(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111913999 B
(45)授权公告日 2024.05.28

(21)申请号 202010513704.4

(22)申请日 2020.06.08

(65)同一申请的已公布的文献号
申请公布号 CN 111913999 A

(43)申请公布日 2020.11.10

(73)专利权人 华南理工大学
地址 510641 广东省广州市天河区五山路381号

(72)发明人 吴兰兰 刘飞

(74)专利代理机构 广州嘉权专利商标事务所有限公司 44205
专利代理师 胡辉

(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 18/23(2023.01)
G06F 18/25(2023.01)
G16B 40/00(2019.01)
G16B 50/00(2019.01)
G16H 50/70(2018.01)

(56)对比文件
CN 107292127 A,2017.10.24
CN 109859796 A,2019.06.07
CN 110379460 A,2019.10.25
KR 101880686 B1,2018.07.20
KR 102042242 B1,2019.11.07
US 2016267235 A1,2016.09.15
US 2017024529 A1,2017.01.26
邹涵 等.胶质瘤患者的生存风险预测模型.国际神经病学神经外科学杂志.2019,第46卷(第1期),1-6.
A. Rose Brannon et al.Molecular Stratification of Clear Cell Renal Cell Carcinoma by Consensus Clustering Reveals Distinct Subtypes and Survival Patterns.Genes & Cancer.2010,第1卷(第2期),152- 163.
(续)

审查员 林粤美

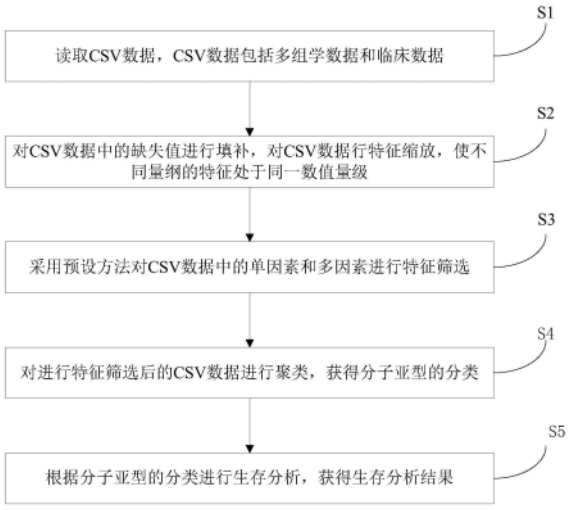
权利要求书2页 说明书7页 附图5页

(54)发明名称

基于多组学与临床数据的统计分析方法、系统和存储介质

(57)摘要

本发明公开了一种基于多组学与临床数据的统计分析方法、系统和存储介质,其中方法包括以下步骤:读取CSV数据;对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级;采用预设方法对所述CSV数据中的单因素和多因素进行特征筛选;对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;根据所述分子亚型的分类进行生存分析,获得生存分析结果。本发明通过对数据预处理、特征筛选、将筛选后的特征进行聚类、将聚类后的结果作为数据的真实标签进行生存分析,实现一个完备的多组学数据融合流程,从而实现肿瘤患者全面的分子分型,可广泛应用于生物信息学的多组学信息领域。



CN 111913999 B

[转续页]

[接上页]

(56)对比文件

Peter J Castaldi et al.Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema.Chronic obstructive pulmonary disease.2014,1-8.

邹涵;王苟思义;叶宁荣;李闫文;黄琦;刘宏伟;熊祖剑;李学军.胶质瘤患者的生存风险预测模型.国际神经病学神经外科学杂志.2019,(01),

陈干霞 等.随机生存森林在大规模基因分型肺癌预后关联性研究中的降维作用.中华疾病控制杂志.2012,第16卷(第7期),621-624.

1.一种基于多组学与临床数据的统计分析方法,其特征在于,包括以下步骤:

读取CSV数据,所述CSV数据包括多组学数据和临床数据;

对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级;

采用预设方法对所述CSV数据中的单因素和多因素进行特征筛选;

对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;

根据所述分子亚型的分类进行生存分析,获得生存分析结果;

所述CSV数据的数据格式为:第一列为time名的标签,第二列为status标签,其它列为特征;

所述采用预设方法对所述CSV数据中的单因素特征筛选,包括:

根据Correlation方法获取所述特征之间的第一相关系数,根据所述第一相关系数和第一预设阈值对所述特征进行筛选;

根据单因素Cox回归方法获取所述特征与生存时间及生存状态之间的第二相关系数,根据所述第二相关系数和第二预设阈值对所述特征进行筛选;

根据logrank test方法获取所述特征与和生存状态之间的第三相关系数,根据所述第三相关系数和第三预设阈值对所述特征进行筛选;

所述对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类,包括:

对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;

采用AMI方法对带有标签的所述CSV数据寻找最好的聚类数;

采用silhouette方法对不带有标签的所述CSV数据寻找最好的聚类数。

2.根据权利要求1所述的一种基于多组学与临床数据的统计分析方法,其特征在于,所述读取CSV数据,包括:

读取需要分析的所述CSV数据;

通过设置随机种子点,从所述CSV数据中获取训练集和测试集。

3.根据权利要求1所述的一种基于多组学与临床数据的统计分析方法,其特征在于,所述对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级,包括:

采用均值和中位数两种方式对所述CSV数据中的缺失值进行填补;

采用标准化和归一化两种方式对缺失值填补后的所述CSV数据进行特征缩放,使不同量纲的特征处于同一数值量级。

4.根据权利要求1所述的一种基于多组学与临床数据的统计分析方法,其特征在于,所述生存分析结果包括生存分析图、logrank test检验得到的p值、风险率HR和置信区间CI以及每个类别对应的时间生存人数。

5.一种基于多组学与临床数据的统计分析系统,其特征在于,包括:

数据读取模块,用于读取CSV数据,所述CSV数据包括多组学数据和临床数据;

