(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117116475 A
(43)申请公布日 2023.11.24

(21)申请号 202310769689.3

(22)申请日 2023.06.27

(71)申请人 开滦总医院
地址 063000 河北省唐山市路北区新华东道57号
申请人 苏州森斯微电子技术有限公司

(72)发明人 元小冬 张萍淑 宋军

(74)专利代理机构 北京维正专利代理有限公司 11508
专利代理师 俞振明

(51)Int.Cl.
G16H 50/30(2018.01)
G16H 50/20(2018.01)
G06F 18/24(2023.01)
G06N 20/20(2019.01)
G06F 18/10(2023.01)
G06F 18/213(2023.01)
G06F 18/214(2023.01)
A61B 5/00(2006.01)
A61B 5/026(2006.01)

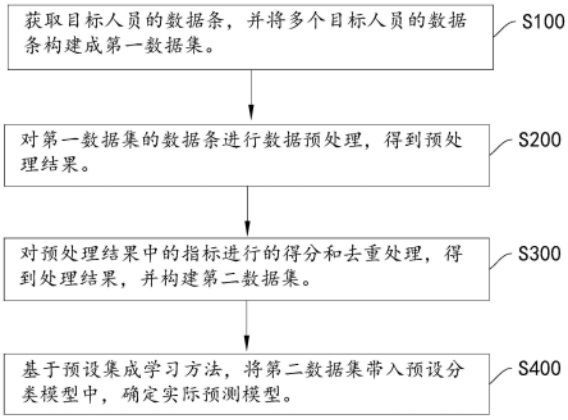
权利要求书2页 说明书11页 附图3页

(54)发明名称

缺血性脑卒中的风险预测方法、系统、终端及存储介质

(57)摘要

本申请涉及一种缺血性脑卒中的风险预测方法、系统、终端及存储介质，其方法包括获取目标人员的数据条，并将多个目标人员的数据条构建成第一数据集；对所述第一数据集的数据条进行数据预处理，得到预处理结果；对所述预处理结果中的指标进行的得分和去重处理，得到处理结果，并构建第二数据集；基于预设集成学习方法，将所述第二数据集带入预设分类模型中，确定实际预测模型。本申请具有对缺血性脑卒中的患者快速的风险预测的效果。



CN 117116475 A

1.一种缺血性脑卒中的风险预测方法,其特征在于:包括;

获取目标人员的数据条,所述数据条包括目标人员的基本信息、睡眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据,并将多个目标人员的数据条构建成第一数据集;

对所述第一数据集的数据条进行数据预处理,得到预处理结果;

对所述预处理结果中的指标进行的得分和去重处理,得到处理结果,并构建第二数据集;

基于预设集成学习方法,将所述第二数据集带入预设分类模型中,确定实际预测模型。

2.根据权利要求1所述的一种缺血性脑卒中的风险预测方法,其特征在于:所述预处理结果包括;

对所述睡眠期生理指标原始数据进行数据清洗、数据平衡和数据降维,得到睡眠期生理指标处理数据;

对所述血液流变学指标原始数据进行数据清洗、数据平衡和数据降维,得到血液流变学指标处理数据;

对所述脑血流动力学指标原始数据进行数据清洗、数据平衡和数据降维,得到脑血流动力学指标处理数据。

3.根据权利要求1所述的一种缺血性脑卒中的风险预测方法,其特征在于:所述对所述预处理结果中的指标进行得分和去重处理,得到处理结果,并构建第二数据集,包括;

基于预设得分规则,分别计算所述第一数据集中的基本信息、睡眠期生理指标、血液流变学指标和脑血流动力学指标之间的分值;

将所述分值按照从小到大的顺序进行排列,得到排列结果;

基于预设计算规则,根据所述排序结果分别计算基本信息、睡眠期生理指标、血液流变学指标和脑血流动力学指标相互之间的关系值;

比较所述关系值与预设关系阈值的大小,得到比较结果;

根据所述比较结果分别对睡眠期生理指标、血液流变学指标和脑血流动力学指标进行去重,得到去重结果,并构建第二数据集。

4.根据权利要求1所述的一种缺血性脑卒中的风险预测方法,其特征在于:所述预设分类模型包括随机森林预测模型、极端随机树预测模型、梯度提升预测模型、决策树预测模型、Bagging集成预测模型、投票集成预测模型、AdaBoost预测模型、SVM预测模型。

5.根据权利要求1所述的一种缺血性脑卒中的风险预测方法,其特征在于:所述基于预设集成学习方法和预设评估指标,将所述第二数据集带入预设分类模型中,确定实际预测模型,包括;

将所述第二数据集内的数据条分为训练集和测试集,训练集包括第一训练集和第二训练集;

将所述第一训练集带入预设分类模型中,对预设分类模型进行训练,得到训练结果;

根据所述训练结果、第二训练集和测试集对预设分类模型进行测试,得到测试结果;

基于预设集成学习方法,根据所述测试结果分别计算预设分类模型中的预设评估指标的评估分数;

根据所述评估分数确定实际预测模型。

6.根据权利要求1所述的一种缺血性脑卒中的风险预测方法,其特征在于:所述预设评估指标包括Accuracy值、Precision值、Recall值、F1-score值和AUC值。

7.根据权利要求5所述的一种缺血性脑卒中的风险预测方法,其特征在于:所述根据所述评估分数确定实际预测模型,包括;

将所述评估分数进行排序,得到排序结果;

根据所述排序结果确定实际预设模型。

8.一种缺血性脑卒中的风险预测系统,其特征在于:包括;

获取模块(21),用于获取目标人员的数据条,所述数据条包括目标人员的基本信息、睡眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据,并将多个目标人员的数据条构建成第一数据集;

预处理模块(22),用于对所述第一数据集的数据条进行数据预处理,得到预处理结果;

