(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111933212 A
(43)申请公布日 2020.11.13

(21)申请号 202010873321.8

(22)申请日 2020.08.26

(71)申请人 腾讯科技(深圳)有限公司
地址 518064 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 邢小涵 杨帆 姚建华

(74)专利代理机构 深圳市智圈知识产权代理事务所(普通合伙) 44351
代理人 韩绍君

(51)Int.Cl.
G16B 20/00(2019.01)
G16H 50/50(2018.01)
G06N 20/00(2019.01)
G06N 3/04(2006.01)

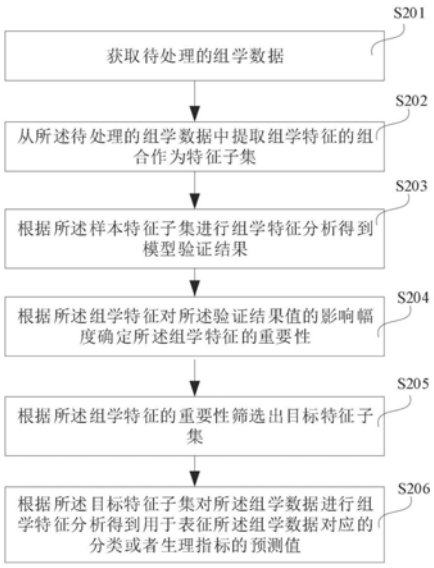
权利要求书2页 说明书14页 附图11页

(54)发明名称

一种基于机器学习的临床组学数据处理方法及装置

(57)摘要

本发明涉及一种基于机器学习的临床组学数据处理方法,包括:获取待处理的组学数据;从待处理的组学数据中提取组学特征的组合作为特征子集;根据所述特征子集进行组学特征分析得到模型验证结果;根据所述组学特征对所述模型验证结果值的影响幅度确定所述组学特征的重要程度值;根据所述组学特征的重要程度值筛选出目标特征子集;以及根据所述目标特征子集对所述组学数据进行组学特征分析得到用于表征所述组学数据对应的分类或者生理指标的预测值。上述方法能够提升组学数据处理的效率。此外,本申请实施例还提供一种临床组学数据处理、服务器及存储介质。



CN 111933212 A

1.一种基于机器学习的临床组学数据处理方法,其特征在于,包括:

获取待处理的组学数据;

从所述待处理的组学数据中提取组学特征的组合作为特征子集;

根据所述特征子集进行组学特征分析得到模型验证结果;

根据所述组学特征对所述模型验证结果的影响幅度确定所述组学特征的重要程度值;

根据所述组学特征的重要程度值筛选出目标特征子集;以及

根据所述目标特征子集对所述组学数据进行组学特征分析得到用于表征所述组学数据对应的分类或者生理指标的预测值。

2.如权利要求1所述的基于机器学习的临床组学数据处理方法,其特征在于,所述根据所述特征子集进行组学特征分析得到模型验证结果包括:

获取已标记的训练样本;

采用超梯度树提升分类器、逻辑回归法、向量机法、全连接网络法、长短期记忆网络法、多图层感知机法、及图卷积神经网络法的任意组合对所述训练数据进行训练得到验证模型;以及

将所述样本特征子集输入所述验证模型得到所述模型验证结果。

3.如权利要求2所述的基于机器学习的临床组学数据处理方法,其特征在于,所述根据所述组学特征对所述验证结果值的影响幅度确定所述组学特征的重要程度值时采用以下方法的任意组合:沙谱利附加解释法、基尼重要性、遗传算法、方差分析、T检验及曼-惠特尼秩和检验。

4.如权利要求2所述的基于机器学习的临床组学数据处理方法,其特征在于,所述验证模型采用超梯度树提升分类器法训练得到,所述方法包括:

将所述训练样本的特征输入梯度树提升分类器,目标函数设置为二分类的逻辑回归,评估指标设置为受试者工作特征曲线的面积,通过参数的自动搜索和交叉验证确定梯度树的参数取值。

5.如权利要求4所述的基于机器学习的临床组学数据处理方法,其特征在于,所述方法还包括:

获取所述组学特征在梯度树中出现的次数及所述验证模型对所述组学特征的评分;以及

根据所述次数及所述评分的加权值确定所述组学特征的重要程度值。

6.如权利要求4所述的基于机器学习的临床组学数据处理方法,其特征在于,所述根据所述组学特征对所述模型验证结果的影响幅度确定所述组学特征的重要程度值包括:

将不同特征子集得到的模型验证结果的受试者工作特征曲线的面积求平均,根据最高的受试者工作特征曲线的面积取值判断所述重要程度值。

7.如权利要求2-6任一项所述的基于机器学习的临床组学数据处理方法,其特征在于,所述方法还包括:

将所述训练样本随机分成多个样本子集;以及

分别采用所述多个样本子集训练所述验证模型。

8.如权利要求1所述的基于机器学习的临床组学数据处理方法,其特征在于,还包括:

根据所述预测值生成图形化解释界面,所述图形化解释界面包括:用指示样本特征重

要程度值可视化的柱状图、用于指示样本聚类情况热图、用于指示差异表达的样本特征的火山图、以及用于指示生物学过程对应的显著性功能点的富集分析图中的任意组合。

9.一种基于机器学习的临床组学数据处理装置，其特征在于，包括：

数据获取模块，用于获取待处理的组学数据；

特征提取模块，用于从所述待处理的组学数据中提取组学特征的组合作为样本特征子集；

分析模块，用于根据所述样本特征子集进行组学特征分析得到验证结果；

重要程度值获取模块，用于根据所述组学特征对所述验证结果值的影响幅度确定所述组学特征的重要程度值；

筛选模块，用于根据所述组学特征的重要程度值筛选出目标特征子集；以及

预测模块，用于根据所述目标特征子集对所述组学数据进行组学特征分析得到预测结果。

10.一种服务器，其特征在于，包括：

一个或多个处理器；

存储器；

一个或多个应用程序，其中所述一个或多个应用程序被存储在所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个程序配置用于执行如权利要求1-8任一项所述的方法。

11.一种计算机可读取存储介质，其特征在于，所述计算机可读取存储介质中存储有程序代码，所述程序代码可被处理器调用执行如权利要求1-8任一项所述的方法。

一种基于机器学习的临床组学数据处理方法及装置

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种基于机器学习的临床组学数据处理方法、装置、服务器及存储介质。

背景技术

[0002] 机器学习(Machine Learning,ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0003] 人体在其生命周期的不同阶段以及疾病发展的不同阶段，其基因表达和蛋白表达可能存在巨大的差异。因此组学(基因组学，转录组学，蛋白组学和代谢组学等)是系统地研究生物学规律的重要工具，同时也可反映出机体所处的生命周期阶段以及疾病发展情况。

[0004] 当前临床样本获得越来越多的组学数据，组学数据在精准医疗中起到至关重要的作用。作为功能的最终执行者，蛋白质功能的改变是所有生理病理过程变化的直接原因，蛋白组学的研究对于疾病的诊断、分型和预测有着不可替代的优势。然而现有的蛋白组学研究多集中在蛋白测序、定量、差异蛋白分析，生物标记(Biomarker)筛选以及功能性分析上，此过程需要大量的人工干预，效率低下。

发明内容

[0005] 有鉴于此，有必要提供一种基于机器学习的临床组学数据处理方法、装置、服务器及存储介质，其可解决现有技术中组学数据处理及运用时处理效率低下的问题。

[0006] 一方面，本申请提供一种基于机器学习的临床组学数据处理方法，其包括以下步骤：

[0007] 获取待处理的组学数据；从所述待处理的组学数据中提取组学特征的组合作为样本特征子集；根据所述样本特征子集进行组学特征分析得到模型验证结果；根据所述组学特征对所述验证结果值的影响幅度确定所述组学特征的重要程度值；根据所述组学特征的重要程度值筛选出目标特征子集；以及根据所述目标特征子集对所述组学数据进行组学特征分析得到用于表征所述组学数据对应的分类或者生理指标的预测值。。

[0008] 第二方面，本申请还提供一种基于机器学习的临床组学数据处理装置，包括：

[0009] 数据获取模块，用于获取待处理的组学数据；

[0010] 特征提取模块，用于从所述待处理的组学数据中提取组学特征的组合作为特征子集；

[0011] 分析模块，用于根据所述特征子集进行组学特征分析得到模型验证结果；

[0012] 重要程度值获取模块，用于根据所述组学特征对所述验证结果值的影响幅度确定

所述组学特征的重要程度值；

[0013] 筛选模块，用于根据所述组学特征的重要程度值筛选出目标特征子集；以及

[0014] 预测模块，用于根据所述目标特征子集对所述组学数据进行组学特征分析得到用于表征所述组学数据对应的分类或者生理指标的预测值。

[0015] 第三方面，本申请实施例还提供一种服务器，包括：

[0016] 一个或多个处理器；

[0017] 存储器；

[0018] 一个或多个应用程序，其中所述一个或多个应用程序被存储在所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个程序配置用于执行上述第一方面提供的基于机器学习的组学数据处理方法。

[0019] 第四方面，本申请实施例还提供一种计算机可读取存储介质，计算机可读取存储介质中存储有程序代码，所述程序代码可被处理器调用执行上述第一方面提供的基于机器学习的组学数据处理方法。

