



(12) 发明专利申请

(10) 申请公布号 CN 111881979 A

(43) 申请公布日 2020. 11. 03

(21) 申请号 202010739336.5

(22) 申请日 2020.07.28

(71) 申请人 复旦大学

地址 200433 上海市杨浦区邯郸路220号

(72) 发明人 张立华 杨鼎康 张沛轩 翟鹏

(74) 专利代理机构 北京纪凯知识产权代理有限公司 11245

代理人 陆惠中 王永伟

(51) Int. Cl.

G06K 9/62 (2006.01)

G06N 3/08 (2006.01)

G06F 9/50 (2006.01)

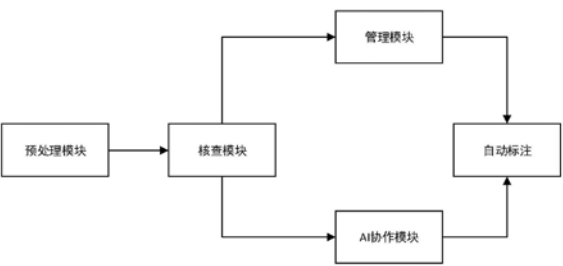
权利要求书1页 说明书5页 附图5页

(54) 发明名称

多模态数据标注装置及包含程序的计算机可读存储介质

(57) 摘要

本发明公开了一种基于患者行为的多模态数据标注装置,包括数据采集设备、标注存储器及处理器,还包括预处理服务器、数据库服务器、AI协作服务器和多图半监督学习算法,该算法由预处理模块、审查模块、管理模块以及AI协作模块四个程序模块组成,旨在利用数据采集设备采集的多模态信息进行信息融合,之后通过融合数据的特征构建成图,然后通过融合图上进行半监督学习,在学习过程中实现患者行为多模态数据的自动标注。本发明克服了训练数据不足和数据冗余较大的问题,实现患者行为多模态数据的自动标注,极大地提高了标注精度。



1. 一种基于患者行为的多模态数据标注装置,包括数据采集设备、标注存储器及处理器,其特征是:还包括预处理服务器、数据库服务器、AI协作服务器和半监督学习算法,该算法由预处理模块、审查模块、管理模块以及AI协作模块四个程序模块组成,旨在利用数据采集设备采集的多模态信息进行信息融合,之后通过融合数据的特征构建成图,然后通过融合多张图集成到一个正则化表达式框架中,同时优化这些图的加权系数,完成多模态数据的自动标注。

2. 如权利要求1所述的基于患者行为的多模态数据标注装置,其特征是:

所述的预处理模块,包括如下步骤:

获取患者行为的不同模态数据,以高层语义和数据集成为基准,将多模态数据构建成符合半监督学习的特征数据集;

根据数据特征中顶点和边的权重随着欧氏距离的变化,确定权重方程和正则表达式;

融合建图,结合多源模态的特征维数选取最优距离度量,生成对应多图。

3. 如权利要求1所述的基于患者行为的多模态数据标注装置,其特征是:

所述的审查模块,包括如下步骤:

利用预处理后的模态信息和距离度量通过AMPQ消息件,与半监督学习的数据反馈实现数据交换;

加入数据编码纠错,结合前向纠错和混合纠错保证数据的完整准确性;

通过控制权重系数实现对算法性能的提升。

4. 如权利要求1所述的基于患者行为的多模态数据标注装置,其特征是:

所述的管理模块,包括如下步骤:

构建若干图表征数据样本之间的连续性关系;

对样本以及相邻样本之间的概念相似概率进行分析和判断;

改变权重相似度的取值。

5. 如权利要求1所述的基于患者行为的多模态数据标注装置,其特征是:

所述的AI协作模块,包括如下步骤:

通过Docker加速矩阵计算速度;

结合ANN阶层完成上述构图检验;

进行传入控制参数更新;

进行半监督学习。

6. 一种包含程序的计算机可读储存介质,其特征是所述程序包括半监督学习算法,该算法由预处理模块、审查模块、管理模块以及AI协作模块四个程序模块组成,旨在利用数据采集设备采集的多模态信息进行信息融合,之后通过融合数据的特征构建成图,然后通过融合多张图集成到一个正则化表达式框架中,同时优化这些图的加权系数,完成多模态数据的自动标注。