构建模块(23),用于对所述预处理结果中的指标进行的得分和去重处理,得到处理结果,并构建第二数据集;

确定模块(24),用于基于预设集成学习方法和预设评估指标,将所述第二数据集带入预设分类模型中,确定实际预测模型。

9.一种终端,其特征在于:包括存储器和处理器,所述存储器上存储有计算机程序,所述处理器执行所述程序时实现如其权利要求1-7中任一项所述的方法。

10.一种计算机可读存储介质,其特征在于:其上存储有计算机程序,所述程序被处理器执行时实现如权利要求1-7中任一项所述的方法。

缺血性脑卒中的风险预测方法、系统、终端及存储介质

技术领域

[0001] 本申请涉及医学技术的领域，尤其是涉及一种缺血性脑卒中的风险预测方法、系统、终端及存储介质。

背景技术

[0002] 脑卒中是一种由脑部血管突然破裂或因血管阻塞导致血液不能流入大脑而引起脑组织损伤的一种急性脑血管疾病。由于脑卒中的病情发展具有不可逆性，一旦发病，药物治疗和康复治疗只能起到减轻症状，预防并发症的效果，尽早发现脑卒中患者及其高危人群，这对于患者和社会来说，具有重要的现实意义。

[0003] 目前国内和国外尚没有在不改变受检者原有睡眠环境习惯、非直接接触式，并适用于家庭和医院等多种生活环境条件的预测方法。因此，对缺血性脑卒中的患者快速的风险预测成为一个亟待解决的问题。

发明内容

[0004] 为了对缺血性脑卒中的患者快速的风险预测，本申请提供一种缺血性脑卒中的风险预测方法、系统、终端及存储介质。

[0005] 本申请目的一是提供一种缺血性脑卒中的风险预测方法。

[0006] 本申请的上述申请目的一是通过以下技术方案得以实现的：

一种缺血性脑卒中的风险预测方法，包括；

获取目标人员的数据条，所述数据条包括目标人员的基本信息、睡眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据，并将多个目标人员的数据条构建成第一数据集；

对所述第一数据集的数据条进行数据预处理，得到预处理结果；

对所述预处理结果中的指标进行的得分和去重处理，得到处理结果，并构建第二数据集；基于预设集成学习方法和预设评估指标，将所述第二数据集带入预设分类模型中，确定实际预测模型。

[0007] 通过采用上述技术方案，通过将目标人员的数据条构建第一数据集之后，对第一数据集进行预处理和处理，并构建第二数据集，基于预设集成学习方法和预设评估指标，将第二数据集带入预设分类模型中，确定实际预测模型，实际预测模型能够对缺血性脑卒中的患者快速的风险预测。

[0008] 本申请在一较佳示例中可以进一步配置为：所述预处理结果包括；

对所述睡眠期生理指标原始数据进行数据清洗、数据平衡和数据降维，得到睡眠期生理指标处理数据；

对所述血液流变学指标原始数据进行数据清洗、数据平衡和数据降维，得到血液流变学指标处理数据；

对所述脑血流动力学指标原始数据进行数据清洗、数据平衡和数据降维，得到脑

血流动力学指标处理数据。

[0009] 本申请在一较佳示例中可以进一步配置为:所述对所述预处理结果中的指标进行得分和去重处理,得到处理结果,并构建第二数据集,包括;

基于预设得分规则,分别计算所述第一数据集中的基本信息、睡眠期生理指标、血液流变学指标和脑血流动力学指标之间的分值;

将所述分值按照从小到大的顺序进行排列,得到排列结果;

基于预设计算规则,根据所述排序结果分别计算基本信息、睡眠期生理指标、血液流变学指标和脑血流动力学指标相互之间的关系值;

比较所述关系值与预设关系阈值的大小,得到比较结果;

根据所述比较结果分别对睡眠期生理指标、血液流变学指标和脑血流动力学指标进行去重,得到去重结果,并构建第二数据集。

[0010] 本申请在一较佳示例中可以进一步配置为:所述预设分类模型包括随机森林预测模型、极端随机树预测模型、梯度提升预测模型、决策树预测模型、Bagging集成预测模型、投票集成预测模型、AdaBoost预测模型、SVM预测模型。

[0011] 本申请在一较佳示例中可以进一步配置为:所述基于预设集成学习方法和预设评估指标,将所述第二数据集带入预设分类模型中,确定实际预测模型,包括;

将所述第二数据集内的数据条分为训练集和测试集,训练集包括第一训练集和第二训练集;将所述第一训练集带入预设分类模型中,对预设分类模型进行训练,得到训练结果;

根据所述训练结果、第二训练集和测试集对预设分类模型进行测试,得到测试结果;

基于预设集成学习方法,根据所述测试结果分别计算预设分类模型中的预设评估指标的评估分数;

根据所述评估分数确定实际预测模型。

[0012] 本申请在一较佳示例中可以进一步配置为:所述预设评估指标包括Accuracy值、Precision值、Recall值、F1-score值和AUC值。

[0013] 本申请在一较佳示例中可以进一步配置为:所述根据所述评估分数确定实际预测模型,包括;

将所述评估分数进行排序,得到排序结果;

根据所述排序结果确定实际预设模型。

[0014] 本申请目的二是提供一种缺血性脑卒中的风险预测系统。

[0015] 本申请的上述申请目的二是通过以下技术方案得以实现的:

一种缺血性脑卒中的风险预测系统,包括;

获取模块,用于获取目标人员的数据条,所述数据条包括目标人员的基本信息、睡眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据,并将多个目标人员的数据条构建成第一数据集;

预处理模块,用于对所述第一数据集的数据条进行数据预处理,得到预处理结果;

构建模块,用于对所述预处理结果中的指标进行的得分和去重处理,得到处理结果,并构建第二数据集;