[0020] 根据以上的基于机器学习的临床组学数据处理方法，通过机器学习模型训练筛选最优特征子集，之后基于此特征子集进行模型的训练和测试。相比于原始特征，此特征子集移除与判断不相关的特征的干扰，同时降低了特征维度，因此能够得到更加准确的预测结果。

[0021] 在模型预测的同时，本方案通过四个模型解释子模块从特征层面，算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法，从特征筛选到模型的训练与解释均为算法自动进行，无需人工干涉，极大的提升了组学数据机器学习的处理效率。

[0022] 为让本发明的上述和其他目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合所附图式，作详细说明如下。

附图说明

[0023] 图1为本申请一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0024] 图2-3为图1所示的方法详细示意图。

[0025] 图4为本申请另一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0026] 图5为本申请另一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0027] 图6为本申请另一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0028] 图7为本申请另一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0029] 图8为本申请另一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0030] 图9为本申请另一个示例性实施例提供的一种基于机器学习的临床组学数据处理方法的流程图。

[0031] 图10为本申请另一个示例性实施例提供的基于机器学习的临床组学数据处理方法的流程图

[0032] 图11为本申请另一个示例性实施例提供一种基于机器学习的临床组学数据处理装置的结构框图。

[0033] 图12为本申请另一个示例性实施例提供一种服务器的结构框图。

[0034] 图13为本申请另一个示例性实施例提供一种存储介质的结构框图。

具体实施方式

[0035] 为更进一步阐述本发明为实现预定发明目的所采取的技术手段及功效,以下结合附图及较佳实施例,对依据本发明的具体实施方式、结构、特征及其功效,详细说明如后。

[0036] 参阅图1到图3,本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法,该方法包括以下步骤:

[0037] 步骤S101,获取已被标定类别标签的训练样本。

[0038] 在一个具体的实施方式中,上述的训练样本为蛋白组学(Proteomics)数据。蛋白质组(Proteome)一词,源于蛋白质(Protein)与基因组(Genome)两个词的组合,意指“一种基因组所表达的全套蛋白质”,即包括一种细胞乃至一种生物所表达的全部蛋白质。蛋白质组学本质上指的是在大规模水平上研究蛋白质的特征,包括蛋白质的表达水平,翻译后的修饰,蛋白与蛋白相互作用等,由此获得蛋白质水平上的关于疾病发生,细胞代谢等过程的整体而全面的认识。

[0039] 蛋白质组的研究不仅能为生命活动规律提供物质基础,也能为众多种疾病机理的阐明及攻克提供理论根据和解决途径。通过对正常个体及病理个体间的蛋白质组比较分析,可以找到某些“疾病特异性的蛋白质分子”,它们可成为新药物设计的分子靶点,或者也会为疾病的早期诊断提供分子标志。

[0040] 在一个具体的实施方式中,上述的训练样本为基因组学(Genomics)数据。基因组学是对生物体所有基因进行集体表征、定量研究及不同基因组比较研究的一门交叉生物学学科。基因组学主要研究基因组的结构、功能、进化、定位和编辑等,以及它们对生物体的影响。

[0041] 在一个具体的实施方式中,上述的训练样本为转录组学数据。转录组学是指一门在整体水平上研究细胞中基因转录的情况及转录调控规律的学科。转录组学是从核糖核酸(Ribonucleic Acid,RNA)水平研究基因表达的情况。转录组即一个活细胞所能转录出来的所有RNA的总和,是研究细胞表型和功能的一个重要手段。以脱氧核糖核酸(DeoxyriboNucleic Acid,DNA)为模板合成RNA的转录过程是基因表达的第一步,也是基因表达调控的关键环节。所谓基因表达,是指基因携带的遗传信息转变为可辨别的表型的整个过程。与基因组不同的是,转录组的定义中包含了时间和空间的限定。同一细胞在不同的生长时期及生长环境下,其基因表达情况是不完全相同的。通常,同一种组织表达几乎相同的一套基因以区别于其他组织,如脑组织或心肌组织等分别只表达全部基因中不同的30%而显示出组织的特异性。

[0042] 在一个具体的实施方式中,上述的训练样本为代谢组学(Metabonomics/Metabolomics)数据。代谢组学是效仿基因组学和蛋白质组学的研究思想,对生物体内所有

代谢物进行定量分析，并寻找代谢物与生理病理变化的相对关系的研究方式，是系统生物学的组成部分。其研究对象大都是相对分子质量1000以内的小分子物质。

[0043] 类别标签是指由工作人员如医生解析不同的测试样本数据而得到，其代表了经专业训练的医生对于样本数据的判定结果。

[0044] 步骤S102，将所述训练样本随机分成多个个子集。

[0045] 步骤S103，将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别训练得到多个子模型，所述子模型输出每个样本特征的重要程度值排序；

[0046] 在一个具体实施方式中，步骤S102具体包括以下步骤：将训练样本平均分成N份，选其中的N-1份作为训练集，剩余的1份作为验证集。可以理解，以上的组合共计N种。

[0047] 例如，将训练样本分成5份，选其中4份作为训练集，其余的1份作为验证集。可以理解，每个训练样本作为验证集，总计有5中情形。进行上述的交叉验证过程，可以防止数据过并拟合，去除不必要的特征。

[0048] 针对以上的N种情形，分别进行训练，则可以各到N个子模型。

[0049] 在一个具体的实施方式中，上述的训练采用超梯度树提升(extremeGradientBoosting，XGBoost)分类器模型，选择树模型为基分类器，将训练集的样本特征输入XGBoost，目标函数设置为二分类的逻辑回归问题，评估指标为受试者工作特征曲线的面积(Receiver Operating Characteristic Area Under the Curve，ROC-AUC)，早停次数例如可设置设置为50步，通过参数的自动搜索和交叉验证确定树的最优棵数，树的最大深度，正则化系数等超参数的取值。可以理解，此处的参数并不限于50，任意合适的参数均可使用。

[0050] 在每个子模型中，完成对XGBoost分类器的训练之后，根据每个特征在树中出现的次数以及XGBoost分类器对各个特征的评分输出每个特征的重要程度值，将各个子模型中的特征重要程度值求平均，然后将所有的输入特征根据平均重要程度从大到小排列。

[0051] 步骤S104，在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用所述特征子集对所述训练集进行训练得到验证子模型，其中N为正整数。

[0052] 验证子模型的训练方法与前面步骤S103中提到的方法相似，不同之处在于，此时只取N个最重要的样本特征作，其他的特征不再作为特征输入训练模型。

[0053] 在一个具体的实施方式中，在每个子模型中，通过实验确定最优特征组合。具体地，每次取前N个最重要的特征作为每个样本的特征子集。本方案依次取N＝1,2,3,4,5,8,10,15，对每种特征子集单独训练一个XGBoost分类器，因此此过程一共训练得到8个XGBoost分类器(分类器的超参搜索以及训练同以上步骤S103。

[0054] 步骤S105，将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果，并根据所述模型验证结果确定最优特征子集。

[0055] 如前所述，验证子模型只采用N个最重要的样本特征进行训练得到，因此在验证阶段，对验证集的每个样本，也取前N个最重要的特征作为其特征子集，然后输入到对应的XGBoost分类器进行结果预测得到模型验证结果。然后，将模型输出的结果，即模型验证结果与人工已经标定的结果进行比对分析、运算得到模型质量的评分。可以理解，模型预测与准确，与标定结果相似度越高，其评分越高。依据此评分结果，即可确定最优的特征子集，亦即，使模型的预测结果与标定结果匹配程度最高的特征子集即为最优的特征子集。

[0056] 步骤S106,对所有所述训练样本,采用所述最优特征子集进行训练得到预测模型。

[0057] 在获取到最优特征子集后,将其输入到XGBoost分类器进行训练,保存训练得到预测模型。

[0058] 步骤S107,将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0059] 对组学数据测试集中的每个样本,取N个最优特征组合作为样本特征子集。之后,将样本特征子集送到预测模型进行结果预测以及输出模型的图形化界面。

[0060] 在一个具体的实施方式中,如图3所示,基于沙谱利附加解释(Shapley Additive Explanation,SHAP)的特征重要程度值分析,根据每个特征对模型预测值的影响确定其重要程度值,并做出重要程度值可视化的柱状图,从而解释模型判断的依据,并且有助于对模型的检验和修正;绘制热图(Heatmap)对取不同特征子集情况下的样本聚类情况可视化,从而对特征重要程度值和最优特征组合的筛选进行解释和验证;绘制火山图(Volcano)直观的显示差异表达的特征,从而对特征重要程度值和最优特征组合的筛选进行解释和验证;通过因本体(Gene Ontology,GO)富集分析得到生物学过程对应的显著性功能点图,为模型提供生物学层面的解释。

[0061] 根据本实施例提供的基于机器学习的临床组学数据处理方法,通过XGBoost分类器的训练筛选最优特征子集,之后基于此特征子集进行模型的训练和测试。相比于原始特征,此特征子集移除与判断不相关的特征的干扰,同时降低了特征维度,因此能够得到更加准确的预测结果。

[0062] 在模型预测的同时,本方案通过四个模型解释子模块从特征层面,算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法,从特征筛选到模型的训练与解释均为算法自动进行,无需人工干涉,极大的提升了组学数据机器学习的处理效率。

[0063] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式,节省了分析成本和等待时间,避免人为实验误差所带来的问题,使用机器学习模型,可以实现对组学数据的自动分析,同时提供模型解释,为理解和检验模型提供了依据。

[0064] 参阅图4,本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法,该方法包括以下步骤:

[0065] 步骤S101,获取已被标定类别标签的训练样本。

[0066] 在一个具体的实施方式中,上述的训练样本为蛋白组学(Proteomics)数据、基因组学数据、转录组学数据、代谢组学数据或其任意组合。

[0067] 步骤S102,将所述训练样本随机分成多个个子集。

[0068] 步骤S201,将所述多个子集分成训练集与验证集,根据训练集与验证集的不同组合方案分别采用逻辑回归模型(Logistic Regression,LR)训练得到多个子模型,所述子模型输出每个样本特征的重要程度值排序。

[0069] 逻辑回归是离散选择法模型之一,属于多重变量分析范畴,是社会学、生物统计学、临床、数量心理学、计量经济学、市场营销等统计实证分析的常用方法。逻辑回归一般用于二分类(Binary Classification)问题中,给定一些输入,输出结果是离散值。例如用逻辑回归实现一个猫分类器,输入一张图片x,预测图片是否为猫,输出该图片中存在猫的概率结果y。从生物学的角度讲:就是一个模型对外界的刺激(训练样本)做出反应,趋利避害(评价标准)。应用于组学数据的特征预测中,可以根据输入的样本数据预测到期望的指标,

例如存活寿命。

[0070] 步骤S104,在每个所述子模型中,依次取N个最重要的样本特征作为每个样本的特征子集,重新采用所述特征子集对所述训练集进行训练得到验证子模型,其中N为正整数。

[0071] 验证子模型的训练方法与前面步骤S103中提到的方法相似,不同之处在于,此时只取N个最重要的样本特征作,其他的特征不再作为特征输入训练模型。

[0072] 步骤S105,将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果,并根据所述模型验证结果确定最优特征子集。

[0073] 步骤S106,对所有所述训练样本,采用所述最优特征子集进行训练得到预测模型。

[0074] 步骤S107,将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0075] 根据本实施例提供的基于机器学习的临床组学数据处理方法,采用逻辑回归模型训练筛选最优特征子集,之后基于此特征子集进行模型的训练和测试。相比于原始特征,此特征子集移除与判断不相关的特征的干扰,同时降低了特征维度,因此能够得到更加准确的预测结果。

[0076] 在模型预测的同时,本方案通过四个模型解释子模块从特征层面,算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法,从特征筛选到模型的训练与解释均为算法自动进行,无需人工干涉,极大的提升了组学数据机器学习的处理效率。

[0077] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式,节省了分析成本和等待时间,避免人为实验误差所带来的问题,使用机器学习模型,可以实现对组学数据的自动分析,同时提供模型解释,为理解和检验模型提供了依据。

[0078] 参阅图5,本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法,该方法包括以下步骤:

[0079] 步骤S101,获取已被标定类别标签的训练样本。

[0080] 在一个具体的实施方式中,上述的训练样本为蛋白组学(Proteomics)数据、基因组学数据、转录组学数据、代谢组学数据或其任意组合。

[0081] 步骤S102,将所述训练样本随机分成多个个子集。

[0082] 步骤S301,将所述多个子集分成训练集与验证集,根据训练集与验证集的不同组合方案分别采用感知机模型(Perceptron)模型训练得到多个子模型,所述子模型输出每个样本特征的重要程度值排序。

[0083] 感知机模型是一种二分类的线性分类器,只能处理线性可分的问题,感知机的模型就是尝试找到一个超平面将数据集分开,在二维空间这个超平面就是一条直线,在三维空间就是一个平面。感知机的分类模型如下:

[0084] $f(x)=\mathrm{sign}(w\cdot x+b)$

[0085] sign函数是指示函数(当$wx+b>0$,$f(x)=+1$;当$wx+b<0$,$f(x)=-1$;感知机的超平面是$wx+b=0$)

[0086] $$\mathbf{sign}(x)=\begin{cases}+1, & x\geqslant 0\\ -1, & x<0\end{cases}$$

[0087] 将上述分段函数整合成$y(wx+b)>0$,则满足该式子的样本点即分类正确的点,不满足的即分类错误的点,目标就是找到这样一组参数w,b使得将训练集中的正类点和负类点

分开。

[0088] 接下来定义损失函数(损失函数是一种衡量损失和错误的程度的函数),可以通过定义分类错误的样本的个数来作为损失函数,但是这种损失函数不是参数w,b的连续可导函数,因此不容易优化。对于误分类的点有-y(wx+b)>0,让所有的误分类点到超平面的距离和最小(注意:感知机的损失函数只针对误分类点,而不是整个训练集):

[0089] $$-\frac{1}{\|w\|}\sum_{x_i\in M} y_i(w\cdot x_i+b)$$

[0090] 其中M是表示误分类的样本集合,当w,b成倍数增大时,并不会改变超平面,||w||的值也会相应的增大,因此令||w||=1不会影响结果。最终的感知机损失函数如下:

[0091] $$L(w,b)=-\sum_{x_i\in M} y_i(w\cdot x_i+b)$$

[0092] 此外,上述的感知机,还可以采用多层结构,即多层感知机(Multi-Layer Perceptron,MLP)模型。

[0093] 步骤S104,在每个所述子模型中,依次取N个最重要的样本特征作为每个样本的特征子集,重新采用所述特征子集对所述训练集进行训练得到验证子模型,其中N为正整数。

[0094] 验证子模型的训练方法与前面步骤S103中提到的方法相似,不同之处在于,此时只取N个最重要的样本特征作,其他的特征不再作为特征输入训练模型。

[0095] 步骤S105,将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果,并根据所述模型验证结果确定最优特征子集。

[0096] 步骤S106,对所有所述训练样本,采用所述最优特征子集进行训练得到预测模型。

[0097] 步骤S107,将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0098] 根据本实施例提供的基于机器学习的临床组学数据处理方法,采用感知机模型训练筛选最优特征子集,之后基于此特征子集进行模型的训练和测试。相比于原始特征,此特征子集移除与判断不相关的特征的干扰,同时降低了特征维度,因此能够得到更加准确的预测结果。

[0099] 在模型预测的同时,本方案通过四个模型解释子模块从特征层面,算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法,从特征筛选到模型的训练与解释均为算法自动进行,无需人工干涉,极大的提升了组学数据机器学习的处理效率。

[0100] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式,节省了分析成本和等待时间,避免人为实验误差所带来的问题,使用机器学习模型,可以实现对组学数据的自动分析,同时提供模型解释,为理解和检验模型提供了依据。

[0101] 参阅图6,本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法,该方法包括以下步骤:

[0102] 步骤S101,获取已被标定类别标签的训练样本。

[0103] 在一个具体的实施方式中,上述的训练样本为蛋白组学(Proteomics)数据、基因组学数据、转录组学数据、代谢组学数据或其任意组合。

[0104] 步骤S102,将所述训练样本随机分成多个个子集。

[0105] 步骤S401,将所述多个子集分成训练集与验证集,根据训练集与验证集的不同组

合方案分别采用支持向量机模型(Support Vector Machine,SVN)模型训练得到多个子模型,所述子模型输出每个样本特征的重要程度值排序。

[0106] 在上述的感知机模型中,目标是将训练集分开,只要是能将样本分开的超平面都满足要求,而这样的超平面有很多。支持向量机本质上和感知机类似,然而要求却更加苛刻,在分类过程中,那些远离超平面的点是安全的,而那些容易被误分类的点是离超平面很近的点,而支持向量机的思想就是要重点关注这些离超平面很近的点,一句话就是在分类正确的同时,让离超平面最近的点到超平面的间隔最大。

[0107] 基于上面的感知机可以将目标表示为:

$$\max_{w,b} \quad \gamma$$

[0108]

$$\text{s.t.} \quad y_i\left(\frac{w}{\|w\|}\cdot x_i+\frac{b}{\|w\|}\right)\geqslant\gamma, \quad i=1,2,\cdots,N$$

[0109] γ是离超平面最近的点的到超平面的几何间隔,将几何间隔用函数间隔替代,可以将式子表示为:

[0110]

$$\max_{w,b} \quad \frac{\hat{\gamma}}{\|w\|}$$

$$\text{s.t.} \quad y_i(w\cdot x_i+b)\geqslant\hat{\gamma}, \quad i=1,2,\cdots,N$$

[0111] γ(帽子)表示的是函数间隔,而函数间隔的取值是会随着w,b成倍数变化而变化的,并不会影响最终的结果,因此令γ(帽子)=1,则我们最终的问题可以表述为:

[0112]

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y_i(w\cdot x_i+b)-1\geqslant 0, \quad i=1,2,\cdots,N$$

[0113] 在这里引出了支持向量机的第一个亮点:最大化间隔,最大化间隔能使得分类更加精确,且该最大间隔超平面是存在且唯一的。

[0114] 上面的问题中的1/2||w||2是凸函数,同时约束不等式是仿射函数,因此这是一个凸二次规划问题,根据凸优化理论,可以借助拉格朗日函数将约束问题转化为无约束的问题来求解,优化函数可以表达为:

[0115]

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i$$

[0116] α_i是拉格朗日乘子,$\alpha_i\geq 0$i=1,2,3,.....,n。

[0117] 根据拉格朗日的对偶性,可以将原始问题转化为对偶问题(只要对偶问题存在,对偶问题的最优化解就是原问题的最优化解,一般对偶问题都比原始问题更容易求解)极大极小问题:

[0118]

$$\max_{\alpha}\min_{w,b} L(w,b,\alpha)$$

[0119] 先对w,b求导求极小问题,可以得到w,b的值:

[0120] $$w=\sum_{i=1}^{N}\alpha_i y_i x_i$$

$$\sum_{i=1}^{N}\alpha_i y_i=0$$

[0121] 将求得的解代入到拉格朗日函数中，可以得到下面的优化函数(将代入后原本的求α的极大问题转换成了极小问题)：

[0122] $$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\alpha_i\geqslant 0,\quad i=1,2,\cdots,N$$

[0123] 因此只需要求得我们的α的值就可以求得我们的w，b的值(求α的常用算法是SMO算法)假设最终求得的α的值为α*，则w，b可以表述成：

[0124] $$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$$

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j)$$

[0125] 引入KTT条件(KTT条件是上面拉格朗日函数求最优解的必要条件)：

[0126] $\alpha_i^*(y_i(w^*\cdot x_i+b^*)-1)=0,i=1,2,\cdots,N$

[0127] $y_i(w^*\cdot x_i+b^*)-1\geq 0,i=1,2,\cdots,N$

[0128] $\alpha_i^*\geq 0,i=1,2,\cdots,N$

[0129] 从KTT条件中可以看出，当yi(w*xi+b*)-1＞0时，αi*＝0；当αi*＞0时，yi(w*xi+b*)-1＝0；

[0130] 结合上面的w，b表达式可以引出支持向量机的第二个亮点：w，b参数只与满足yi(w*xi+b*)-1＝0的样本有关，而这些样本点就是离最大间隔超平面最近的点，将这些点称之为支持向量。因此很多时候支持向量在小样本集分类时也能表现的很好，也正是因为这个原因。另外需注意α向量的个数是和训练集数量相等的，对于大的训练集，会导致所需要的参数数量增多，因此SVM在处理大的训练集时会比其他常见的机器学习算法要慢。

[0131] 步骤S104，在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用所述特征子集对所述训练集进行训练得到验证子模型，其中N为正整数。

[0132] 验证子模型的训练方法与前面步骤S103中提到的方法相似，不同之处在于，此时只取N个最重要的样本特征作，其他的特征不再作为特征输入训练模型。

[0133] 步骤S105，将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果，并根据所述模型验证结果确定最优特征子集。

[0134] 步骤S106，对所有所述训练样本，采用所述最优特征子集进行训练得到预测模型。

[0135] 步骤S107,将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0136] 根据本实施例提供的基于机器学习的临床组学数据处理方法,采用支持向量机模型训练筛选最优特征子集,之后基于此特征子集进行模型的训练和测试。相比于原始特征,此特征子集移除与判断不相关的特征的干扰,同时降低了特征维度,因此能够得到更加准确的预测结果。

[0137] 在模型预测的同时,本方案通过四个模型解释子模块从特征层面,算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法,从特征筛选到模型的训练与解释均为算法自动进行,无需人工干涉,极大的提升了组学数据机器学习的处理效率。

[0138] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式,节省了分析成本和等待时间,避免人为实验误差所带来的问题,使用机器学习模型,可以实现对组学数据的自动分析,同时提供模型解释,为理解和检验模型提供了依据。

[0139] 参阅图7,本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法,该方法包括以下步骤:

[0140] 步骤S101,获取已被标定类别标签的训练样本。

[0141] 在一个具体的实施方式中,上述的训练样本为蛋白组学(Proteomics)数据、基因组学数据、转录组学数据、代谢组学数据或其任意组合。

[0142] 步骤S102,将所述训练样本随机分成多个个子集。

[0143] 步骤S501,将所述多个子集分成训练集与验证集,根据训练集与验证集的不同组合方案分别采用全连接神经网络模型训练得到多个子模型,所述子模型输出每个样本特征的重要程度值排序。

[0144] 全连接的意思即为多层的神经网络,下一层的每个节点都和上一层的所有节点相连,构成一个感知机模型。这种全连接的网络是相对来说参数最多的神经网络。根据上述所述步骤,单层感知机就处理过程相似。定义好各层的激活函数后,模型就建立好了。然后根据是二分类、多分类或回归来定义损失函数,然后使用梯度下降即可。

[0145] 步骤S104,在每个所述子模型中,依次取N个最重要的样本特征作为每个样本的特征子集,重新采用所述特征子集对所述训练集进行训练得到验证子模型,其中N为正整数。

[0146] 验证子模型的训练方法与前面步骤S103中提到的方法相似,不同之处在于,此时只取N个最重要的样本特征作,其他的特征不再作为特征输入训练模型。

[0147] 步骤S105,将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果,并根据所述模型验证结果确定最优特征子集。

[0148] 步骤S106,对所有所述训练样本,采用所述最优特征子集进行训练得到预测模型。

[0149] 步骤S107,将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0150] 根据本实施例提供的基于机器学习的临床组学数据处理方法,采用全连接神经网络训练筛选最优特征子集,之后基于此特征子集进行模型的训练和测试。相比于原始特征,此特征子集移除与判断不相关的特征的干扰,同时降低了特征维度,因此能够得到更加准确的预测结果。

[0151] 在模型预测的同时,本方案通过四个模型解释子模块从特征层面,算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法,从特征筛选到模型的训练与解释均为算法自动进行,无需人工干涉,极大的提升了组学数据机器学习的处理效率。

[0152] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式，节省了分析成本和等待时间，避免人为实验误差所带来的问题，使用机器学习模型，可以实现对组学数据的自动分析，同时提供模型解释，为理解和检验模型提供了依据。

[0153] 参阅图8，本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法，该方法包括以下步骤：

[0154] 步骤S101，获取已被标定类别标签的训练样本。

[0155] 在一个具体的实施方式中，上述的训练样本为蛋白组学(Proteomics)数据、基因组学数据、转录组学数据、代谢组学数据或其任意组合。

[0156] 步骤S102，将所述训练样本随机分成多个个子集。

[0157] 步骤S601，将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用长短期记忆网络模型(Long Short-Term Memory，LSTM)训练得到多个子模型，所述子模型输出每个样本特征的重要程度值排序。

[0158] LSTM是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。LSTM是解决循环神经网络中存在的“梯度消失”问题而提出的，是一种特殊的循环神经网络。最常见的一个例子就是：当我们要预测“the clouds are in the(...)″的时候，这种情况下，相关的信息和预测的词位置之间的间隔很小，会使用先前的信息预测出词是”sky”。但是如果想要预测”I grew up in France...I speak fluent(...)”，语言模型推测下一个词可能是一种语言的名字，但是具体是什么语言，需要用到间隔很长的前文中France，在这种情况下，因为“梯度消失”的问题，并不能利用间隔很长的信息，然而，LSTM在设计上明确避免了长期依赖的问题，这主要归功于LSTM精心设计的“门”结构(输入门、遗忘门和输出门)消除或者增加信息到细胞状态的能力，使得LSTM能够记住长期的信息。

[0159] 步骤S104，在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用所述特征子集对所述训练集进行训练得到验证子模型，其中N为正整数。

[0160] 验证子模型的训练方法与前面步骤S103中提到的方法相似，不同之处在于，此时只取N个最重要的样本特征作，其他的特征不再作为特征输入训练模型。

[0161] 步骤S105，将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果，并根据所述模型验证结果确定最优特征子集。

[0162] 步骤S106，对所有所述训练样本，采用所述最优特征子集进行训练得到预测模型。

[0163] 步骤S107，将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0164] 根据本实施例提供的基于机器学习的临床组学数据处理方法，采用长短期记忆网络模型训练筛选最优特征子集，之后基于此特征子集进行模型的训练和测试。相比于原始特征，此特征子集移除与判断不相关的特征的干扰，同时降低了特征维度，因此能够得到更加准确的预测结果。

[0165] 在模型预测的同时，本方案通过四个模型解释子模块从特征层面，算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法，从特征筛选到模型的训练与解释均为算法自动进行，无需人工干涉，极大的提升了组学数据机器学习的处理效率。

[0166] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式，节省了分析成本和等待时间，避免人为实验误差所带来的问题，使用机器学习模型，可以实现对组学数据

的自动分析,同时提供模型解释,为理解和检验模型提供了依据。

[0167] 参阅图9,本申请一个示例性实施例提供一种基于机器学习的临床组学数据处理方法,该方法包括以下步骤:

[0168] 步骤S101,获取已被标定类别标签的训练样本。

[0169] 在一个具体的实施方式中,上述的训练样本为蛋白组学(Proteomics)数据、基因组学数据、转录组学数据、代谢组学数据或其任意组合。

[0170] 步骤S102,将所述训练样本随机分成多个个子集。

[0171] 步骤S701,将所述多个子集分成训练集与验证集,根据训练集与验证集的不同组合方案分别采用图卷神经网络模型(Graph Convolutional Network,GCN)训练得到多个子模型,所述子模型输出每个样本特征的重要程度值排序。

[0172] 以GNN为基础的图卷积神经网络GCN是对卷积神经网络在图领域的自然推广。它能同时对节点特征信息与结构信息进行端对端学习,是目前对图数据学习任务的较优选择。GCN精妙地设计了一种从图数据中提取特征的方法,从而让我们可以使用这些特征去对图数据进行节点分类(node classification)、图分类(graph classification)、边预测(link prediction),还可以顺便得到图的嵌入表示(graph embedding)。

[0173] 步骤S104,在每个所述子模型中,依次取N个最重要的样本特征作为每个样本的特征子集,重新采用所述特征子集对所述训练集进行训练得到验证子模型,其中N为正整数。