数据预处理模块,用于对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级;

数据降维模块,用于采用预设方法对所述CSV数据中的单因素和多因素进行特征筛选;

数据聚类模块,用于对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分

类;

生存分析模块,用于根据所述分子亚型的分类进行生存分析,获得生存分析结果;

所述CSV数据的数据格式为:第一列为time名的标签,第二列为status标签,其它列为特征;

所述采用预设方法对所述CSV数据中的单因素特征筛选,包括:

根据Correlation方法获取所述特征之间的第一相关系数,根据所述第一相关系数和第一预设阈值对所述特征进行筛选;

根据单因素Cox回归方法获取所述特征与生存时间及生存状态之间的第二相关系数,根据所述第二相关系数和第二预设阈值对所述特征进行筛选;

根据logrank test方法获取所述特征与和生存状态之间的第三相关系数,根据所述第三相关系数和第三预设阈值对所述特征进行筛选;

所述对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类,包括:

对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;

采用AMI方法对带有标签的所述CSV数据寻找最好的聚类数;

采用silhouette方法对不带有标签的所述CSV数据寻找最好的聚类数。

6.一种基于多组学与临床数据的统计分析系统,其特征在于,包括:

至少一个处理器;

至少一个存储器,用于存储至少一个程序;

当所述至少一个程序被所述至少一个处理器执行,使得所述至少一个处理器实现权利要求1-4任一项所述的一种基于多组学与临床数据的统计分析方法。

7.一种存储介质,其中存储有处理器可执行的指令,其特征在于,所述处理器可执行的指令在由处理器执行时用于执行如权利要求1-4任一项所述一种基于多组学与临床数据的统计分析方法。

基于多组学与临床数据的统计分析方法、系统和存储介质

技术领域

[0001] 本发明涉及生物信息学的多组学信息领域,尤其涉及一种基于多组学与临床数据的统计分析方法、系统和存储介质。

背景技术

[0002] 整合定义为不同的组学数据(多组学)结合的过程,其中,组学数据包括基因组学定义的突变,转录组学确定的mRNA水平,蛋白质组学确定的蛋白质丰度和类型,以及表观基因组确定的甲基化图谱,代谢组学确定的代谢产物水平,元数据(如临床结果),组织学概况以及一系列数字成像分析等许多方面,相较于单个孤立的组学,整合的数据可提供具有更高信息能力的全局图。组学数据整合最前沿的领域之一是癌症生物学:在这里,数据整合的实施允许进行例如肿瘤分类以及随后的侵袭性和结果预测,从而支持选择个性化治疗。

[0003] 而在癌症生物学中,癌症基因组数据整合的相关研究最值得关注,因为癌症基因组具有大量的躯体获得性畸变。这些畸变会导致基因组不稳定,DNA中的基因组失调和扩增,mRNA的过表达以及基因产物的改变。此外,癌症基因组的异质性表现出不同个体之间在不同癌症类型中发生的各种生物学过程和表型。

[0004] 近年来,许多大规模的癌症基因组项目逐渐聚集了各种各样的基因组规模的数据,以便使用高分辨率的下一代测序平台发现生物学和生物医学知识。例如,癌症基因组图谱(TCGA)项目揭示了近10,000个癌症患者样品的大量表观基因组,基因组和转录组信息的遗传图景。因此,这是一个很好的机会,可以充分利用这些基因组规模的数据来发现可能的癌症亚型,从而推进诊断,预后和治疗策略。但是,融合大量基因组数据源有两个挑战。首先,高维数据通常具有不可避免的特征,即样本量远小于基因量。因此,很难使用常规的确定性方法来分析这些数据集。

[0005] 幸运的是,在快速发展的实验技术创新的推动下,目前已经出现了一系列的计算框架和算法,例如,著名的方法iCluster+,它通过对不同数据平台中x和w之间关系的不同建模方法进行假设,扩展了iCluster。它允许不同的数据类型,包括二进制、连续、分类和顺序数据与不同的建模假设,包括逻辑斯谛、正态线性、多对数和泊松分布;贝叶斯一致性聚类(BCC)方法是一种灵活的聚类方法,能够同时对各种数据源的相关性和异构性进行建模。它允许来自每个数据源的对象进行单独的集群,并对单独的集群进行事后集成。一致性聚类用于对源特定的结构建模以及确定总体聚类;相似网络融合(SNF)方法的目的是发现患者亚群。SNF通过为每种数据类型构建一个样本网络(而不是基因组特征)来集成不同的数据类型,然后将这些网络融合成一个综合网络。它有两个主要的数据集成步骤:首先,它为每个数据类型构造一个逐个样本的相似度矩阵,作为一个单独的网络。相似矩阵有助于识别通用的集群和网络。它还检测支持网络中每个连接的不同类型的数据。然后,利用消息传递理论的非线性方法(KNN和图扩散),SNF融合不同的相似矩阵和网络,使组合网络在每次迭代过程中更加连贯;Lemon-Tree是另一种非监督的模块网络重建方法。在从表达式数据矩阵中找到共同表达的集群之后,Lemon-Tree通过集成方法帮助识别一致模块和上游调控

