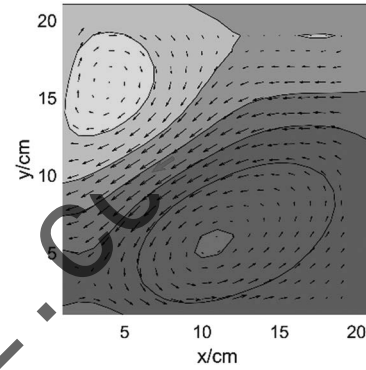


图2 二维电流密度图

Fig.2 Two-dimensional current density diagram

然后,求单周期36通道心磁图中TT间隔时间内,所有二维电流密度图中最大电流密度的平均值,如公式(2)所示:

$$\bar{V} = \frac{1}{N} \sum_{k=1}^N \vec{v}_{\max}(k) \quad (2)$$

其中, N 为TT间隔的时间长度, k 为采样时间, $\bar{V}$ 为所有电流密度图中最大电流密度的平均值。

从最大电流密度的平均值 $\bar{V}$ 中提取四个参数,分别为:

(1) $\bar{V}$ 的长度: $V = \sqrt{V_x^2 + V_y^2}$;

(2) $\bar{V}$ 向量的角度: $\theta = \arctan\left(\frac{V_y}{V_x}\right)$;

(3) 由 $\bar{V}$ 向量构成矩形的周长: $P = 2(V_x + V_y)$;

(4) 由 $\bar{V}$ 向量构成矩形的面积: $A = V_x \times V_y$ 。

其中, V_x 、 V_y 是 $\bar{V}$ 在 x 、 y 方向上的分量。由于坐标系和测量系统相对于测量者都处于相同的位置,因此不同的测量者产生的这些参数具有可比性。

3 方法(Methods)

3.1 基于加权分类的机器学习方法

选择使用传统机器学习模型中的SVM、LR、KNN、Adaboost和XGBoost来对提取的特征进行学习分析,因为这五种学习模型应用于中小型数据集,并且能够处理高维数据。另外,由于试验的样本中非患者与患者的人数比不平衡,因此使用带有加权分类的方法进行机器学习,从而使模

型更加注重样本数量少的类别。如果类别的样本数量多,那么它的权重就低,反之则权重就高。

$$\omega_j = \frac{n}{kn_j} \quad (3)$$

其中, ω_j 为类别 j 对应权值, n 为数据总数, k 为类别数量, 即数据有 k 个种类, n_j 是类别 j 的数据个数。

使用以上五种带有样本加权的机器学习模型对标准化后的训练样本进行交叉验证, 并对预测样本进行预测。分别绘制五种学习模型测试样本的接受者操作特征曲线(Receiver Operator Characteristic, ROC), 如图3所示。

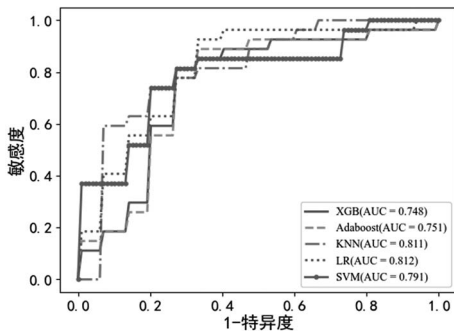


图3 五种学习模型分别对应的ROC曲线图

Fig.3 ROC curves of five learning models

ROC曲线是通过敏感度与1-特异度在各种阈值下绘制的。ROC曲线越靠近左上角, 说明模型的预测精度越高, 通常采用曲线下的面积(Area under Curve, AUC)值表示预测结果的概率。

3.2 基于加权的结合(Stacking)学习

从统计方面来看, 由于学习任务的假设空间往往很大, 可能有多个假设在训练集上达到同等性能, 此时若使用单学习器可能因误选而导致泛化性能不佳, 结合多个学习器则会减小这一风险, 这一策略就是结合策略^[15]。在Stacking学习中, 个体学习器称为初级学习器, 用于结合的学习器称为次级学习器。在上面的五种机器学习模型中, 支持性向量机SVM、LR和KNN的结果较好, 所以设计了使用LR和KNN为初级学习器, SVM为次级学习器。其中作为初级学习器的LR和KNN仍然进行样本加权处理, 而SVM不用进行加权处理。Stacking学习具体步骤如下:

第一步初级训练阶段: 分别使用加权后的LR和KNN学习器对训练样本进行交叉验证学习, 使用每一折中训练数据训练模型, 对每一折中的验证数据的标签进行预测。这样在交叉验证后, 可以得到全部折中验证数据标签的预测值。把这些预测值按照顺序进行堆叠, 形成训练数据的新特征。同时, 在每一次折中使用训练数据训练出模型后, 对测试集数据的标签进行预测。交叉验证之后得到全部的预测标签值,

并把全部的预测标签值进行平均, 得到每个测试样本唯一的预测值, 用来作为测试集新特征。

第二步次级分类阶段: 使用核函数为高斯核函数的SVM作为次级学习器, 没有进行样本加权处理。使用第一步中得出的训练数据新特征作为次级学习器的训练数据, 训练出模型后, 对第一步得出的测试集新特征的标签进行预测。不管是第一阶段还是第二阶段, 数据的标签没有变化。

绘制Stacking学习模型测试样本的接受者操作特征曲线(ROC), 如图4所示。

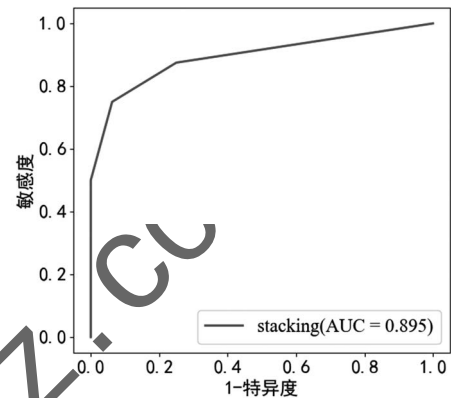


图4 Stacking学习模型分别对应的ROC曲线图

Fig.4 ROC curves of stacking learning model

4 结果与分析(Results and analysis)

我们使用准确率、敏感度、特异度和AUC值作为主要参考的评价指标。此外, 在医学中敏感度为真阳性率, 即实际患者试验结果为阳性的比例; 特异度为真阴性率, 即非患者试验结果为阴性的比例。敏感度和特异度在理论上与疾病患病率无关, 仅由诊断试验本身决定, 不受外界因素的影响。

在表1中, 记录了没有进行样本加权处理的五种学习模型对应ROC曲线下的面积AUC的值, 以及五种学习模型处于最佳阈值时的准确率、敏感度和特异度。

表1 五种学习模型的四个评价指标

Tab.1 Four evaluation indexes of five learning models

分类器	准确率	敏感度	特异度	AUC
SVM	63.1%	77.7%	73.2%	75.4%
LR	62.5%	92.6%	66.8%	80.5%
KNN	70.1%	74.1%	79.8%	81.1%
Adaboost	63.8%	88.9%	66.7%	75.1%
XGBoost	54.1%	81.4%	72.7%	74.1%

在表2中, 记录了进行样本加权的五种学习模型对应ROC曲线下的面积AUC的值, 以及五种学习模型处于最佳阈值时

的准确率、敏感度和特异度。

表2 样本加权的五种学习模型的四个评价指标

Tab.2 Four evaluation indexes of five learning models weighted by samples

分类器	准确率	敏感度	特异度	AUC
SVM	72.2%	81.5%	73.2%	79.1%
LR	73.6%	92.6%	66.8%	81.2%
KNN	71.6%	74.1%	79.8%	81.1%
Adaboost	68.1%	88.9%	66.7%	75.1%
XGBoost	66.7%	81.4%	72.7%	74.8%