[0174] 验证子模型的训练方法与前面步骤S103中提到的方法相似,不同之处在于,此时只取N个最重要的样本特征作,其他的特征不再作为特征输入训练模型。

[0175] 步骤S105,将所述验证集的每个样本的所述特征子集输入到所述验证子模型中进行验证得到模型验证结果,并根据所述模型验证结果确定最优特征子集。

[0176] 步骤S106,对所有所述训练样本,采用所述最优特征子集进行训练得到预测模型。

[0177] 步骤S107,将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0178] 根据本实施例提供的基于机器学习的临床组学数据处理方法,采用图卷神经网络模训练筛选最优特征子集,之后基于此特征子集进行模型的训练和测试。相比于原始特征,此特征子集移除与判断不相关的特征的干扰,同时降低了特征维度,因此能够得到更加准确的预测结果。

[0179] 在模型预测的同时,本方案通过四个模型解释子模块从特征层面,算法层面和生物学层面对模型的判断提供了依据。本发明的整个算法,从特征筛选到模型的训练与解释均为算法自动进行,无需人工干涉,极大的提升了组学数据机器学习的处理效率。

[0180] 上述方案摆脱了基于统计学检验方法和人工判断为主的分析方式,节省了分析成本和等待时间,避免人为实验误差所带来的问题,使用机器学习模型,可以实现对组学数据的自动分析,同时提供模型解释,为理解和检验模型提供了依据。

[0181] 参阅图10,其示出了本申请一个示例性实施例提供的基于机器学习的临床组学数据处理方法的流程图,该方法包括:

[0182] 步骤S201,获取待处理的组学数据。

[0183] 此处的组学数据是指蛋白组不字数据、基因组学数据、转录组学数据或者代谢组学数据。

[0184] 步骤S202,从所述待处理的组学数据中提取组学特征的组合作为特征子集。

[0185] 样本特征是指计算机可识别的数据特征，其可以表征为一个数据范围、向量、数据组合、图形特征等等。样本特征子集里可以包括一个或者多个样本特征。

[0186] 步骤S203，根据所述特征子集进行组学特征分析得到模型验证结果。

[0187] 如上所述，可以预告采用已经标记好训练数据，采用机器学习的方式训练得到验证模型，将步骤S202中获取得到的样本特征子集输入到该验证模型中，即可得到模型验证结果。

[0188] 步骤S204，根据所述组学特征对所述验证结果值的影响幅度确定所述组学特征的重要程度值。

[0189] 在一个具体的实施方式中，上述的机器学习采用超梯度树提升(extremeGradientBoosting，XGBoost)分类器模型。完成对XGBoost分类器的训练之后，根据每个特征在树中出现的次数以及XGBoost分类器对各个特征的评分输出每个特征的重要程度值，将各个子模型中的特征重要程度值求平均，然后将所有的输入特征根据平均重要程度从大到小排列。可以理解，此处的重要程度值获取方法并不限于采用上述的方法，其还可以是以下方法：沙谱利附加解释法、基尼重要性、遗传算法、方差分析、T检验及曼-惠特尼秩和检验的任意组合。

[0190] 步骤S205，根据所述组学特征的重要程度值筛选出目标特征子集。

[0191] 如前所述，验证模型可采用N个最重要的样本特征进行训练得到，因此在验证阶段，对验证集的每个样本，也取前N个最重要的特征作为其特征子集，然后输入到对应的XGBoost分类器进行结果预测得到模型验证结果。然后，将模型输出的结果，即模型验证结果与人工已经标定的结果进行比对分析、运算得到模型质量的评分。从而确定最优的特征子集，即上述的目标特征子集。

[0192] 步骤S206，根据所述目标特征子集对所述组学数据进行组学特征分析得到用于表征所述组学数据对应的分类或者生理指标的预测值。

[0193] 在获取目标特征子集后，采用目标特征子集进行分析，即可将组学数据进行分类，或者分析得到与组学数据对应的生理指标数值。

[0194] 根据本实施例提供的方法，通过采用机器学习的方法对组学数据进行处理，极大的提升了组学数据的处理效率，而通过对目标组学特征的筛选，可以移除与判断不相关的特征的干扰，同时降低了特征维度，因此能够得到更加准确的预测结果。

[0195] 参阅图11，其示出了本申请一个示例性实施例提供的基于机器学习的临床组学数据处理装置的结构框图。该装置包括：

[0196] 样本获取模块101，用于获取已被标定类别标签的训练样本；

[0197] 样本分拆模块102，用于将所述训练样本随机分成多个个子集；

[0198] 交叉验证模块103，用于将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别训练得到多个子模型，所述子模型输出每个样本特征的重要程度值排序；

[0199] 验证子模型获取模块104，用于在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用所述特征子集对所述训练集进行训练得到验证子模型，其中N为正整数；

[0200] 最优特征子集获取模块105，用于将所述验证集的每个样本的所述特征子集输入

到所述验证子模型中进行验证得到模型验证结果，并根据所述模型验证结果确定最优特征子集；

[0201] 预测模型训练模块106，用于对所有所述训练样本，采用所述最优特征子集进行训练得到预测模型；以及

[0202] 预测模块107，用于将待测试样本的所述最优特征子集输入所述预测模型获取预测结果。

[0203] 通过上述的装置，摆脱了基于统计学检验方法和人工判断为主的分析方式，节省了分析成本和等待时间，避免人为实验误差所带来的问题，使用机器学习模型，可以实现对组学数据的自动分析，同时提供模型解释，为理解和检验模型提供了依据。

[0204] 请参阅图12，其示出了本申请实施例提供的一种服务器的结构框图。该服务器100可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上处理器(central processing units，CPU)11和一个或一个以上的存储器12，其中，所述存储器12中存储有至少一条指令，所述至少一条指令由所述处理器11加载并执行以实现上述各个方法实施例提供的方法。当然，该服务器还可以具有有线或无线网络接口、键盘以及输入输出接口等部件，以便进行输入输出，该服务器还可以包括其他用于实现设备功能的部件，在此不做赘述。

[0205] 请参考图13，其示出了本申请实施例提供的一种计算机可读取存储介质的结构框图。该计算机可读取存储介质200中存储有程序代码，所述程序代码可被处理器调用执行上述方法实施例中所描述的方法。

[0206] 计算机可读取存储介质200可以是诸如闪存、EEPROM(电可擦除可编程只读存储器)、EPROM、硬盘或者ROM之类的电子存储器。可选地，计算机可读取存储介质200包括非易失性计算机可读取存储介质(Non-Transitory Computer-Readable Storage Medium)。计算机可读取存储介质200具有执行上述方法中的任何方法步骤的程序代码201的存储空间。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。程序代码201可以例如以适当形式进行压缩。

[0207] 以上所述，仅是本发明的较佳实施例而已，并非对本发明作任何形式上的限制，虽然本发明已以较佳实施例揭示如上，然而并非用以限定本发明，任何本领域技术人员，在不脱离本发明技术方案范围内，当可利用上述揭示的技术内容做出些许更动或修饰为等同变化的等效实施例，但凡是未脱离本发明技术方案内容，依据本发明的技术实质对以上实施例所作的任何简介修改、等同变化与修饰，均仍属于本发明技术方案的范围内。

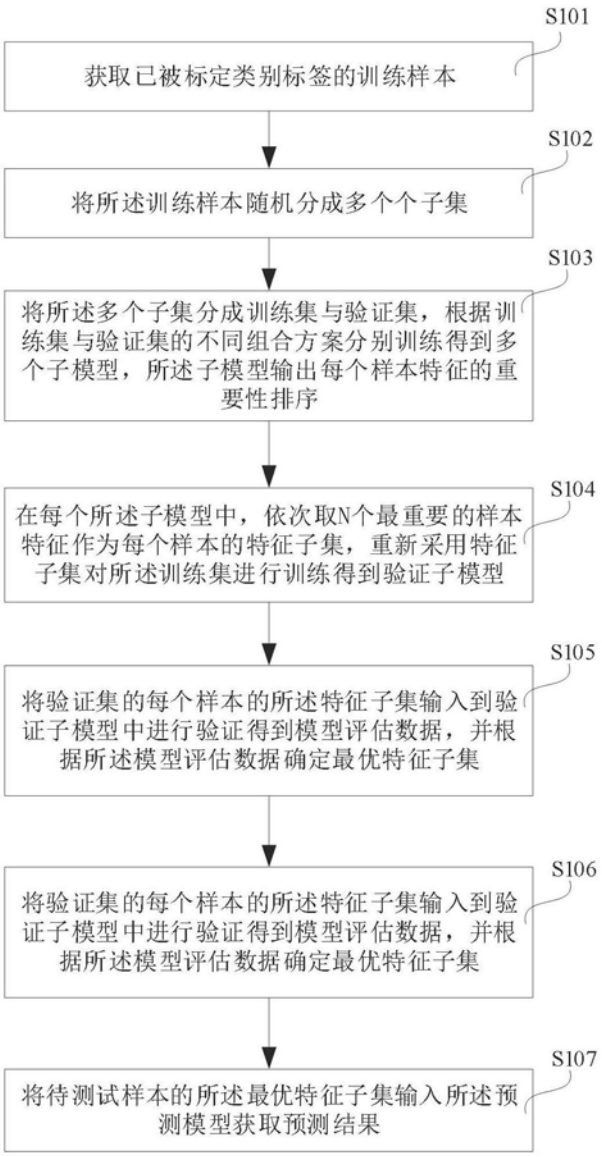

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S103
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图1