程序。首先利用基因表达矩阵,通过基于模型的吉布斯采样器推断共表达基因簇。通过谱边聚类算法将共表达基因的一致性模块与基因聚类结果进行融合。另一方面,额外的候选调节器类型的数据,如miRNA表达,CNV和甲基化数据与一致性模块结合,以推断由决策树结构计算的调节得分。上述模块学习和调节器分配步骤的分离提供了更多的灵活性,允许与其他方法结合;特征选择多内核学习(FSMKL)是实现多内核学习监督式学习的另一种方法。这种新的方案使用统计得分为特征选择每个数据类型每条通路。通过引入基于临床协变量的额外核函数,提高了癌症检测的预测精度。多核学习使用基于路径的核来构造决策函数依赖于各种不同类型的输入数据(基因表达和CNV)的分类器。每种类型的数据(组学)被封装到一个称为基本内核的对象中;一个复合内核被构建为这些基本内核的线性组合。为了进一步将生物信息整合到该算法中,不仅单个特征(如基因)被独立地用于构造核函数,而且还将已知具有KEGG通路成员关系的特定基因组合在一起,得到其他碱基核函数。在特征选择步骤之后,确定最适合于内核的决策函数,形成基本内核的综合决策函数。该方法以基于路径的信息构建核为先验知识,在其他基于核的方法中脱颖而出。路径隶属度是FSMKL将样本分组为不同类群的中心标准,与其他方法的基本统计前提相比,它带来了更多的生物学知识。结合临床因素和高通量特征到分类器也带来了预测准确性的能力。

[0006] 虽然这些已有的方法针对不同的目标都取得了一定的成效,但是这些现有方法通常是以程序包的形式展现出来,其方法很少结合临床数据,并且专注于解决特定问题,方法固定不能供使用者选择。这些局限性对于前线的非编程的医务人员和科学家来讲是一个巨大的挑战。

发明内容

[0007] 为了解决上述技术问题,本发明的目的是提供一种基于多组学与临床数据的统计分析方法、系统和存储介质。

[0008] 本发明所采用的技术方案是:

[0009] 一种基于多组学与临床数据的统计分析方法,包括以下步骤:

[0010] 读取CSV数据,所述CSV数据包括多组学数据和临床数据;

[0011] 对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级;

[0012] 采用预设方法对所述CSV数据中的单因素和多因素进行特征筛选;

[0013] 对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;

[0014] 根据所述分子亚型的分类进行生存分析,获得生存分析结果。

[0015] 进一步,所述CSV数据的数据格式为:第一列为time名的标签,第二列为status标签,其它列为特征。

[0016] 进一步,所述读取CSV数据,包括:

[0017] 读取需要分析的所述CSV数据;

[0018] 通过设置随机种子点,从所述CSV数据中获取训练集和测试集。

[0019] 进一步,所述对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级,包括:

[0020] 采用均值和中位数两种方式对所述CSV数据中的缺失值进行填补;

[0021] 采用标准化和归一化两种方式对缺失值填补后的所述CSV数据进行特征缩放,使不同量纲的特征处于同一数值量级。

[0022] 进一步,所述采用预设方法对所述CSV数据中的单因素特征筛选,包括:

[0023] 根据Correlation方法获取所述特征之间的第一相关系数,根据所述第一相关系数和第一预设阈值对所述特征进行筛选;

[0024] 根据单因素Cox回归方法获取所述特征与生存时间及生存状态之间的第二相关系数,根据所述第二相关系数和第二预设阈值对所述特征进行筛选;

[0025] 根据logrank test方法获取所述特征与和生存状态之间的第三相关系数,根据所述第三相关系数和第三预设阈值对所述特征进行筛选。

[0026] 进一步,所述对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类,包括:

[0027] 对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;

[0028] 采用AMI方法对带有标签的所述CSV数据寻找最好的聚类数;

[0029] 采用silhouette方法对不带有标签的所述CSV数据寻找最好的聚类数。

[0030] 进一步,所述生存分析结果包括生存分析图、logrank test检验得到的p值、风险率HR和置信区间CI以及每个类别对应的时间生存人数。

[0031] 本发明所采用的另一技术方案是:

[0032] 一种基于多组学与临床数据的统计分析系统,包括:

[0033] 数据读取模块,用于读取CSV数据,所述CSV数据包括多组学数据和临床数据;

[0034] 数据预处理模块,用于对所述CSV数据中的缺失值进行填补,对所述CSV数据行特征缩放,使不同量纲的特征处于同一数值量级;

[0035] 数据降维模块,用于采用预设方法对所述CSV数据中的单因素和多因素进行特征筛选;

[0036] 数据聚类模块,用于对进行特征筛选后的所述CSV数据进行聚类,获得分子亚型的分类;

[0037] 生存分析模块,用于根据所述分子亚型的分类进行生存分析,获得生存分析结果。

[0038] 本发明所采用的另一技术方案是:

[0039] 一种基于多组学与临床数据的统计分析系统,包括:

[0040] 至少一个处理器;

[0041] 至少一个存储器,用于存储至少一个程序;

[0042] 当所述至少一个程序被所述至少一个处理器执行,使得所述至少一个处理器实现上所述方法。

[0043] 本发明所采用的另一技术方案是:

[0044] 一种存储介质,其中存储有处理器可执行的指令,所述处理器可执行的指令在由处理器执行时用于执行如上所述方法。

[0045] 本发明的有益效果是:本发明通过对数据预处理、特征筛选、将筛选后的特征进行聚类、将聚类后的结果作为数据的真实标签进行生存分析,实现一个完备的多组学数据融合流程,从而实现肿瘤患者全面的分子分型,能够促进精准医学的发展,拓宽在生物标志物发现方面的视野。

附图说明

[0046] 图1是本发明实施例一种基于多组学与临床数据的统计分析方法的步骤流程图；

[0047] 图2是本发明实施例中Cox模型训练样本相关系数图；

[0048] 图3是本发明实施例中Cox模型测试样本相关系数图；

[0049] 图4是本发明实施例中Lasso模型均方误差随惩罚系数的变化图；

[0050] 图5是本发明实施例中Lasso模型特征回归系数随惩罚系数的变化图；

[0051] 图6是本发明实施例中聚类方法寻找最好聚类数的结果示意图；

[0052] 图7是本发明实施例中系统选择聚类数进行生存分析的结果图；

[0053] 图8是本发明实施例用户选择聚类数进行生存分析的结果图；

[0054] 图9是本发明实施例一种基于多组学与临床数据的统计分析系统的结构示意图。

具体实施方式

[0055] 下面详细描述本发明的实施例,所述实施例的示例在附图中示出,其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的,仅用于解释本发明,而不能理解为对本发明的限制。