确定模块,用于基于预设集成学习方法和预设评估指标,将所述第二数据集带入预设分类模型中,确定实际预测模型。

[0016] 本申请目的三是提供一种终端。

[0017] 本申请的上述申请目的三是通过以下技术方案得以实现的:

一种终端,包括存储器和处理器,所述存储器上存储有计算机程序,所述处理器执行所述程序时实现上述任一种缺血性脑卒中的风险预测方法。

[0018] 本申请目的四是提供一种可读存储介质,能够存储相应的程序。

[0019] 本申请的上述申请目的四是通过以下技术方案得以实现的:

一种计算机可读存储介质,其上存储有计算机程序,所述程序被处理器执行时实现上述任一种缺血性脑卒中的风险预测方法。

[0020] 综上所述,本申请包括以下有益技术效果:

通过将目标人员的数据条构建第一数据集之后,对第一数据集进行预处理和处理,并构建第二数据集,基于预设集成学习方法和预设评估指标,将第二数据集带入预设分类模型中,确定实际预测模型,实际预测模型能够对缺血性脑卒中的患者快速的风险预测。

附图说明

[0021] 图1是本申请实施例一种缺血性脑卒中的风险预测方法的流程示意图。

[0022] 图2是本申请实施例一种缺血性脑卒中的风险预测系统的系统示意图。

[0023] 图3是本申请实施例的终端的结构示意图。

[0024] 附图标记说明:21、获取模块;22、预处理模块;23、构建模块;24、确定模块;301、CPU;302、ROM;303、RAM;304、总线;305、I/O接口;306、输入部分;307、输出部分;308、存储部分;309、通信部分;310、驱动器;311、可拆卸介质。

具体实施方式

[0025] 以下结合附图对本申请作进一步详细说明。

[0026] 本具体实施例仅仅是对本申请的解释,其并不是对本申请的限制,本领域技术人员在阅读完本说明书后可以根据需要对本实施例做出没有创造性贡献的修改,但只是在本申请的权利要求范围内都受到专利法的保护。

[0027] 为使本申请实施例的目的、技术方案和优点更加清楚,下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本申请的保护范围。

[0028] 众所周知,缺血性脑卒中作为我国发病率高、致残率高、致死率高的重大疾病之一,目前已成为了整个社会关注的焦点健康问题之一,而提前预防和尽早干预的健康管理手段是改变其预后的主要途径。因此,通过高新技术方法对人体个体的生理状态监测,早期对脑卒中疾病的发生进行预警和预防已成为当前急需解决的医药卫生领域的重大科学问题之一。

[0029] 近年来,深度学习由于其强大的表示学习能力,已经在通信、工业、计算机网络安全以及医学等领域取得了成功,这也为其在辅助诊断脑卒中疾病上提供了先验理论基础。

目前,不少研究者进行基于深度学习的脑卒中预测研究。然而,现有的研究工作存在以下问题:首先,所采用的深度学习网络大多都忽视了数据分布对神经网络分类性能的影响,对于输入模型的数据“一视同仁”地进行残差拟合,而后判决输出;其次,现有的大多数研究仅对正常人样本和脑卒中患者进行分类,这样的操作忽略了脑卒中疾病的病变过程;此外,现有的脑卒中评价准则对脑卒中高危人群的关注不够,导致脑卒中高危人群对脑卒中疾病不够重视,进而易导致病情恶化。而机体的血液流变学和脑血流动力学作为血栓发病前和形成过程中最重要的病理生理学基础和直接参与性机制,虽然目前已经广泛应用于临床做为此类疾病辅助诊断的重要工具,但因其需要昂贵的专业设备和专业性很强的技术人员,而限制了将其应用于医院以外的大众日常生活环境进行缺血性脑卒中的预测和报警。同时,深度学习方法由于其对于特征提取的“不可解释性”、“大数据”特征,因而并完全不适合此类研究,而实践证明与临床紧密研判相结合的多维度“小数据”了的智能方法更适用于此类研究。

[0030] 尽管针对疾病预测预防的研究很多,针对心脑血管疾病和脑卒中的研究也不少,但目前在脑卒中的危险等级预测仍然存在以下三方面的问题:

(1)模型预测能力不足。深度学习网络的预测需要精准的大数据提供支撑。现有的脑卒中数据,具有采集标准不统一、样本数据特征缺失、忽略地区差异进行数据统一记录等特点,导致相同类型的脑卒中样本数量不足,模型的准确度较低。高的预测准确率,能够为医生提供更为准确的辅助信息,为医生诊断病人病情提供更有效的辅助信息。

[0031] (2)忽略脑卒中疾病的病变过程导致模型对脑卒中疾病的过程监控程度不够。现有的深度学习脑卒中模型仅仅专注于脑卒中疾病于正常人之间的区分,忽略了脑卒中疾病的发病过程其实是一个渐变的过程,并非一蹴而就。在卒中前就已经表现出一些早期征兆。因此,对脑卒中疾病进行过程监控,能够有效防控脑卒中疾病。因此单一类型数据(脑卒中样本与正常样本)建立的模型很难实现对脑卒中疾病的过程监督

(3)对脑卒中疾病特征进行无差别式的训练建模。脑卒中数据具有高混叠(同类样本具有较大的类内方差且不同类样本之间距离较近)的分布情况。一般的深度学习模型,在建立过程中,都是对所有的特征进行残差拟合,很不适合处理具有高混叠的脑卒中数据。这就导致了深度神经网络对于现有的脑卒中样本的分类效果不理想。

[0032] 因此,本申请提供一种缺血性脑卒中的风险预测方法、系统、终端及存储介质,其能够对缺血性脑卒中的患者快速的风险预测。

[0033] 步骤S100:获取目标人员的数据条,并将多个目标人员的数据条构建成第一数据集。

[0034] 具体的,目标人员的数据条包括目标人员的基本信息、睡眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据。其中,目标人员的基本信息通过入院登记获得,目标人员的睡眠期生理指标原始数据通过无接触式睡眠监测床垫采集,目标人员的血液流变学指标原始数据和脑血流动力学指标原始数据均通过实验室检测采集。将目标人员的数据条进行集合,构建第一数据集。