在表3中，记录了进行样本加权处理的Stacking学习模型对应ROC曲线下的面积AUC的值，以及模型处于最佳阈值时的准确率、敏感度和特异度。

表3 样本加权的Stacking学习模型的四个评价指标

Tab.3 Four evaluation indexes of stacking learning model weighted by samples

分类器	准确率	敏感度	特异度	AUC
Stacking	87.5%	75%	85.7%	89.4%

从表1至表3可知：

(1)采用样本加权处理的单个机器学习较没有进行样本加权的机器学习的预测效果好。

(2)采用样本加权后的Stacking学习模型能够同时提高准确率、特异度和AUC值，相比采用样本加权处理的单个机器学习和未进行样本加权处理的单个机器学习预测效果好。

5 结论(Conclusion)

本研究中，我们从心磁数据中提取除极阶段(TT间隔)的数据，求解电流密度图，提取磁场特征参数。为了解决非平衡数据分类中易于倾向判别为多数类的问题，本文首先利用进行样本加权处理的SVM、LR、KNN、Adaboost和XGBoost五种学习模型进行学习分析，进而诊断心肌梗死。在此基础上，观察到SVM、KNN和LR这三种学习模型的分类效果较好，所以设计了使用带有加权属性的KNN和LR作为初级学习器，使用不带加权属性的SVM作为次级学习器进行结合学习，并对样本不平衡的心磁信号数据进行分类。该算法能够有效减少数据不平衡对单个分类器分类效果的影响。

本文使用实测MCG数据做了三组实验的比较，结果显示：采用样本加权处理的单个机器学习较没有进行样本加权处理的机器学习的预测效果好；采用样本加权后的Stacking学习模型能够获得较好的预测效果。

参考文献(References)

[1] HAILER B, CHAIKOVSKY I, AUTH-EISERNITZ S, et al. The value of magnetocardiography in patients with and without relevant stenoses of the coronary arteries using an unshielded system[J]. Pacing and Clinical Electrophysiology: PACE, 2005, 28(1):8-16.

[2] HAILER B, LEEUWEN P V, CHAIKOVSKY I, et al. The value of magnetocardiography in the course of coronary intervention[J]. Ann Noninvasive Electrocardiol, 2015, 10(2):188-196.

[3] PARK J W, HILL P M, CHUNG N, et al. Magnetocardiography predicts coronary artery disease in patients with acute chest pain[J]. Annals of Noninvasive Electrocardiology, 2005, 10(3):312-323.

[4] STEINBERG B A, ROGUIN A, WATKINS S P, et al. Magnetocardiogram recordings in a nonshielded environment—reproducibility and ischemia detection[J]. Annals of Noninvasive Electrocardiology, 2005, 10(2):152-160.

[5] TOLSTRUP K, MADSEN B E, RUIZ J A, et al. Non-invasive resting magnetocardiographic imaging for the rapid detection of ischemia in subjects presenting with chest pain[J]. Cardiology, 2006, 106(4):270-276.

[6] PARK J W, LEITHÄUSER B, HILL P, et al. Resting magnetocardiography predicts 3-year mortality in patients presenting with acute chest pain without ST segment elevation[J]. Annals of Noninvasive Electrocardiology, 2008, 13(2):171-179.

[7] KANDORI A, OGATA K, MIYASHITA T, et al. Subtraction magnetocardiogram for detecting coronary heart disease[J]. Annals of Noninvasive Electrocardiology, 2010, 15(4):360-368.

[8] KANGWANARIYAKUL Y, NANTASENAMAT C, TANTIMONGCOLWAT T, et al. Data mining of magnetocardiograms for prediction on ischemic heart disease[J]. Excli Journal, 2010, 9:82-95.

[9] TAN J H, HAGIWARA Y, PANG W, et al. Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals[J]. Computers in Biology and Medicine, 2018, 94:19-26.

[10] OGATA K, KANDORI A, WATANABE Y, et al. Repolarization spatial-time current abnormalities in patients with coronary heart disease[J]. Pacing and Clinical Electrophysiology, 2009,

32(4):516–524.

[11] KANDORI A, OGATA K, MIYASHITA T, et al. Subtraction magnetocardiogram for detecting coronary heart disease[J]. Annals of Noninvasive Electrocardiology, 2010, 15(4):360–368.