[0056] 在本发明的描述中,需要理解的是,涉及到方位描述,例如上、下、前、后、左、右等指示的方位或位置关系为基于附图所示的方位或位置关系,仅是为了便于描述本发明和简化描述,而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作,因此不能理解为对本发明的限制。

[0057] 在本发明的描述中,若干的含义是一个或者多个,多个的含义是两个以上,大于、小于、超过等理解为不包括本数,以上、以下、以内等理解为包括本数。如果有描述到第一、第二只是用于区分技术特征为目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量或者隐含指明所指示的技术特征的先后关系。

[0058] 本发明的描述中,除非另有明确的限定,设置、安装、连接等词语应做广义理解,所属技术领域技术人员可以结合技术方案的具体内容合理确定上述词语在本发明中的具体含义。

[0059] 如图1所示,本实施例提供了一种基于多组学与临床数据的统计分析方法,为技术与非技术人员研究生物信息学提供技术支持,包括但不限于以下步骤:

[0060] S1、读取CSV数据,CSV数据包括多组学数据和临床数据。

[0061] 本实施例中,以胶质瘤为例将其多组学数据(基因表达,甲基化表达和miRNA表达)和临床数据进行整合,数据格式为:第一列为time名的标签,第二列为status标签,其它列为特征,并设置随机种子点为1从而使分割的测试及训练集的数据,包括在后面的模型训练的过程中在设置相同种子点的情况下每次运行的结果相同。后续选择设置分割训练测试数据集的大小0.70,即将数据集分割为70％的训练集,30％的测试集。

[0062] S2、对CSV数据中的缺失值进行填补,对CSV数据行特征缩放,使不同量纲的特征处于同一数值量级。

[0063] 先对CSV数据中的异常值进行清除,并通过均值和中位数两种方式对缺失值进行填补,包括使用每列的特征值的均值对NaN数据进行填补、使用每列的特征值的中位数对NaN数据进行填补。接着通过两种方式对缺失值填补后的数据进行特征缩放,使不同量纲的

特征处于同一数值量级,减少方差大的特征的影响,使后面的模型更准确,该方法包括Standardization和MinMaxScaler,Standardization方法是将特征缩放到均值为0,方差为1的范围,MinMaxScaler是将特征缩放到0和1之间的范围。其中,Standardization为:标准化,标准化后会使每个特征中的数值平均变为0,将每个特征的值都减掉原始资料中该特征的平均,标准差变为1。MinMaxScaler为:归一化,把有量纲表达式变成无量纲表达式,便于不同单位或量级的指标能够进行比较和加权。归一化是一种简化计算的方式,即将有量纲的表达式,经过变换,化为无量纲的表达式,成为纯量。

[0064] 在本实施例中,先将胶质瘤的多组学数据异常值去除,并通过使用每列的特征值的均值对NaN数据进行填补,接着通过Standardization对缺失值填补后的数据进行特征缩放,即将特征缩放到均值为0,方差为1的范围,使不同量纲的特征处于同一数值量级,减少方差大的特征的影响,使后面的模型更准确。

[0065] S3、采用预设方法对CSV数据中的单因素和多因素进行特征筛选。

[0066] 对于单因素分析提供Correlation方法、单因素Cox回归方法和logrank test方法来进行特征筛选;对于多因素分析提供Cox回归方法和Lasso的特征选择方法进行特征筛选。其中,Correlation为:相关性。相关性分析是指对两个或多个具备相关性的变量元素进行分析,从而衡量两个变量因素的相关密切程度。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析。Cox为:比例风险回归模型,是一种半参数回归模型。该模型以生存结局和生存时间为应变量,可同时分析众多因素对生存期的影响,能分析带有截尾生存时间的资料,且不要求估计资料的生存分布类型。logrank test为:对数秩复检验,常用于临床百判断疗效,检验不同生存分析曲线之间的差异性是否显著。

[0067] 对于单因素分析,提供Correlation方法并设置相关性阈值来寻找特征与特征之间的关系将相关系数小于阈值的特征筛选出来。同时,也提供单因素Cox回归方法来寻找特征与生存时间和生存状态之间的关系将p值小于等于阈值具有统计学意义的的特征筛选出来。还提供logrank test方法来寻找特征与生存时间和生存状态之间的关系将p值小于等于阈值具有统计学意义的的特征筛选出来。

[0068] 对于多因素分析,提供Cox回归来进行特征筛选,并保存筛选后的特征或计算Cox风险值radscore。还提供Lasso的特征选择方法进行特征筛选,它是基于Cox模型的Lasso自适应方法。每次的特征筛选都会在上一次方法的基础上对剩下的特征进行再一次特征筛选,在每次的单因素或多因素特征筛选后会画出特征与特征之间的关系的系数图和样本特征的热力图,并在进行Lasso特征筛选时给出Lasso path图,可以看出各个回归系数随惩罚系数的变化、看出自变量退出模型的先后顺序,从而提供该方法的可视化结果。

[0069] 在特征筛选时,对胶质瘤的多组学数据选择单因素分析中的Cox回归方法来寻找特征与生存时间和生存状态之间的关系,将p值小于等于阈值具有统计学意义的特征筛选出来。在Cox回归方法特征筛选的基础上,对胶质瘤的多组学数据的多因素分析选择Lasso方法进行特征筛选,在每次的单因素或多因素特征筛选后会画出特征与特征之间的关系的系数图和样本特征的热力图,并在进行Lasso特征筛选时给出Lasso path图,可以看出各个回归系数随惩罚系数的变化、看出自变量退出模型的先后顺序,从而提供该方法的可视化结果,如图2-图5所示。其中,图2是本发明实施例中Cox模型训练样本相关系数图,图3是本发明实施例中Cox模型测试样本相关系数图,图4是本发明实施例中Lasso模型均方误差随