[0035] 由于数据条的数量多,会存在以下问题:

(1)数据条体量较大,但因睡眠期生理指标原始数据采集设备的部署条件有限而限制了其进一步增多;

(2)数据条类别不均衡,因采集客观环境造成阳性数据条明显多于阴性数据条(阳性数据条为目标人员是高危人员的数据条,阴性数据条为目标人员是正常人员的数据条);

(3)指标具有多样性,第一数据集所含指标共30个,数据类型包括分类数据和定量数据。各个指标存在不同程度的缺失与异常,数据多样性造成数据的量纲存在较大差异,同时不可忽略的是每条数据条的维度较高,从中提取较为关键的、贡献度大的指标,进行降低维度十分重要。

[0036] 步骤S200:对所第一数据集的数据条进行数据预处理,得到预处理结果。

[0037] 具体的,在对目标人员的数据条内的数据进行采集之后,会出现数据的缺失、重复、互斥和极端变异,因此,需要对第一数据集内的数据条的数据进行数据预处理。

[0038] 数据预处理的作用是清洗清洗噪声数据、无关数据、处理遗漏数据、填补缺失特征值、识别删除孤立点等。合理处理这些异常数据能够为机器学习提供合理而有效的学习数据。

[0039] 其中,预处理结果包括:

对所述睡眠期生理指标原始数据进行数据清洗、数据平衡和数据降维,得到睡眠期生理指标处理数据;

对所述血液流变学指标原始数据进行数据清洗、数据平衡和数据降维,得到血液流变学指标处理数据;

对所述脑血流动力学指标原始数据进行数据清洗、数据平衡和数据降维,得到脑血流动力学指标处理数据。

[0040] 数据清洗主要体现在对数据条内数据缺失进行删除或补齐,对数据的删除适用于缺失的数据在第一数据集内所有数据的占比较少的情况,对数据的补齐适用于所有情况。其中,常用的填补方式有均值插补、同类均值插补和建模预测等,均值插补根据同一类指标的均值插补到空缺的特征中。

[0041] 如,睡眠期生理指标包括睡眠期平均呼吸次数,A目标人员的睡眠期平均呼吸次数为70次/分,B目标人员的睡眠期平均呼吸次数为60次/分,C目标人员的睡眠期平均呼吸次数为65次/分,D目标人员的睡眠期平均呼吸次数为缺失数值,则对A目标人员睡眠期平均呼吸次数、B目标人员的睡眠期平均呼吸次数和C目标人员的睡眠期平均呼吸次数进行平均值计算,平均值为65,因此,D目标人员的睡眠期平均呼吸次数为65次/分。

[0042] 数据清洗还体现在对数据条内数据的重复值进行删除,仅保留一个数据条。

[0043] 如,睡眠期生理指标包括睡眠期平均呼吸次数、睡眠期最慢呼吸次数和睡眠期最快呼吸次数,A目标人员的睡眠期平均呼吸次数为70次/分、睡眠期最慢呼吸次数为65次/分和睡眠期最快呼吸次数为75次/分,B目标人员的睡眠期平均呼吸次数为70次/分、睡眠期最慢呼吸次数为65次/分和睡眠期最快呼吸次数为75次/分,则将B目标人员的数据条进行删除。

[0044] 数据平衡主要体现在数据条内的数据的互斥,对于互斥的数据,如果能够找出其中的正确数据则对其进行保留,尽可能多的保留数据条,为建模提供更多的数据条。而对于无法进行判断孰对孰错的数据条,在多数情况下选择将二者都去掉的方式以保证数据条的正确性与可靠性。

[0045] 如,睡眠期生理指标包括睡眠期平均呼吸次数,A目标人员的睡眠期平均呼吸次数

为40次/分,B目标人员的睡眠期平均呼吸次数为90次/分,无法确定哪个数据相对正确,因此,将A目标人员的数据条和B目标人员的数据条全部删除。

[0046] 数据降维主要体现在数据条内数的数据的极端变异,对于一些具有极端变异性的数据和产生于观测、记录、计算错误(或试验方法和条件的偶然偏离)的数据,将对应的数据条删除。离群点检测是异常检测中最常用的方法之一。离群点检测在各行业里面引用都非常广泛。

[0047] 如,如,睡眠期生理指标包括睡眠期平均呼吸次数,A目标人员的睡眠期平均呼吸次数为5分/次,在实际情况中,人的睡眠期平均呼吸次数不可能为5分/次,因此,将A目标人员的数据条进行删除。

[0048] 同时,还会分别对目标人员的基本信息、睡眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据进行数据变换,数据变换是将数据转换为适合分析与实验的形式,提高数据处理速度,便于数据的后续的使用。

[0049] 如,将基本信息中的性别、吸烟史、饮酒史、高血压史、糖尿病史的“无”和“有”转化为“0”和“1”。

[0050] 基本信息分别是性别、年龄、身高、体重、吸烟史、饮酒史、高血压史和糖尿病史。

[0051] 睡眠期生理指标分别是睡眠总时长、入睡后清醒总时长、睡眠潜伏期、睡眠效率、睡眠期平均呼吸次数、睡眠期最慢呼吸次数、睡眠期最快呼吸次数、睡眠期平均心率、睡眠期醉眠心率、睡眠期最快心率、R期潜伏期、R期睡眠时长、R期睡眠占比、浅睡眠时长、浅睡眠占比、深睡眠时长、深睡眠占比、总呼吸暂停次数、呼吸变异指数、心率变异指数和心脑血管健康指数。