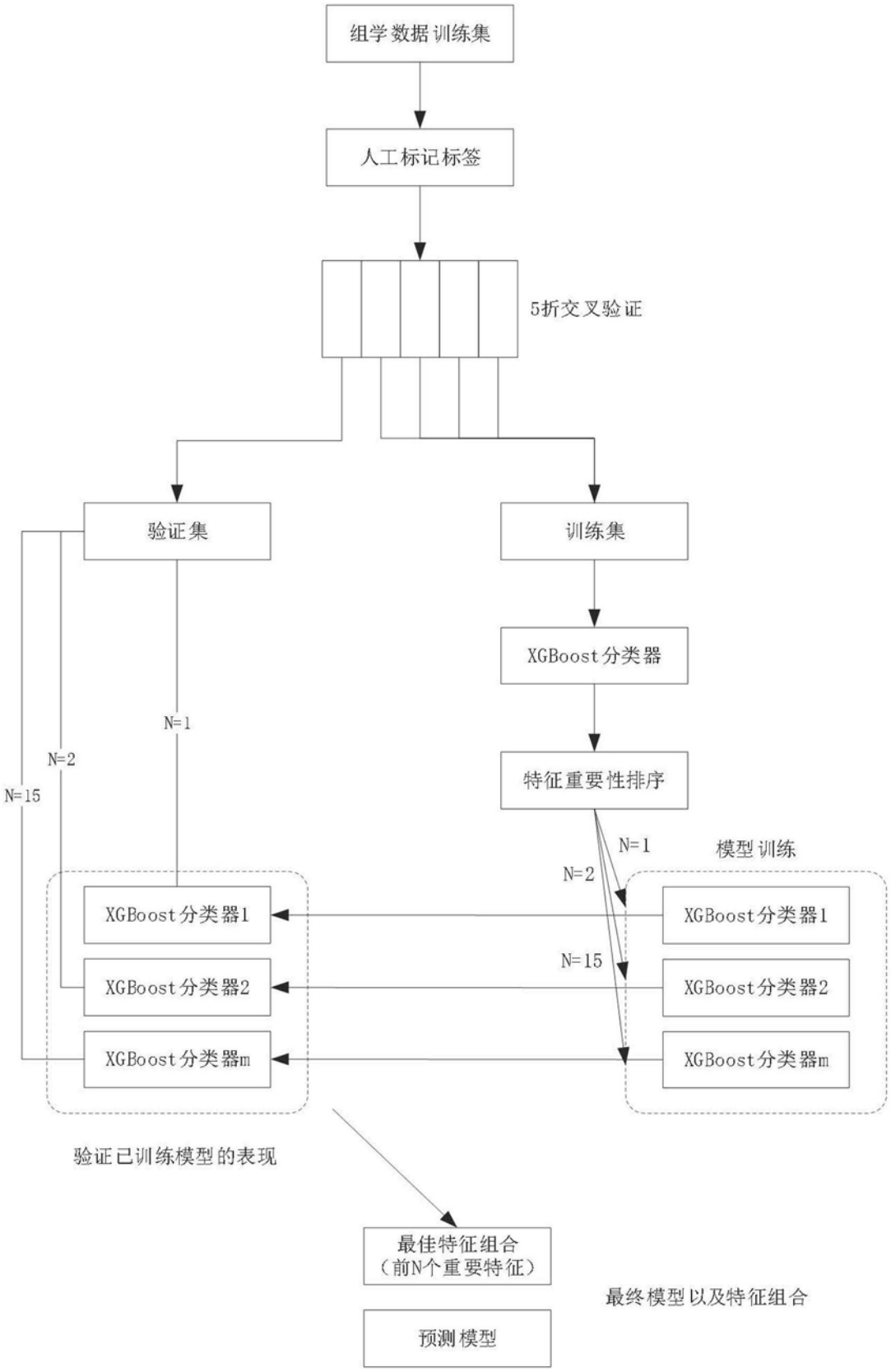

组学数据训练集
人工标记标签
5折交叉验证
验证集
训练集
XGBoost分类器
特征重要性排序
N=1
N=2
N=15
模型训练
XGBoost分类器1
XGBoost分类器2
XGBoost分类器m
验证已训练模型的表现
最佳特征组合
（前N个重要特征）
预测模型
最终模型以及特征组合

图2

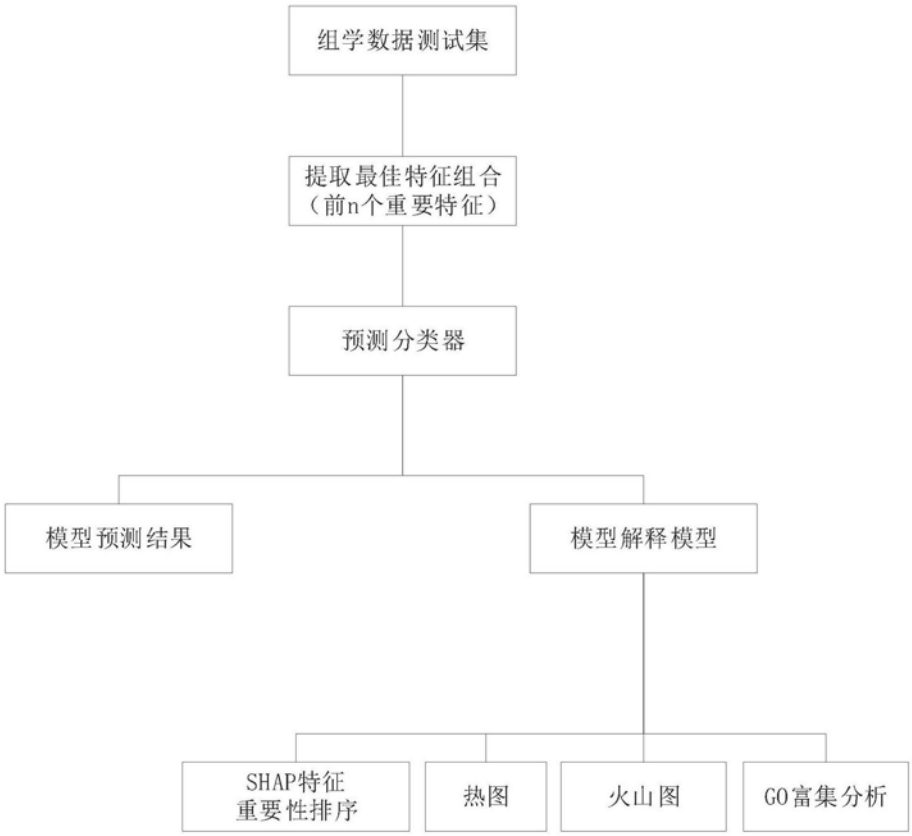

组学数据测试集
提取最佳特征组合
（前n个重要特征）
预测分类器
模型预测结果
模型解释模型
SHAP特征
重要性排序
热图
火山图
GO富集分析

图3

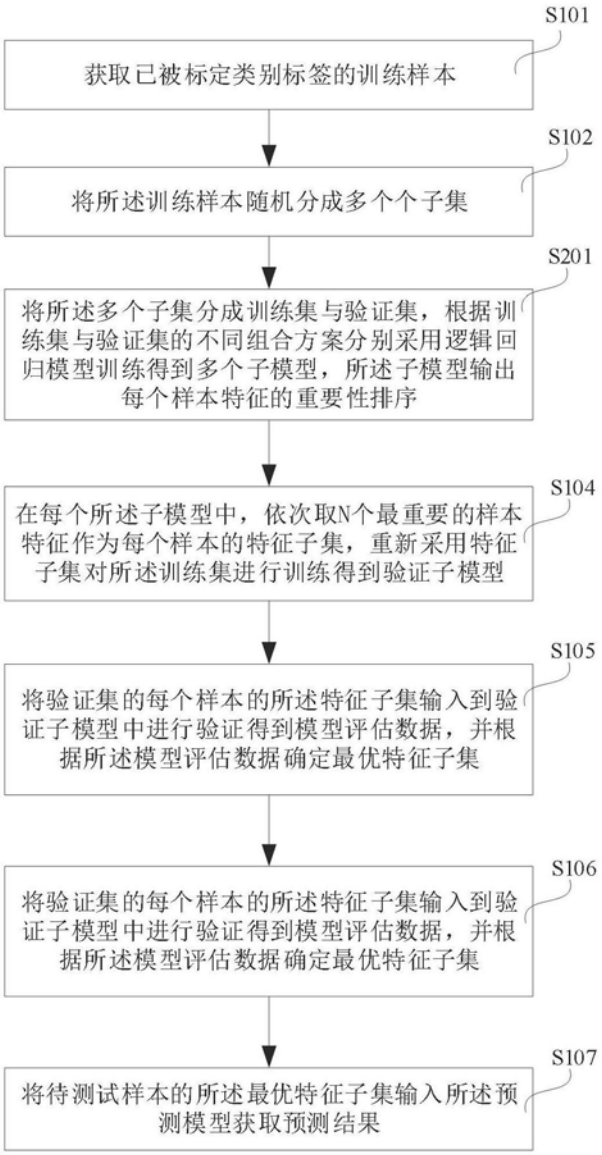

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S201
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用逻辑回归模型训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图4