惩罚系数的变化图,图5是本发明实施例中Lasso模型特征回归系数随惩罚系数的变化图。

[0070] S4、对进行特征筛选后的CSV数据进行聚类,获得分子亚型的分类。

[0071] 对筛选后的特征提供K-Means聚类,从而获得分子亚型的分类。并针对有标签(癌症分型明确)的数据采用AMI方法寻找最好的聚类数,针对无标签的数据采用silhouette方法寻找最好的聚类数。由于胶质瘤的数据没有肿瘤分类标签,所有采用silhouette方法寻找最好的聚类数为3,如图6所示,同时也设置了类别数为4,从而得到两种分子亚型标签。

[0072] S5、根据分子亚型的分类进行生存分析,获得生存分析结果。

[0073] 针对S4步骤得到的两种结果分别做生存分析,得到生存分析图,logranktest检验得到的p值,每组的风险率HR和置信区间CI,每个类别对应的时间生存人数。其得到的p值均小于等于0.0001,说明其每组之间的差异显著,效果好。但二者相比较,通过silhouette方法寻找的聚类数的效果更好,其每组分类之间没有交叉,具有显著性的差异,如图7-图8所示,其中,图7是本发明实施例中系统选择聚类数进行生存分析的结果图,图8是本发明实施例用户选择聚类数进行生存分析的结果图。

[0074] 综上所述,本实施例的方法,至少具有如下有益效果:1)为前线的非编程的医务人员和科学家在研究多组学数据方面提供了技术支持和帮助;2)该流程提供了多样化的特征选择方法,在使用时,可根据数据特点来选择适合自己的模型;3)该流程可以对比不同的聚类结果,并将生存分析效果好的k值作为亚型分类结果,从而为前线的医生科学家提供一个特定癌症的分子亚型的参考;4)该流程提供的特征筛选后的结果可为研究人员提供一些与肿瘤分期密切相关的基因作为组学与临床表型之间联系的参考,从而有助于建立个性化的癌症治疗计划。

[0075] 如图9所示,本实施例还提供了一种基于多组学与临床数据的统计分析系统,包括:

[0076] 数据读取模块,用于读取CSV数据,CSV数据包括多组学数据和临床数据;

[0077] 数据预处理模块,用于对CSV数据中的缺失值进行填补,对CSV数据行特征缩放,使不同量纲的特征处于同一数值量级;

[0078] 数据降维模块,用于采用预设方法对CSV数据中的单因素和多因素进行特征筛选;

[0079] 数据聚类模块,用于对进行特征筛选后的CSV数据进行聚类,获得分子亚型的分类;

[0080] 生存分析模块,用于根据分子亚型的分类进行生存分析,获得生存分析结果。

[0081] 本实施例的一种基于多组学与临床数据的统计分析系统,可执行本发明方法实施例所提供的一种基于多组学与临床数据的统计分析系统方法,可执行方法实施例的任意组合实施步骤,具备该方法相应的功能和有益效果。

[0082] 本实施例还提供了一种基于多组学与临床数据的统计分析系统,包括:

[0083] 至少一个处理器;

[0084] 至少一个存储器,用于存储至少一个程序;

[0085] 当至少一个程序被至少一个处理器执行,使得至少一个处理器实现上所述方法。

[0086] 本实施例的一种基于多组学与临床数据的统计分析系统,可执行本发明方法实施例所提供的一种基于多组学与临床数据的统计分析系统方法,可执行方法实施例的任意组合实施步骤,具备该方法相应的功能和有益效果。

[0087] 本实施例还提供了一种存储介质,其中存储有处理器可执行的指令,处理器可执行的指令在由处理器执行时用于执行如上所述方法。

[0088] 本实施例的一种存储介质,可执行本发明方法实施例所提供的一种基于多组学与临床数据的统计分析系统方法,可执行方法实施例的任意组合实施步骤,具备该方法相应的功能和有益效果。

[0089] 可以理解的是,上文中所公开方法中的全部或某些步骤、系统可以被实施为软件、固件、硬件及其适当的组合。某些物理组件或所有物理组件可以被实施为由处理器,如中央处理器、数字信号处理器或微处理器执行的软件,或者被实施为硬件,或者被实施为集成电路,如专用集成电路。这样的软件可以分布在计算机可读介质上,计算机可读介质可以包括计算机存储介质(或非暂时性介质)和通信介质(或暂时性介质)。如本领域普通技术人员公知的,术语计算机存储介质包括在用于存储信息(诸如计算机可读指令、数据结构、程序模块或其他数据)的任何方法或技术中实施的易失性和非易失性、可移除和不可移除介质。计算机存储介质包括但不限于RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光盘存储、磁盒、磁带、磁盘存储或其他磁存储装置、或者可以用于存储期望的信息并且可以被计算机访问的任何其他的介质。此外,本领域普通技术人员公知的是,通信介质通常包含计算机可读指令、数据结构、程序模块或者诸如载波或其他传输机制之类的调制数据信号中的其他数据,并且可包括任何信息递送介质。

[0090] 上面结合附图对本发明实施例作了详细说明,但是本发明不限于上述实施例,在所述技术领域普通技术人员所具备的知识范围内,还可以在不脱离本发明宗旨的前提下作出各种变化。