[0052] 血液流变学指标包括低切1秒全血粘度、低切5秒全血粘度、低切30秒全血粘度、全血高切200秒全血粘度、血浆粘度和红细胞压积。

[0053] 脑血流动力学指标包括左大脑中动脉收缩期峰值血流速度、左大脑中动脉搏动指数、左大脑后动脉收缩期峰值血流速度、左大脑后动脉搏动指数、左大脑中动脉收缩期峰值血流速度、左大脑中动脉狭窄程度、左大脑后动脉收缩期峰值血流速度、右大脑中动脉收缩期峰值血流速度、右大脑中动脉搏动指数、右大脑后动脉收缩期峰值血流速度、右大脑后动脉搏动指数、右大脑中动脉收缩期峰值血流速度、右大脑中动脉狭窄程度和右大脑后动脉收缩期峰值血流速度。

[0054] 步骤S300:对预处理结果中的指标进行得分和去重处理,得到处理结果,并构建第二数据集。

[0055] 具体的,基于预设得分规则,分别计算第一数据集中的基本信息、睡眠期生理指标、血液流变学指标和脑血流动力学指标之间的分值;将分值按照从小到大的顺序进行排列,得到排列结果。

[0056] 利用sklearn的内置函数SelectKBest函数计算数据集中各指标的得分,该函数内部为分类特征的评价提供了集中方法,包括卡方检验、样本方差F值、离散类别交互信息。随后将分值按照从小到大的顺序进行排列。

[0057] 然后,基于预设计算规则,计算第一数据集中的基本信息、睡眠期生理指标、血液流变学指标和脑血流动力学指标之间的关系值;比较关系值与预设关系阈值的大小,得到比较结果;根据比较结果分别对睡眠期生理指标、血液流变学指标和脑血流动力学指标进

行去重,得到去重结果。

[0058] 即利用Spearman方法计算第一数据集中所有指标间相关关系,从而将相关性大于预设关系阈值的指标进行去重,进一步达到指标降维的目的,随后构建第二数据集。

[0059] 步骤S400:基于预设集成学习方法和预设评估指标,将第二数据集带入预设分类模型中,确定实际预测模型。

[0060] 具体的,将第二数据集内的数据条分为训练集和测试集,训练集包括第一训练集和第二训练集;将第一训练集带入预设分类模型中,对预设分类模型进行训练,得到训练结果;根据训练结果、第二训练集和测试集对预设分类模型进行测试,得到测试结果;基于预设集成学习方法,根据所述测试结果分别计算预设分类模型中的预设评估指标的评估分数;根据评估分数确定实际预测模型。

[0061] 预设分类模型包括随机森林预测模型、极端随机树预测模型、梯度提升预测模型、决策树预测模型、Bagging集成预测模型、投票集成预测模型、AdaBoost预测模型、SVM预测模型。

[0062] 将第二数据集内的数据条带入上述八个预设分类模型中,通过预设集成学习方法确定一个实际预测模型。

[0063] 预设集成学习方法是解决单一分类器稳定性的重要手段。预设集成学习方法能够将多个分类器进行组合,使用投票等方法将它们的预测结果融合到一起,并得到最后的预测结果。集成中的分类器称为基分类器,根据基分类器的生成方式可以将集成分为两类:异构的(heterogeneous)和同构的(homogeneous)。

[0064] 集成中基分类器成员的数量并不是越多越好,增加基分类器有时徒增计算量却不能提高集成的准确性,甚至会降低集成的性能。集成的准确性和泛化能力在很大程度上依赖于成员分类器的性能及其多样性。研究表明,组成集成的基分类器的间隔、准确性、多样性等是获得一个较好集成的重要因素。

[0065] 可以知道的是,在预设分类模型学习过程中,需要人工选择的参数称为超参数,比如随机森林中的决策树的个数,正则项中常数大小等等,都需要人工确定,超参数选择不恰当,会出现过拟合或过拟合的问题。选择超参数的方式有两个途径,一个是凭经验微调,另一个是选择不同大小的参数,带入模型中,挑选表现最好的参数。

[0066] 可以知道的是,在机器学习建模过程中,通常是将数据条分为训练集和测试集。测试集是与训练独立的数据,完全不参与训练,用于最终模型的评估,即实际预测模型的评估。在训练过程中,经常会出现过拟合的问题,就是模型可以很好的匹配训练数据,却不能很好在预测训练集外的数据。如果此时就使用测试数据来调整模型参数,就相当于在训练时已知部分测试数据的信息,会影响最终评估结果的准确性。通常的做法是在训练数据中分出一部分做为验证数据,用来评估模型的训练效果,即将训练集分为第一训练集和第二训练集。验证数据取自训练数据,但不参与训练,这样可以相对客观的评估模型对于训练集之外数据的匹配程度。

[0067] 可以知道的是,机器学习分类模型常用评价指标有Accuracy值、Precision值、Recall值、F1-score值和AUC值等,并且二分类模型与多分类模型的评估有所差异,每一种指标的关注点都有所不同,所以综合实际预测模型的各项性能是较为合适的。

[0068] 混淆矩阵是分类算法的一种标准表示方式,它给出了数据条的实际类别以及分类

模型的预测类别之间的映射关系,是评估分类模型性能的常用方法。混淆矩阵隐含了丰富的类信息,包括类内部的聚合度和类之间的离散度,同时也为类别的划分提供了重要的参考。由点(0,0)、(0,1)、(1,1)、(1,0)围成的1*1矩形区域称为ROC空间。ROC(Receiver OperatingCharacteristic)曲线则是由点(FP rate,TP rate)链接而成的一条曲线,每个(FP rate,TP rate)点对应一个分类模型,并能从中评估分类模型的性能。ROC曲线又称接受者操作特征曲线,该曲线最早应用于雷达信号检测领域,用于区分信号与噪声。后来人们将其用于评价模型的预测能力,ROC曲线是基于混淆矩阵得出的,改变阈值只是不断地改变预测的正负样本数,即TPR和FPR,但是曲线本身是不会变的。其中,FPR表示模型虚报的响应程度,而TPR表示模型预测响应的覆盖程度。模型的虚报越少越好,覆盖越多越好,因此TPR越高,同时FPR越低,即ROC曲线越陡,那么模型的性能就越好。