多模态数据标注装置及包含程序的计算机可读存储介质

技术领域

[0001] 本发明涉及一种基于患者行为的多模态数据标注装置,还涉及一种包含程序的计算机可读存储介质。

背景技术

[0002] 随着机器学习、行为识别等科学技术的飞速发展,很多医院对于患者行为数据的收集和存储方式变得更加多元化,多模态用来表示不同形态的数据形式,一般表示成病例单的文本信息、基于计算机视觉的表情或情绪图片、病情叙述的音频信号、人体运动三维捕捉的视频以及生理信号的混合数据等等。多模态患者行为标注技术通过现代人机物共融医疗服务采集到的患者图像信息、语音信息、状态信息,包括通过可穿戴传感器设备获取的心跳,姿态,血氧浓度信息等,即利用全方位的不同角度相关信息,描述患者的行为,并及时地对健康风险及重要性进行评估。

[0003] 数据标注作为人工智能学习数据进行加工的一种重要行为,其标注得越精确对算法模型训练的效果就越好。然而由于大规模多模态数据库的匮乏,目前市面上大多数采用的是单模态数据标注方式,传统的数据标注算法有基于拉普拉斯图的GRF算法和基于归一化拉普拉斯的LLGC算法等,现有的标注系统往往仅针对单一的患者行为数据,采用传统的算法进行标注,通常固定在特定的数据领域和概念数据集中,采取特定领域的先验知识,通过提取单一的简单底层特征进行分析判断,完成半自动标注。

[0004] 然而针对复杂医疗场景下的异构多模态患者行为数据,传统算法的标注结果往往突显出性能欠佳和准确率低下的缺陷。

[0005] 同时,现有的标注系统缺乏数据核查和管理环节,在对数据进行简单的预处理后便开始进行特征提取,没有考虑到数据在交换传输过程中容易产生的误码、漏传以及误传等问题;

[0006] 此外,在管理层面仅依靠人工核查和简单数据库管理,形式单一且缺乏准确度,往往需要对对应先验知识储备的人员实现,数据库之间联系不够紧密,缺乏不同模态数据的统一数据库支持和部署,导致成为患者行为诊断和病情评估分析技术上的一大阻碍。

发明内容

[0007] 本发明解决的技术问题是提供一种基于患者行为的多模态数据标注装置,它克服了训练数据不足和数据冗余较大的问题,实现患者行为多模态数据的自动标注,极大地提高了标注精度。

[0008] 本发明的技术方案是,一种基于患者行为的多模态数据标注装置,包括数据采集设备、标注存储器及处理器,其特征是:还包括预处理服务器、数据库服务器、AI协作服务器和多图半监督学习算法,该算法由预处理模块、审查模块、管理模块以及AI协作模块四个程序模块组成,旨在利用数据采集设备采集的多模态信息进行信息融合,之后通过融合数据的特征构建成图,然后通过融合多张图集成到一个正则化表达式框架中,同时优化这些图

的加权系数,完成多模态数据的自动标注。

[0009] 所述的预处理模块,包括如下步骤:

[0010] 获取患者行为的不同模态数据,以高层语义和数据集成为基准,将多模态数据构建成符合半监督学习的特征数据集;

[0011] 根据数据特征中顶点和边的权重随着欧氏距离的变化,确定权重方程和正则表达式;

[0012] 融合建图,结合多源模态的特征维数选取最优距离度量,生成对应多图。

[0013] 所述的审查模块,包括如下步骤:

[0014] 利用预处理后的模态信息和距离度量通过AMPQ消息件,与半监督学习的数据反馈实现数据交换;

[0015] 加入数据编码纠错,结合前向纠错和混合纠错保证数据的完整准确性;

[0016] 通过控制权重系数实现对算法性能的提升。

[0017] 所述的管理模块,包括如下步骤:

[0018] 构建若干图表征数据样本之间的连续性关系;

[0019] 对样本以及相邻样本之间的概念相似概率进行分析和判断;

[0020] 改变权重相似度的取值。

[0021] 所述的AI协作模块,包括如下步骤:

[0022] 通过Docker加速矩阵计算速度;

[0023] 结合ANN阶层完成上述构图检验;

[0024] 进行传入控制参数更新;

[0025] 进行半监督学习。

[0026] 本发明的技术效果是:

[0027] 本发明提供了一种基于患者行为的多模态数据标注系统,包括实现患者行为自动标注的多图半监督学习标注方法MG-SSL的实现和标注预处理、标注审查、标注管理以及标注AI协作四个模块标注系统层次体系的构建。与传统的LLGC等数据标注领域算法相比,MG-SSL算法将多模态数据构建成图,由单图扩展到多图进行基于融合图的半监督学习,很好克服了训练数据不足和数据冗余较大的问题,同时充分考虑到了异构数据之间的距离度量选择与时间相关性,为此具有极强的可扩展性,很容易扩展到多类问题,实现不同医疗场景环境下的多模态数据的自动标注工作;

[0028] 另一方面,本发明通过四个阶段的模块化标注系统层次体系建立,极大的提高了患者行为异构数据结构复杂难点下的标注精度,充分调用了CPU与内存资源,将服务部署到最合适的服务器;采取的多种通讯系统中的数据信道纠错方式,解决了分布式架构中底层与上层数据交换中的可靠性问题;同时利用深度学习辅助智能标注,结合边缘智能思想,通过协同终端设备与边缘服务器,来整合二者的计算本地性与强计算能力的互补性优势,从而达到显著降低深度学习模型推理的延迟与能耗的目的,为患者行为的医疗感知分析系统提供了可靠的数据支撑。

附图说明

[0029] 图1为本发明提供的基于患者行为的多模态数据标注系统结构示意图。

[0030] 图2为本发明提供的多图半监督学习方法MG-SSL。

[0031] 图3为本发明提供的预处理模块流程图。

[0032] 图4为本发明提供的审查模块流程图。

[0033] 图5为本发明提供的管理模块流程图。

[0034] 图6为本发明提供的AI协作模块流程图。

具体实施方式

[0035] 现结合附图和实施例对本发明作进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0036] 参看图1~图2,本发明的基于患者行为的多模态数据标注系统,包括医疗仪器、视频监控、诊断记录设备以及声学设备,预处理服务器、标注存储器等硬件,和多图半监督学习方法MG-SSL,该算法由预处理、审查、管理、和AI协作等程序模块组成。下面分别叙述。

[0037] 其中预处理模块的功能特点是充分考虑多模态数据的异构性和复杂性,通过提取数据特征、构建正则表达式和权重矢量等生成数据之间的距离度量,即初步实现患者行为多模态数据的预处理,供下一步多模态数据的传输和运用;

[0038] 审查模块的功能特点是在数据交换传输阶段实现对预处理后数据的传输纠错,利用纠错编码修正错误数据,确保数据的正确性;

[0039] 管理模块的功能特点是结合多模态和构建的距离度量完成多样本建图,即考虑不同模态样本数据之间的时序性,通过分析样本概念语义完成权重相似度的确定;

[0040] AI协作模块的功能特点是检验构图合理性和进行半监督学习,结合管理模块的建图,完成多模态数据的自动标注。即采用Docker加速算法计算速度,同时结合ANN阶层网络更新控制参数,实现构图协作和权重系数衡量,利用管理模块构建的多图进行半监督学习,最终完成多模态数据的标注。

[0041] 综合以上四个模块所述,多图半监督学习方法MG-SSL指的是利用患者行为多模态数据之间的时空关联性,实现多模态数据的融合和样本距离度量的确定,在此基础上将多模态转换为多图,将使用大量未标记数据进行训练的半监督学习方式迁移到多图训练中,完成多模态数据的自动标注。

[0042] 1) 标注系统首先实现多模态数据的采集,通过医疗仪器、视频监控、诊断记录设备以及声学设备等医疗场景下的医疗设备实现患者行为多模态数据采集。在数据层面理解,可认为是多种数据类型的组合,如图片、数值、音频、符号等等。

[0043] 之后在预处理服务器中,在标注预处理模块通过对上述采集到的患者行为下的多模态数据进行初步建图和融合操作,实现多模态数据的预处理,其具体步骤参看图3所示的如下步骤:

[0044] 图3为本发明提供的预处理模块流程图,具体步骤如下:

[0045] 1.1标注预处理阶段从不同信息源中提取获得患者不同模态后,分析高层语义之间的数据关系,进行数据集成;

[0046] 1.2给定符合半监督学习特征的数据集之后提取数据底层特征,即认为每个数据样本是图上的一个顶点;结合数据特征通过边将有标签和无标签数据点进行连接,每一对顶点之间都有边相连,边的权重随欧式距离增加而降低,根据欧氏距离的变化,选取合适的

权重方程。

[0047] 1.3确定半监督学习的正则化框架表达,取W表示 $n \times n$ 数据之间的仿真,其中 W_{ij} 表示样本 x_i, y_j 之间的相似度。令D为对角阵,确定正则化表达式。

[0048] 1.4通过多模态数据建图得到的G张图,改变正则表达式,获取权重矢量,规范化图拉普拉斯算子满足融合图的权重约束关系如下:

$$[0049] \quad f = \left(I + \frac{\sum_{g=1}^G \alpha_g L_g}{\sum_{g=1}^G \alpha_g \mu_g} \right)^{-1} Y$$

$$[0050] \quad \text{等同于先融合 } L_g, \mu_g \text{ 为 } L_0 = \sum_{g=1}^G \alpha_g L_g, \mu_g = \sum_{g=1}^G \alpha_g \mu_g$$

[0051] 上述公式里其中f表示正则表达式, L_g, μ_g 表示图的拉普拉斯算子, α 表示权重改变系数,后续进行系数替换即可。

[0052] 利用获取患者行为的M个模态,样本 X_i 在其模态下分别表示为,

$$[0053] \quad X_i^1, X_i^2, X_i^3, \dots, X_i^M$$

[0054] 通过特征空间维数选取最优D种距离度量集合空间,从M个模态与D种距离度量可生成 $M \times D$ 张图。其中 $W_{(m-1) \times D+k}$ 是从第m个模态和第k中距离度量生成的图。

[0055] 2)通过预处理服务器提取预处理后的患者行为多模态数据之后,通过标注存储器进行审核,实现纠检传输差错,降低误码率,提高数据传输质量。

[0056] 图4为本发明提供的审查模块流程图,其具体步骤如下:

[0057] 2.1将产生的模态信息和距离度量通过AMPQ消息件传入下层接口,与半监督学习的数据反馈实现数据交换;(对应数据交换)

[0058] 2.2在这一过程中,不断的通过前向纠错和混合纠错的方式,校验传入的数据编码,出现错误编码的时候将信息正反馈至数据交换层,实现数据核查,保证数据的正确性和一致性。(对应传输纠错)

[0059] 2.3传输纠错后通过控制权重系数实现对算法性能的提升改变。当G很小时,采用交互验证方法设置大小;当G很大时,将其作为系数直接集成至正则化框架中,同时优化f和 α 。(对应控制权重系数)

[0060] 3)图5为本发明提供的管理模块流程图,具体步骤如下:

[0061] 3.1标注管理阶段考虑数据的时间不间断性。由于患者行为多模态数据中存在时间维度的关联,为此考虑构建Q张图表征数据样本之间的连续性关系。(对应时间连续性)

[0062] 3.2当选取Q为某合适值时,第一张图考虑每两个相邻样本之间的关系,即认为每个样本和它相邻的样本有较高概率有相同的概念。这时返回原始样本重新进行判断;当判断与下N个相邻样本之间的概念相同的概率较低时,进入下一步相似度改变步骤;(对应中间判断结构)

[0063] 3.3另一张图考虑每个样本和相邻的6个样本的关系,权重根据6个样本的位置而

定,其与相邻样本之间的关系分别表示为维度上的变化范围为1时,权重相似度取值为1;当变化范围为2时,权重相似度取值为1/2;当变化范围为3时,权重相似度取值为1/4;当变化范围为其他时,权重相似度取值为0。实际还可以设计更多的图来表示多模态数据之间的时空关联性,以此改变权重相似度的取值,鉴于MG-SSL的开放性,很容易供后续本领域技术人员集成进去。(对应改变权重相似度)