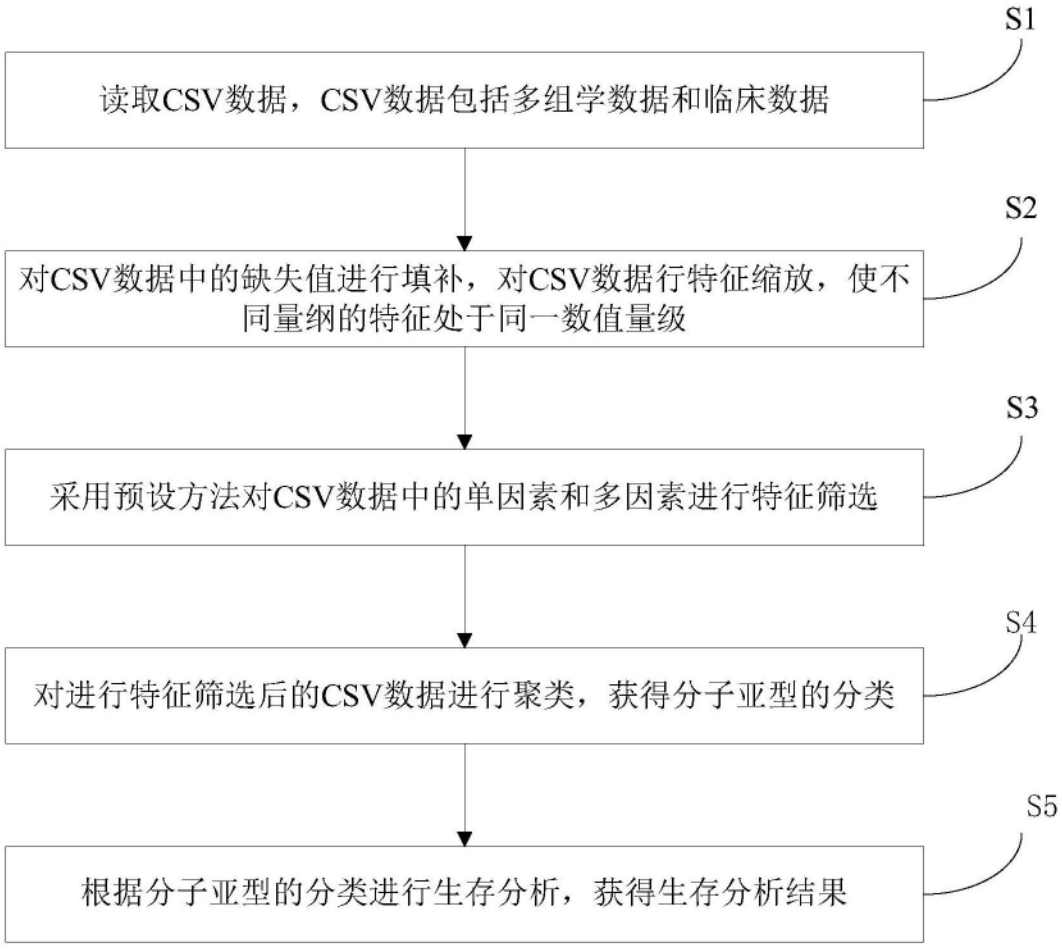

S1
读取CSV数据，CSV数据包括多组学数据和临床数据
S2
对CSV数据中的缺失值进行填补，对CSV数据行特征缩放，使不同量纲的特征处于同一数值量级
S3
采用预设方法对CSV数据中的单因素和多因素进行特征筛选
S4
对进行特征筛选后的CSV数据进行聚类，获得分子亚型的分类
S5
根据分子亚型的分类进行生存分析，获得生存分析结果

图1

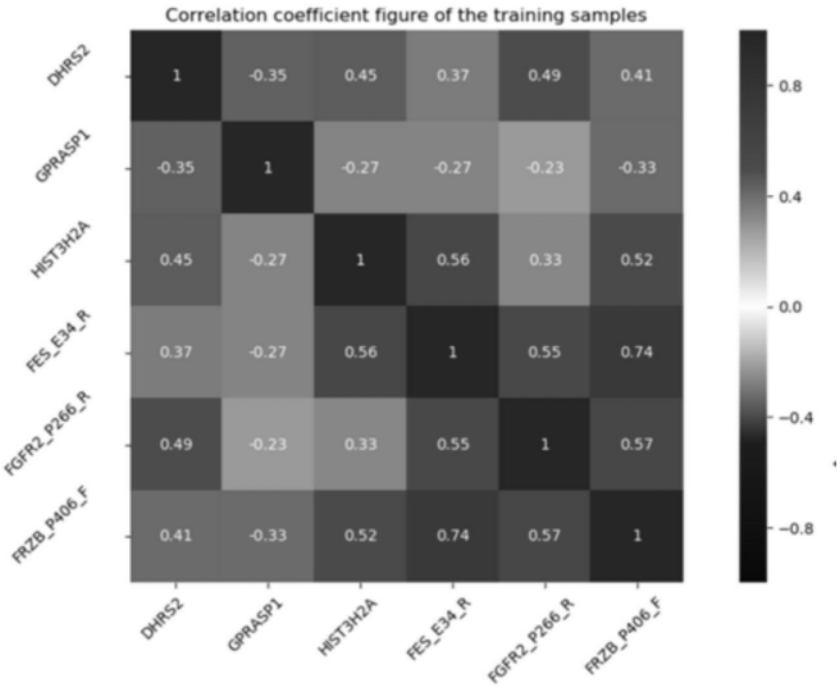

Correlation coefficient figure of the training samples
DHRS2
GPRASP1
HIST3H2A
FES_E34_R
FGFR2_P266_R
FRZB_P406_F
1 -0.35 0.45 0.37 0.49 0.41
-0.35 1 -0.27 -0.27 -0.23 -0.33
0.45 -0.27 1 0.56 0.33 0.52
0.37 -0.27 0.56 1 0.55 0.74
0.49 -0.23 0.33 0.55 1 0.57
0.41 -0.33 0.52 0.74 0.57 1
0.8
0.4
0.0
−0.4
−0.8