[0069] 如,数据条有100个,分类结束后得到的混淆矩阵为:

实际类别\预测类别	正例	负例
正例	30(True Positive,TP)	5(False Negative,FN)
负例	10(False Positive,FP)	55(True Negative,TN)

表一

参照表一可以知道的是,有30个属于正例的数据条被正确预测为了正例,有5个正例的数据条被错误预测为了负例,有10个属于负例的数据条被错误预测为了正例,有55个负例的数据条被错误预测为了负例。可以理解为100个数据条中有35个阳性数据条和65个阴性数据条,其中,有30个阳性数据条被预测正确,有55个阴性数据条被预测正确。

[0070] 当ROC曲线间出现交叉时,通过简单的比较ROC曲线凸起程度就很难区分实际预测模型的性能,所以使用曲线下面积AUC(Area Under The Curve)来量化ROC曲线,并通过比较AUC值大小来评判实际预测模型性能。AUC的值越大,说明实际预测模型的性能越好。

[0071] $$AUC = \frac{1}{P*N}\sum_{i=1}^{P}\sum_{j=1}^{N} G\ (P_i^+ > N_j^-) \qquad (1.1)$$

$$G = \begin{cases} 1, & P_i^+ > N_j^- \\ 0, & P_i^+ < N_j^- \end{cases} \qquad (1.2)$$

$$SE = \sqrt{\frac{AUC(1-AUC)+(P-1)(Q_1-AUC^2)+(N-1)(Q_2-AUC^2)}{P*N}} \qquad (1.3)$$

公式1.1、公式1.2和1.3中可理解为P个正例和N个负例,若正例多于负例则为1,否则为0,计算加权平均值即可得到AUC值,AUC的误差公式如1.3所示,AUC仍继承ROC曲线的诸多优势,能评估实际预测模型性能、与类分布无关、独立于阈值等。其中,$Q_1 = \frac{AUC}{2} - AUC, Q_2 = \frac{2*AUC^2}{1+AUC}$。

[0072] $$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1.4)$$

$$Percision = \frac{TP}{TP+FP} \qquad (1.5)$$

$$Recall = \frac{TP}{TP+FN} \qquad (1.6)$$

$$F1-score = \frac{2*Percision*Recall}{Percision+Recall} \qquad (1.7)$$

公式(1.4)为Accuracy值的计算公式,公式(1.5)为Precision值的计算公式、公式(1.6)为Recall值的计算公式、公式(1.7)为F1-score值的计算公式。

[0073] 通过Accuracy值、Precision值、Recall值、F1-score值和AUC值能够确定影响脑缺性脑卒中的影响因素。

[0074] 如,选择临床确诊的TIA患者15例、急性脑梗死患者50例。我们首先筛选影响急性脑梗死与一过性脑缺血的血液流变学主要风险因素,从而为后继的建模测警提供依据。在临床脑卒中患者为了辅助诊断常规检测的血液流变学指标主要为低切1秒全血粘度(mPa.s)、低切5秒全血粘度(mPa.s)、低切30秒全血粘度(mPa.s)、高切200秒全血粘度(mPa.s)、血浆粘度、红细胞压积(MM/H)。在判断血液流变学指标是否在急性脑梗死发病过程中起作用,对其属性特征进行相关性分析。低切1秒全血粘度(mPa.s)、低切5秒全血粘度(mPa.s)、低切30秒全血粘度(mPa.s)之间有极大的相关性,可以选择保留其中的一个指标即可;我们根据此次相关分析结果和文献回顾以及以前的临床经验,确定保留或重点关注其中的“低切30秒全血粘度”用做后面的分析和建模之用。同样,血浆粘度与红细胞压积之间也存在着极大的相关性,我们只保留其中的“红细胞压积”在上述预实验的基础上,进行第二阶段实验,主要分析血液流变学指标与急性缺血性脑卒中的关系。本部分实验选择正常志愿者组15例,TIA患者组15例,急性脑梗死患者组25例。首先,对于血液流变学预测TIA进行分析,以正常组和TIA组为基础应用svm算法进行建模,血液流变学指标对于TIA这类发生急性脑梗死的高危险人群的预测作用很弱,预测结果中TIA患者的准确率仅50%,召回率为25%,而且对其预测模型的综合评价效果F1值为0.33,因此血液流变学指标对于预测和判断TIA发病的作用不明显。其次,对于血液流变学预测急性脑梗死进行分析,以正常组和急性脑梗死组为基础应用svm算法进行建模,血液流变学指标对于发生急性脑梗死的高危险人群的预测作用较强,结果中预测急性脑梗死患者的准确率虽然也为50%,但召回率达到100%,而且对其预测模型的综合评价效果F1值上升到0.67,因此,血液流变学指标对急性脑梗死发病具有明显的预测和判断价值。同时,在本阶段过程中我们进一步在TIA组(15例)和急性脑梗死组(50例)之间分析其血液流变学指标对于TIA这种高风险状态发展成为急性脑梗死的预测作用,其中分类为TIA组是发生急性脑梗死的高危状态人群和急性脑梗死组,以血液流变学指标为特征应用svm算法建模,其模型的评价血液流变学指标对于TIA这类发生急性脑梗死的高危险人群没有明显的判断作用;但对于急性脑梗死人群的预测结果可见脑梗死患者的召回率为100%,这表示本模型对于临床诊断为脑梗死患者的判断程度达到了100%的程度,而且对此模型预测效果综合评价的F1值很高,因此血液流变学指标对于急性脑梗死的发病具有重要的预测和判断作用,特别是其中的红细胞压积、高切200秒全血粘度、低切30秒全血粘度在个体的急性脑梗死发病过程中发挥着重要的作用和具有明显的预报价值。