[12] KWON H, KIM K, LEE Y H, et al. Non-invasive magnetocardiography for the early diagnosis of coronary artery disease in patients presenting with acute chest pain[J]. Circulation Journal, 2010, 74(7):1424–1430.

[13] COHEN D, EDELSACK E A, Zimmerman J E.

Magnetocardiograms taken inside a shielded room with a superconducting point-contact magnetometer[J]. Appl. Phys. Lett., 1970, 16(7):278–280.

[14] HOSAKA H, COHEN D. Part IV visual determination of generators of the magnetocardiogram[J]. Electrocardiol, 1976, 9(4):426–432.

[15] 周志华.机器学习[M].北京:清华大学出版社,2016:183–185.

作者简介:

赵永鹏(1994–), 男, 硕士生.研究领域: 信号与信息处理.

朱俊杰(1981–), 男, 博士, 讲师.研究领域: 信号与信息处理.

(上接第9页)

- (3)编写迪杰斯特拉算法程序, 作为循环体。
- (4)输出最佳路径坐标, 并存储到数组中。
- (5)将最佳路径系列坐标显示在DataGridView中。

4.3 结果评估

给定起点和终点, 测试程序自动生成的最短路径测试结果如表1所示。Path_name是最短路径的名称, 越复杂的热流道系统, 管道连接的数量越多, 因此加以命名, 以便区分; Point_name是点的名字, 便于人工检测最短路径是否正确; X、Y、Z是点的三个坐标; Type是点的类型, 用于程序识别点的种类。经过检测, 程序获得的最短路径与人工设计的路径完全吻合, 结果正确。

表1 最短路径结果表

Tab.1 Results table of the shortest path

Path_name	Point_name	X	Y	Z	Type
Path0	Star	-11.78	-140	-45	
Path0	交点	-11.78	-238	0	x
Path0	Kagu1	52	-238	0	*
Path0	Kagu2	237.5	-238	0	*
Path0	Dian3	287.5	-238	0	01
Path0	Kagu4	287.5	-188.5	0	*
Path0	End	252.5	-45	-160	

5 结论(Conclusion)

本文基于热流道系统三维模型, 对三维软件Solidworks进行二次开发, 开发一个插件, 能够实现在三维软件中自动计算并生成硬管管道最短路径, 最后将最短路径结果以表格(DataGridView)的形式显示在三维软件中。最短路径的获得是根据图论中典型的单源最短路径算法——迪杰斯特拉算法

(Dijkstra)实现的^[8]。自动生成的硬管管道热流道系统不仅降低了系统的成本, 而且提高了系统的质量和品质, 同时降低了工人安装管道的复杂程度, 并且硬管管道相比软管也更加美观。

参考文献(References)

[1] 赵美勇, 宋思睿. 三种最短路算法的比较[J]. 数码世界, 2019 (06):77.

[2] 刘汝佳. 算法竞赛入门经典[M]. 2版. 北京: 清华大学出版社, 2014:56.

[3] 赖志刚. 求解最短路问题的一个计算机算法解析[J]. 中国新通信, 2019, 21(20):133.

[4] 曾庆红, 杨桥艳. 最短路问题算法综述[J]. 保山学院学报, 2019, 38(05):44–46.

[5] (美)CORMEN T H, LEISERSON C E, RIVEST R L, 等. 算法导论[M]. 殷建平, 徐云, 王刚, 等, 译. 北京: 机械工业出版社, 2012:45–46.

[6] (美)SEdgeWICK R, WAYNE K. 算法[M]. 4版. 谢路云, 译. 北京: 人民邮电出版社, 2012:23–25.

[7] 吴昕宇, 罗雪颖. 基于Dijkstra算法的旅游路线规划研究[J]. 数码世界, 2019(07):33–34.

[8] QU S X. Research on dynamic route guidance system based on comparison of shortest path algorithms[J]. International Core Journal of Engineering, 2020, 6(12):12–14.

作者简介:

王遵义(1993–), 男, 硕士生.研究领域: 计算机辅助设计与智能制造.

仲梁维(1962–), 男, 硕士, 教授.研究领域: 计算机辅助设计, 企业信息化.本文通讯作者.