图2

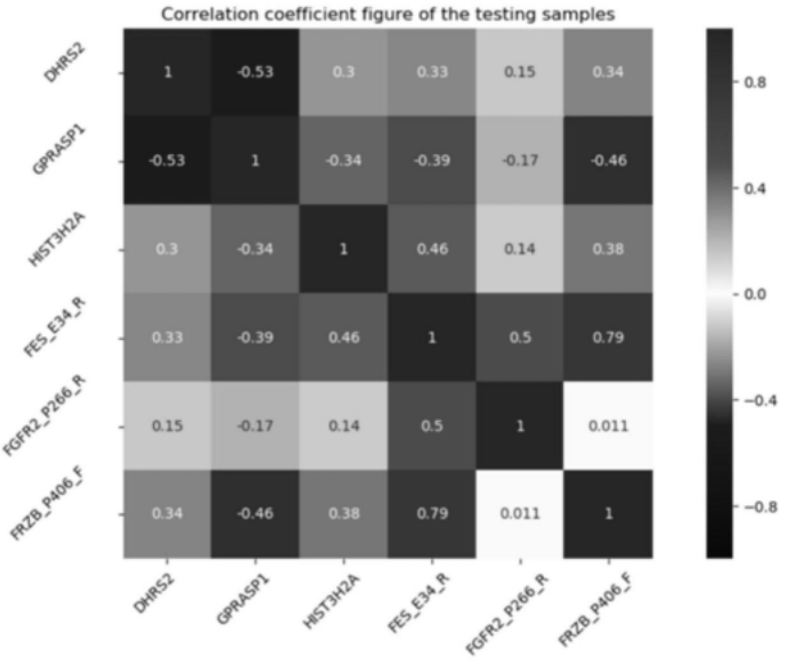

Correlation coefficient figure of the testing samples
DHRS2
GPRASP1
HIST3H2A
FES_E34_R
FGFR2_P266_R
FRZB_P406_F
1 -0.53 0.3 0.33 0.15 0.34
-0.53 1 -0.34 -0.39 -0.17 -0.46
0.3 -0.34 1 0.46 0.14 0.38
0.33 -0.39 0.46 1 0.5 0.79
0.15 -0.17 0.14 0.5 1 0.011
0.34 -0.46 0.38 0.79 0.011 1
0.8
0.4
0.0
−0.4
−0.8

图3

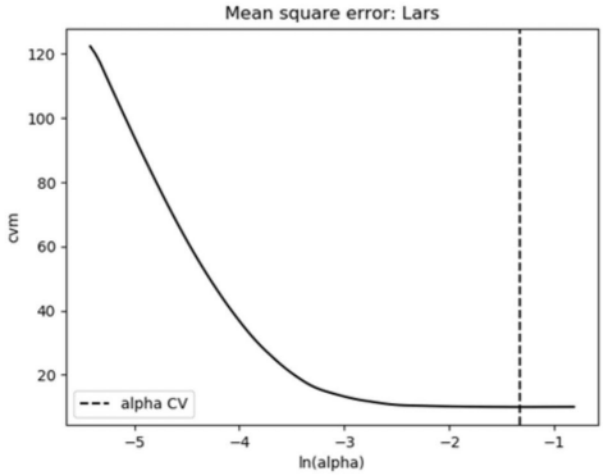

Mean square error: Lars
120
100
80
60
40
20
cvm
alpha CV
−5
−4
−3
−2
−1
ln(alpha)

图4

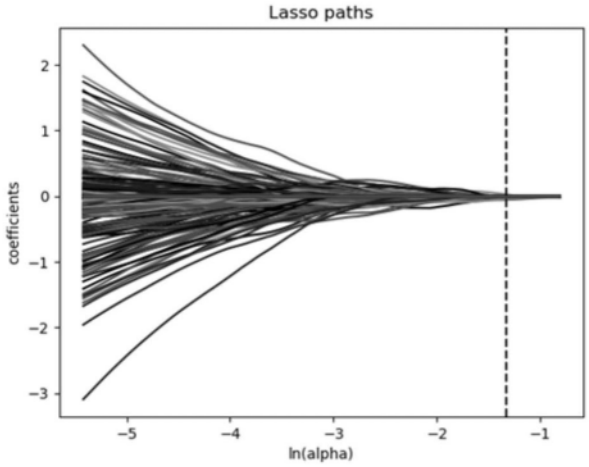

Lasso paths
2
1
0
−1
−2
−3
coefficients
−5
−4
−3
−2
−1
ln(alpha)