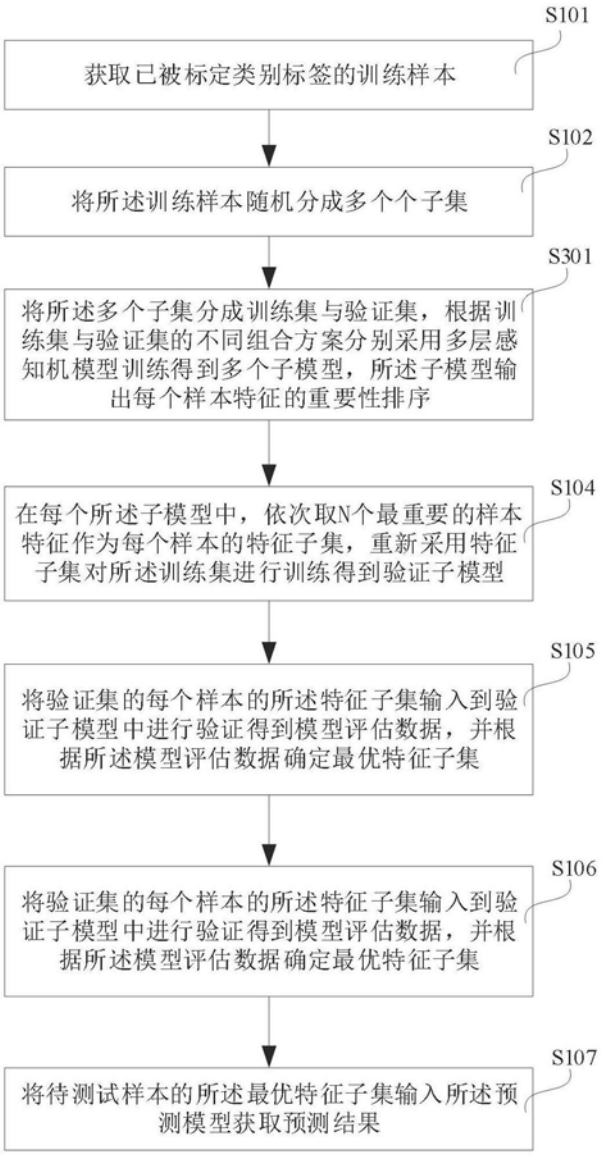

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S301
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用多层感知机模型训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图5

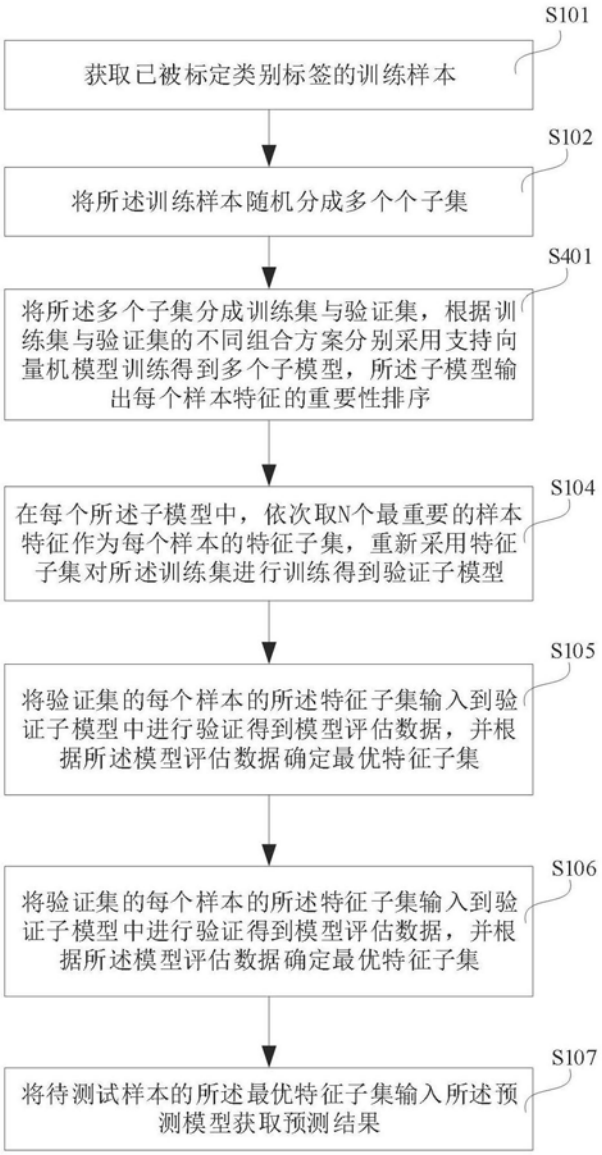

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S401
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用支持向量机模型训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图6

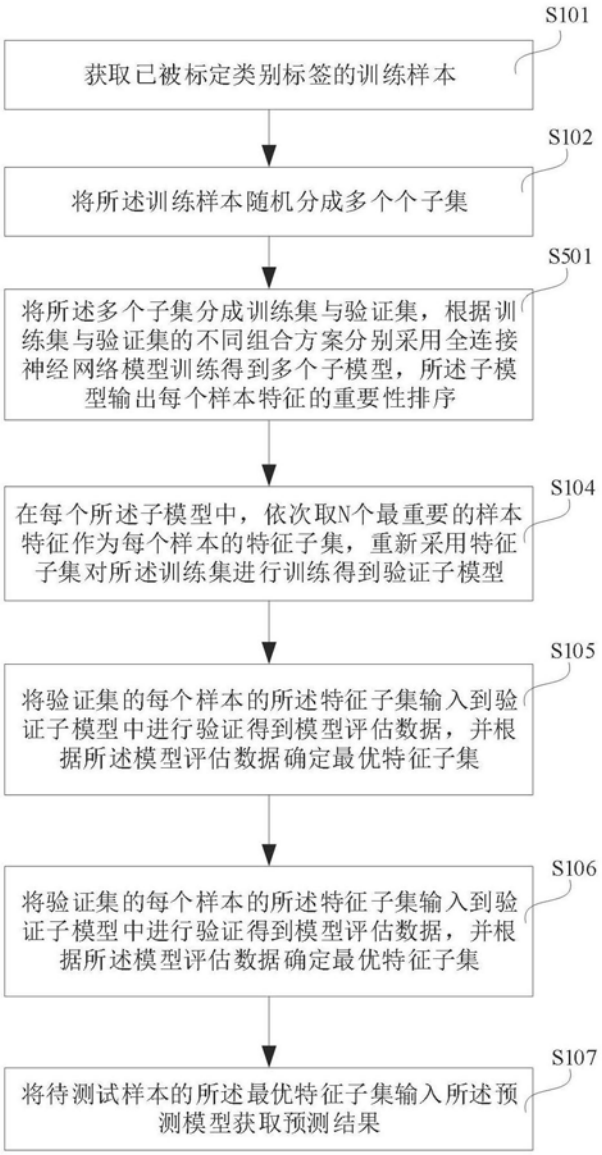

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S501
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用全连接神经网络模型训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图7

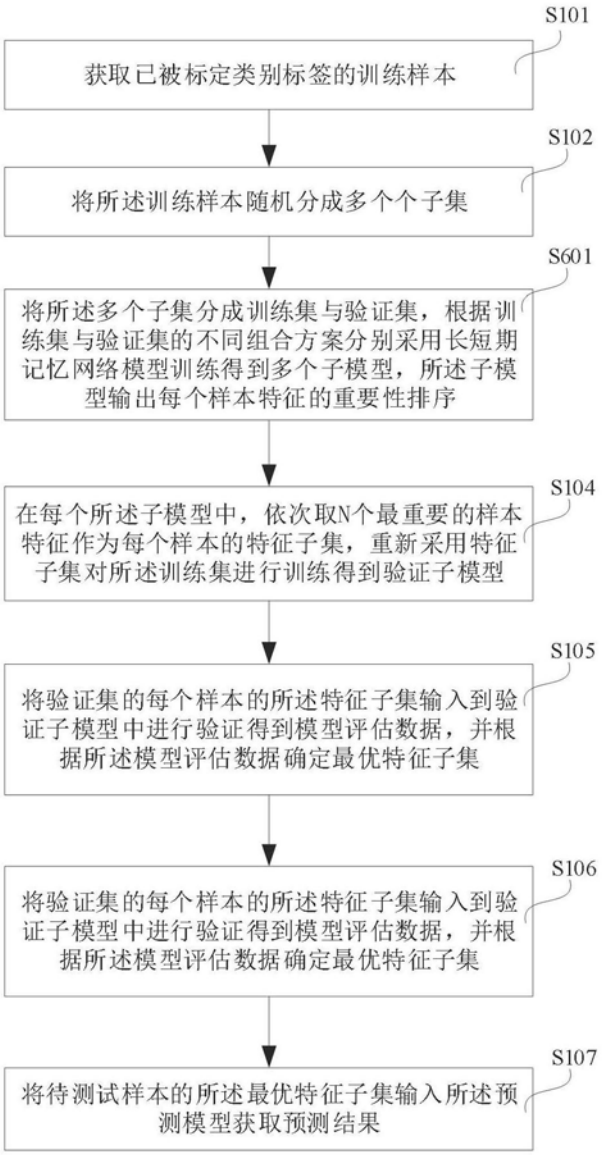

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S601
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用长短期记忆网络模型训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图8