[0075] 综上所述,通过将目标人员的数据条构建第一数据集之后,对第一数据集进行预处理和处理,并构建第二数据集,基于预设集成学习方法和预设评估指标,将第二数据集带入预设分类模型中,确定实际预测模型,实际预测模型能够对缺血性脑卒中的患者快速的风险预测。

[0076] 参照图2,一种缺血性脑卒中的风险预测系统,包括:

获取模块21,用于获取目标人员的数据条,所数据条包括目标人员的基本信息、睡

眠期生理指标原始数据、血液流变学指标原始数据和脑血流动力学指标原始数据,并将多个目标人员的数据条构建成第一数据集;

预处理模块22,用于对第一数据集的数据条进行数据预处理,得到预处理结果;

构建模块23,用于对预处理结果中的指标进行的得分和去重处理,得到处理结果,并构建第二数据集;

确定模块24,用于基于预设集成学习方法和预设评估指标,将第二数据集带入预设分类模型中,确定实际预测模型。

[0077] 参照图3,终端包括中央处理单元(CPU)301,其可以根据存储在只读存储器(ROM)302中的程序或者从存储部分加载到随机访问存储器(RAM)303中的程序而执行各种适当的动作和处理。在RAM 303中,还存储有系统操作所需的各种程序和数据。CPU 301、ROM 302以及RAM 303通过总线304彼此相连。输入/输出(I/O)接口305也连接至总线304。

[0078] 以下部件连接至I/O接口305:包括键盘、鼠标等的输入部分306;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分307;包括硬盘等的存储部分308;以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分309。通信部分309经由诸如因特网的网络执行通信处理。驱动器310也根据需要连接至I/O接口305。可拆卸介质311,诸如磁盘、光盘、磁光盘、半导体存储器等,根据需要安装在驱动器310上,以便于从其上读出的计算机程序根据需要被安装入存储部分308。

[0079] 特别地,根据本申请的实施例,上文参考流程图图1描述的过程可以被实现为计算机软件程序。例如,本申请的实施例包括一种计算机程序产品,其包括承载在机器可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分309从网络上被下载和安装,和/或从可拆卸介质311被安装。在该计算机程序被中央处理单元(CPU)301执行时,执行本申请的系统中限定的上述功能。

[0080] 需要说明的是,本申请所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一种或多种导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等,或者上述的任意合适的组合。

[0081] 附图中的流程图和框图,图示了按照本申请各种实施例的系统、方法和计算机程

序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，前述模块、程序段、或代码的一部分包含一种或多种用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0082] 描述于本申请实施例中所涉及到的单元或模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元或模块也可以设置在处理器中，例如，可以描述为：一种处理器接获取模块21、预处理模块22、构建模块23和确定模块24。其中，这些单元或模块的名称在某种情况下并不构成对该单元或模块本身的限定，例如，预处理模块22还可以被描述为“用于对第一数据集的数据条进行数据预处理，得到预处理结果的模块”。

[0083] 作为另一方面，本申请还提供了一种计算机可读存储介质，该计算机可读存储介质可以是上述实施例中描述的电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中的。上述计算机可读存储介质存储有一个或者多个程序，当上述前述程序被一个或者一个以上的处理器用来执行描述于本申请的数据加密传输方法。

[0084] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的申请范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离前述申请构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中申请的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