图5

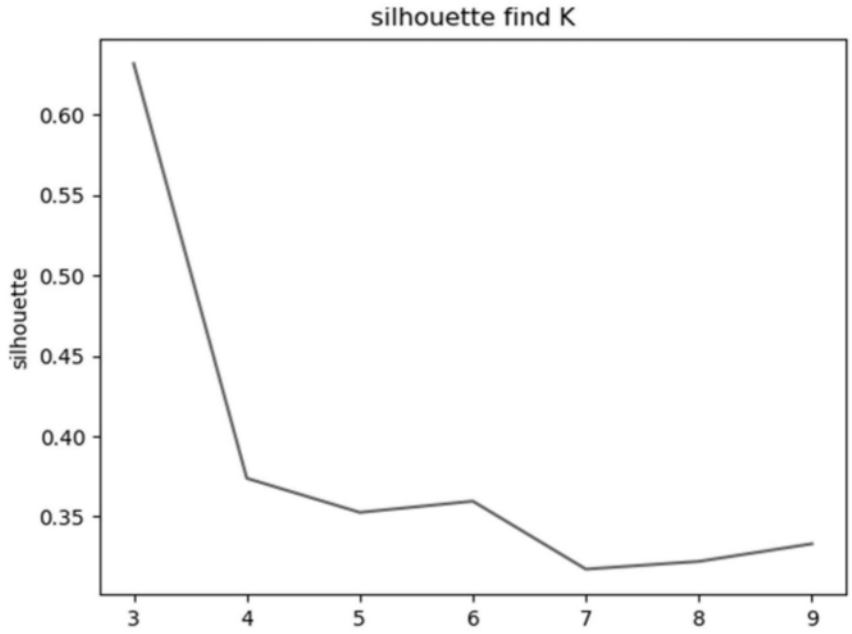

silhouette find K
silhouette
0.60
0.55
0.50
0.45
0.40
0.35
3
4
5
6
7
8
9

图6

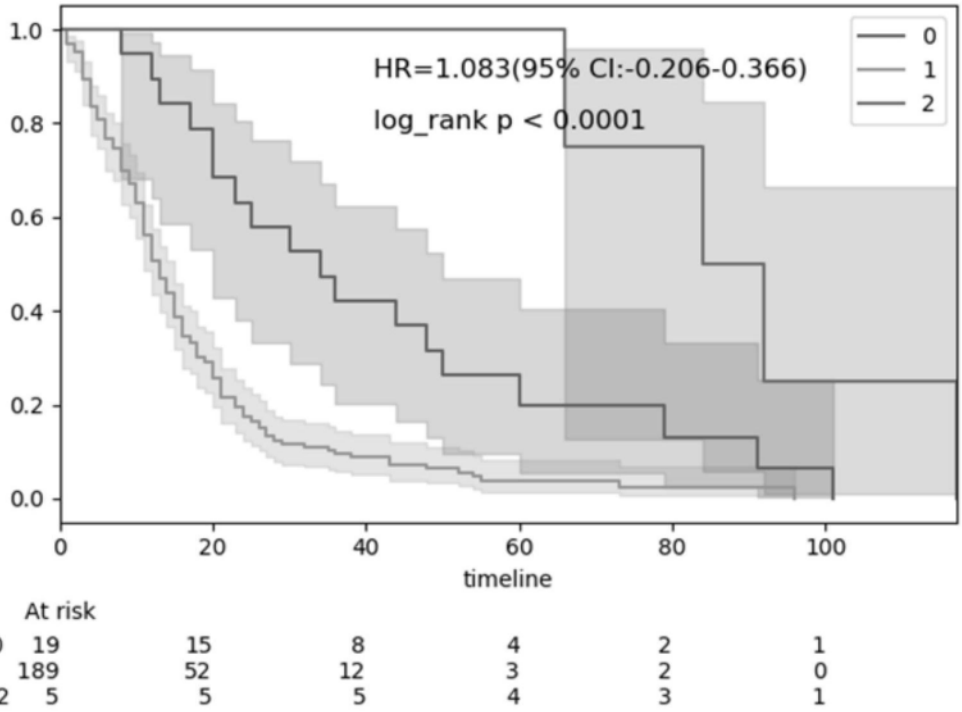

HR=1.083(95% CI:-0.206-0.366)
log_rank p < 0.0001
0
1
2
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
timeline
At risk
0 19 15 8 4 2 1
1 189 52 12 3 2 0
2 5 5 5 4 3 1

图7

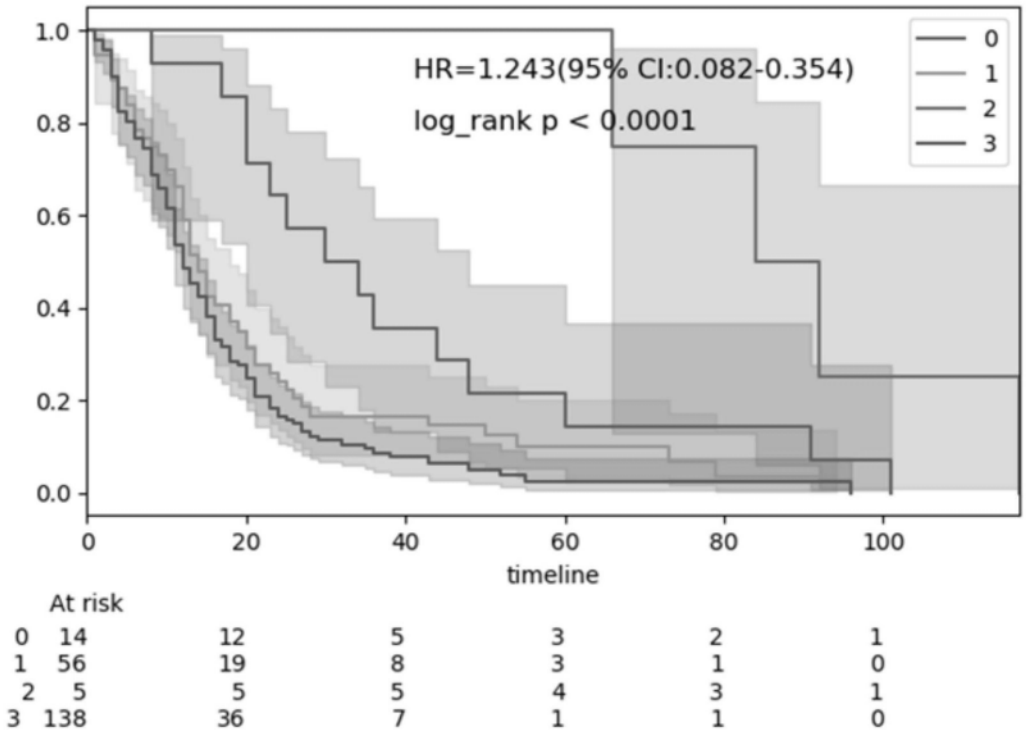

HR=1.243(95% CI:0.082-0.354)
log_rank p < 0.0001
0
1
2
3
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
timeline
At risk
0 14 12 5 3 2 1
1 56 19 8 3 1 0
2 5 5 5 4 3 1
3 138 36 7 1 1 0

图8


数据读取模块
数据预处理模块
数据降维模块
数据聚类模块
生存分析模块

图9