[0064] 4) 与关系型数据库并行工作的是标注AI协作单元,其利用轻量级的Docker虚拟技术加速本地开发,通过Docker能够快速搭建好开发和运行环境,并且该环境可以直接传递给测试和产品部署,实现自动化测试和持续集成部署。通过获取关系型数据库里的患者行为多模态数据,利用网络结构基于深度学习嵌入标注服务器,深度学习使用数据对其构筑中的参数进行更新以达成训练目标,实现在自动标注数据的同时进行在线学习。

[0065] 图6为本发明提供的AI协作模块流程图,具体步骤如下:

[0066] 启用Docker加速算法矩阵计算速度,使用相应环境的Docker镜像实现运行服务,同时利用生成的ANN阶层进行构图检验,为适应不同模态数据的更新要求,进行传入控制参数更新,以此辅助数据的提取和筛选。通过比较各图中被赋予的权重与使用这些图进行半监督学习的性能之间的关系,将权重与性能都实现规一化处理。MG-SSL能够考虑各图之间的互补性,赋予权重后能使得各图较好的协作,而不是简单的由单模态各图性能决定权重大小。之后多张图集成到一个正则化表达式框架中,同时优化这些图的加权系数,完成多模态数据的自动标注。

[0067] 本实例构建的患者行为多模态数据标注系统体系,有效的规范了数据的采集、录入和标注操作,实现了自动标注,为医疗操作的合规性评估和患者病情诊断感知分析打下了坚实的基础。