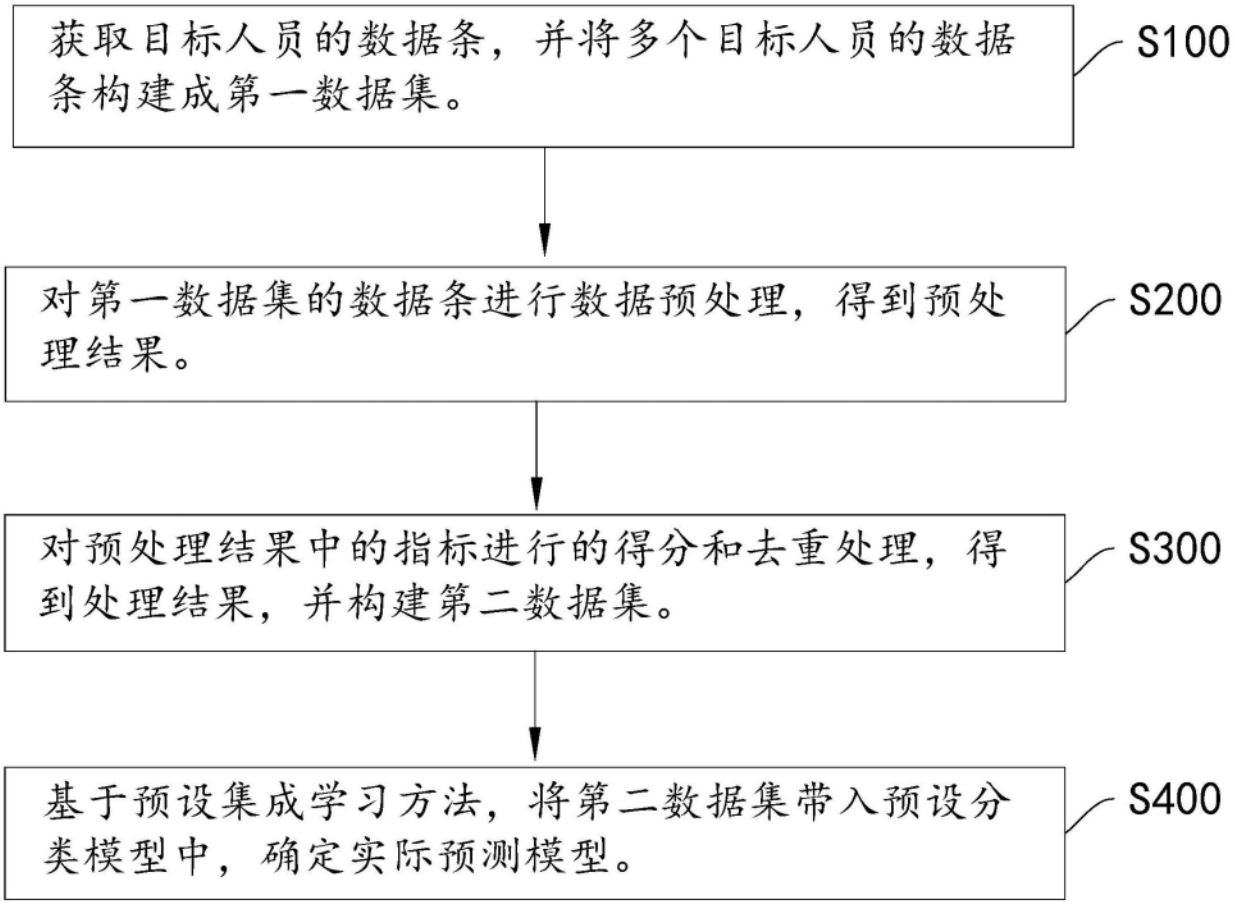

获取目标人员的数据条，并将多个目标人员的数据条构建成第一数据集。
S100
对第一数据集的数据条进行数据预处理，得到预处理结果。
S200
对预处理结果中的指标进行的得分和去重处理，得到处理结果，并构建第二数据集。
S300
基于预设集成学习方法，将第二数据集带入预设分类模型中，确定实际预测模型。
S400

图1

图2

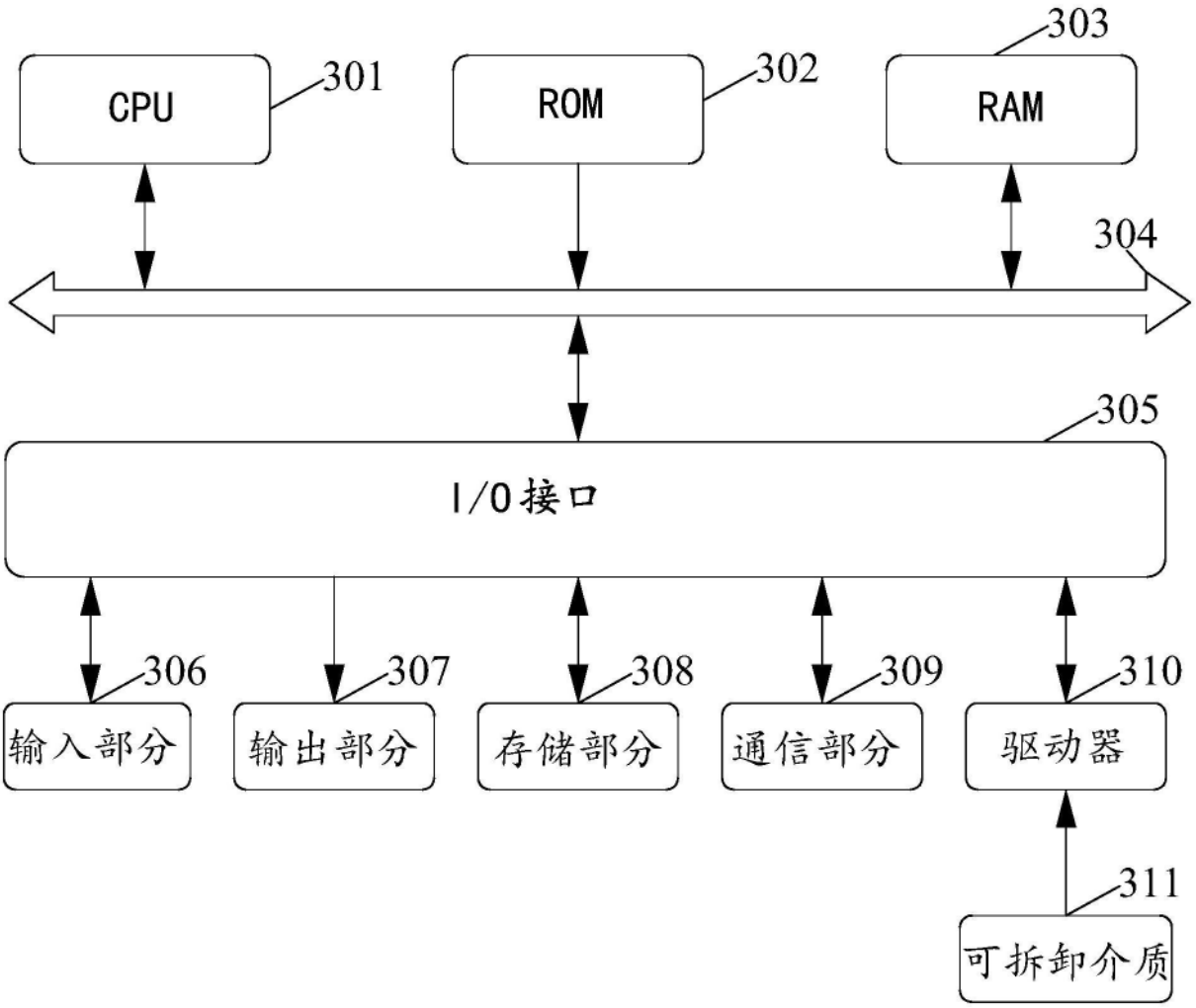

CPU
301
ROM
302
RAM
303
304
I/O接口
305
306
307
308
309
310
输入部分
输出部分
存储部分
通信部分
驱动器
311
可拆卸介质

图3