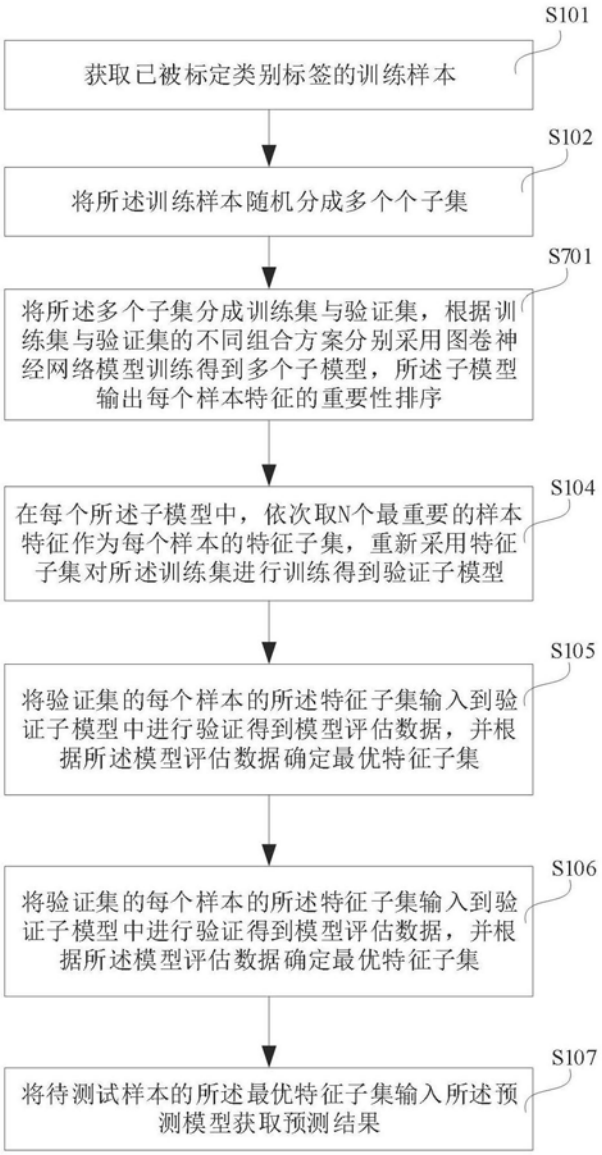

S101
获取已被标定类别标签的训练样本
S102
将所述训练样本随机分成多个个子集
S701
将所述多个子集分成训练集与验证集，根据训练集与验证集的不同组合方案分别采用图卷神经网络模型训练得到多个子模型，所述子模型输出每个样本特征的重要性排序
S104
在每个所述子模型中，依次取N个最重要的样本特征作为每个样本的特征子集，重新采用特征子集对所述训练集进行训练得到验证子模型
S105
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S106
将验证集的每个样本的所述特征子集输入到验证子模型中进行验证得到模型评估数据，并根据所述模型评估数据确定最优特征子集
S107
将待测试样本的所述最优特征子集输入所述预测模型获取预测结果

图9

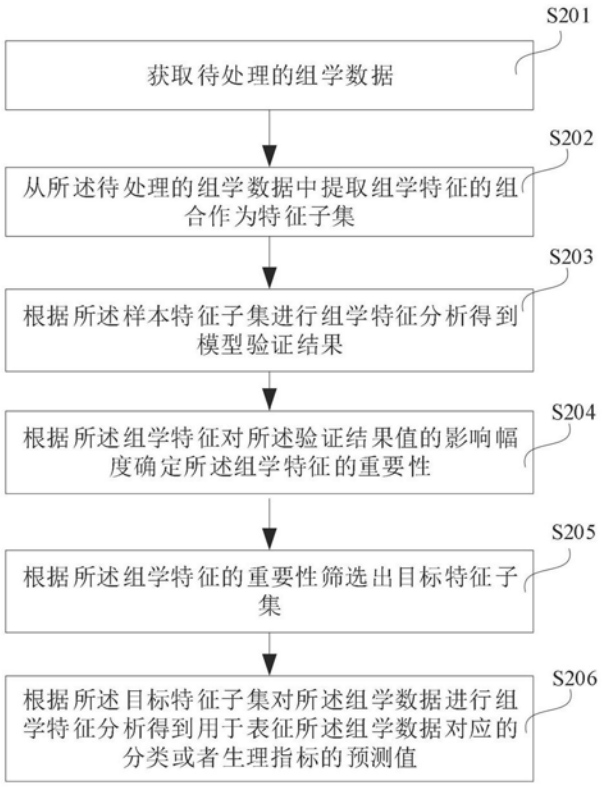

S201
获取待处理的组学数据
S202
从所述待处理的组学数据中提取组学特征的组合作为特征子集
S203
根据所述样本特征子集进行组学特征分析得到模型验证结果
S204
根据所述组学特征对所述验证结果值的影响幅度确定所述组学特征的重要性
S205
根据所述组学特征的重要性筛选出目标特征子集
S206
根据所述目标特征子集对所述组学数据进行组学特征分析得到用于表征所述组学数据对应的分类或者生理指标的预测值

图10

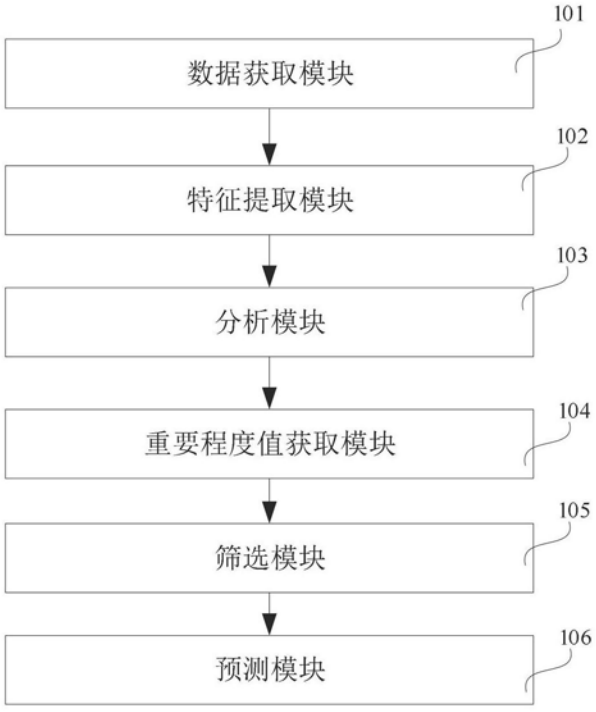

101
数据获取模块
102
特征提取模块
103
分析模块
104
重要程度值获取模块
105
筛选模块
106
预测模块

图11

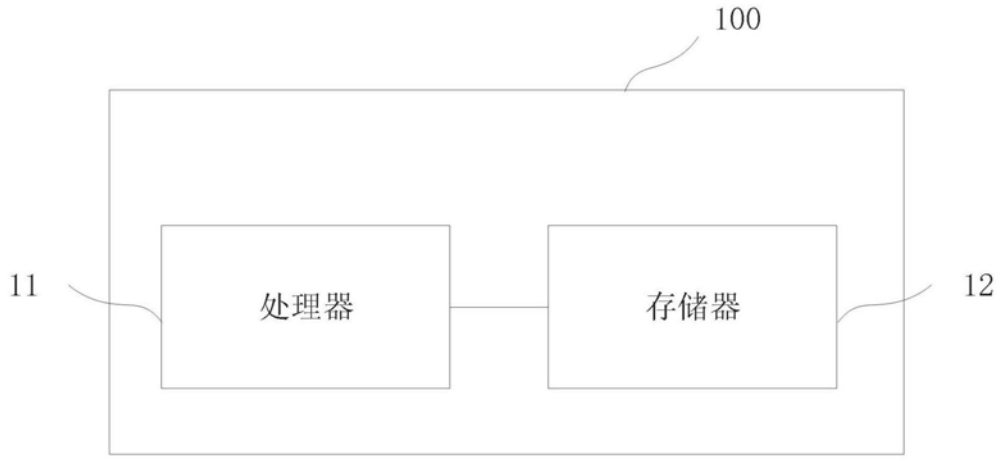


图12

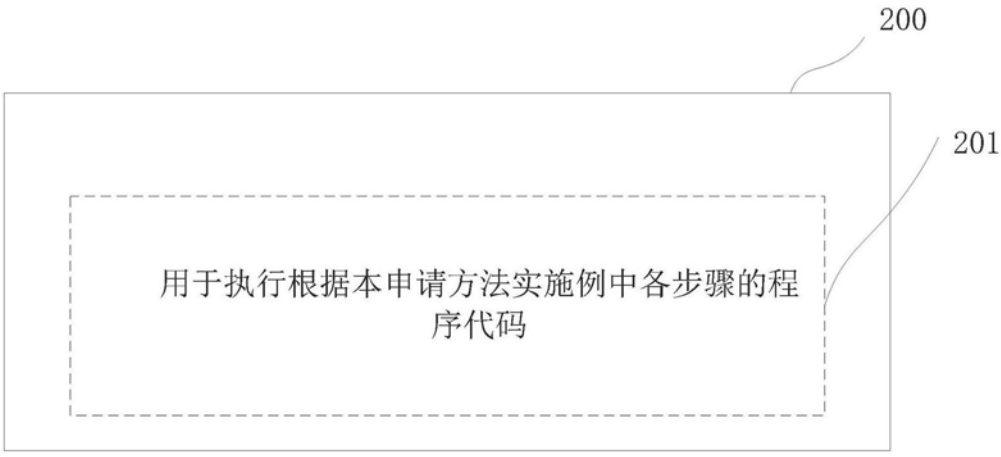


图13