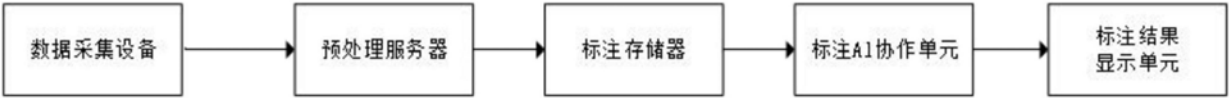


图1

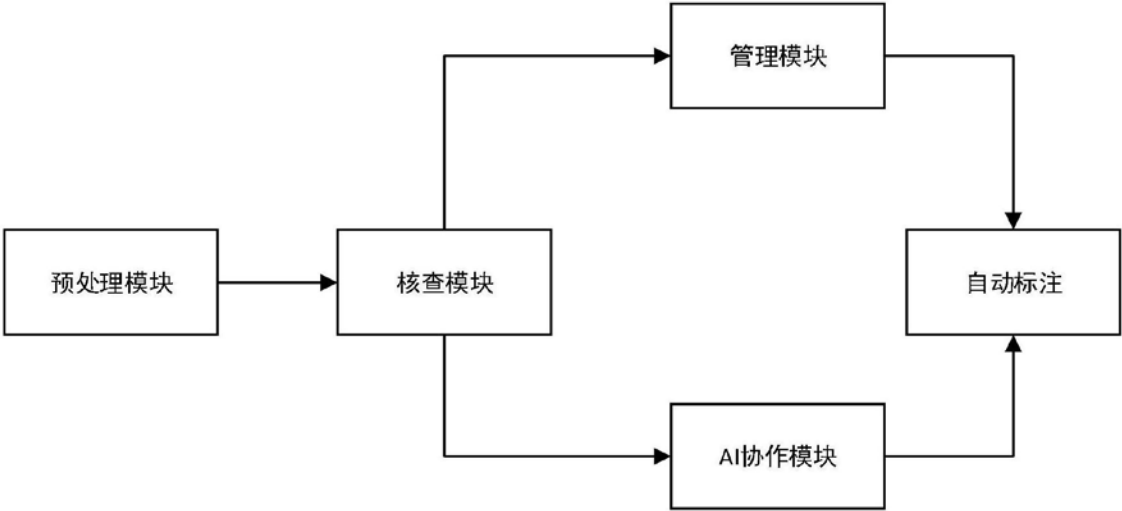


图2



图3



图4

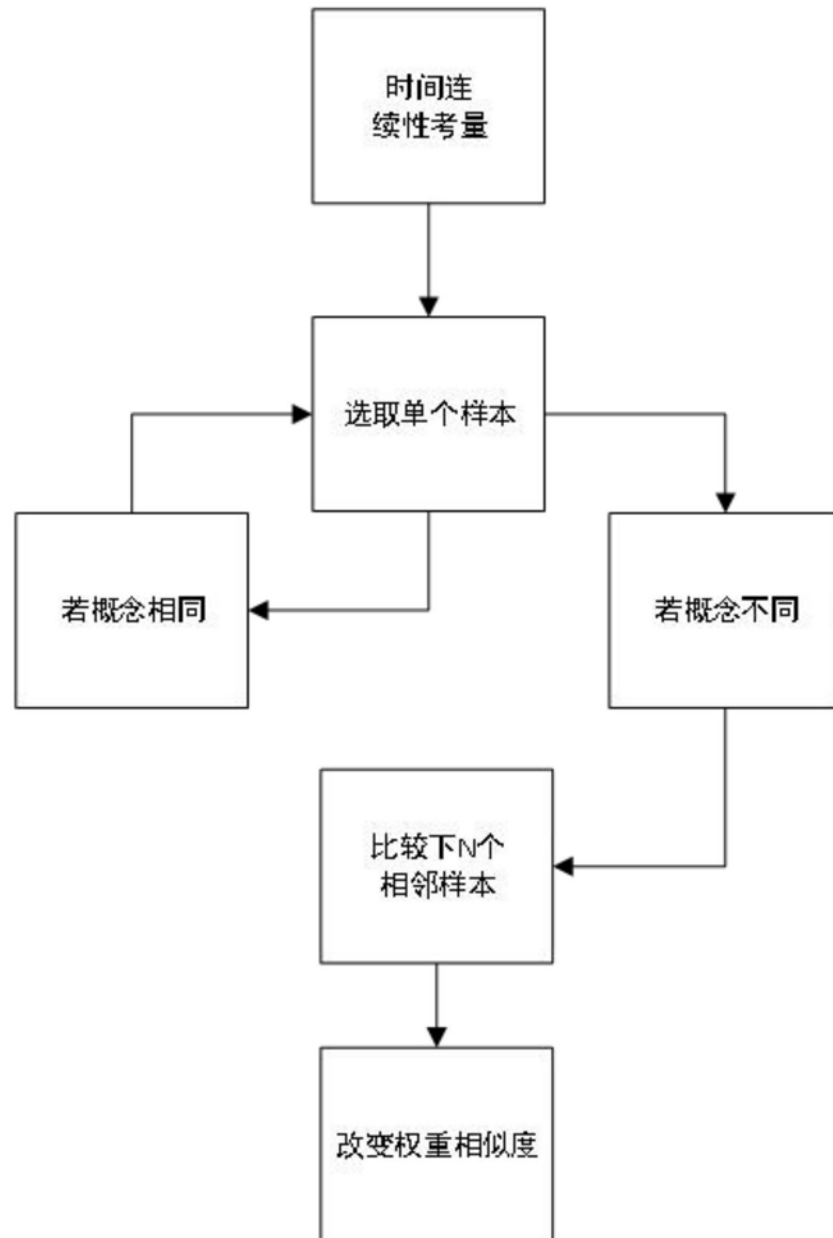


图5

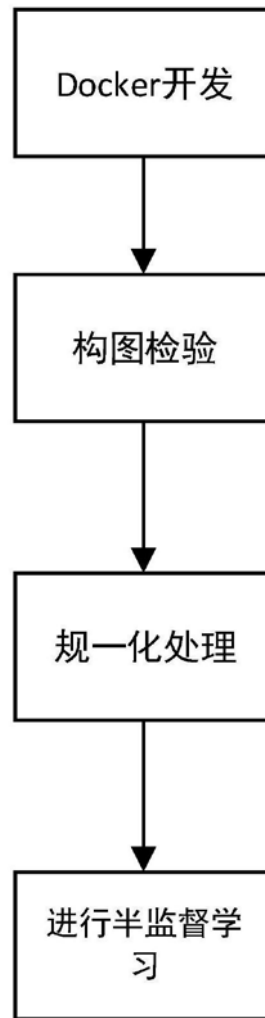


图6