(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115565610 A
(43)申请公布日 2023.01.03

(21)申请号 202211198172.5

(22)申请日 2022.09.29

(71)申请人 四川大学
地址 610065 四川省成都市一环路南一段24号

(72)发明人 李冰 章乐 袁勇

(74)专利代理机构 北京创赋致远知识产权代理有限公司 11972
专利代理师 邱晓宁

(51)Int.Cl.
G16B 40/00(2019.01)
G16H 50/30(2018.01)

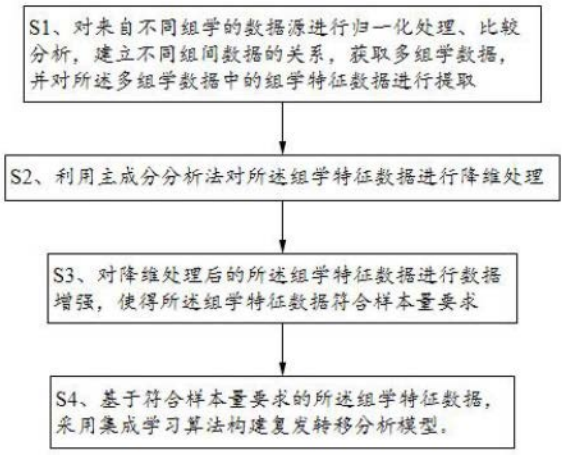
权利要求书2页 说明书9页 附图4页

(54)发明名称

基于多组学数据的复发转移分析模型建立方法及系统

(57)摘要

本发明提供了一种基于多组学数据的复发转移分析模型建立方法及系统。本发明的基于多组学数据的复发转移分析模型建立方法利用多组学数据信息，从多层次对复发转移的数据进行分析，使得对复发转移数据的分析更为全面有效，同时，通过对多组学数据进行系统化的组学特征数据的选择和降维处理，有效利用和筛选了不同组学的数据，对进行复发转移分析模型建立的多组学数据进行了质量控制，最终综合多种经典机器学习模型，提高了复发转移分析模型的准确性。



CN 115565610 A

1.一种基于多组学数据的复发转移分析模型建立方法,其特征在于,包括以下步骤:

S1、对来自不同组学的数据源进行归一化处理、比较分析,建立不同组间数据的关系,获取多组学数据,并对所述多组学数据中的组学特征数据进行提取;

S2、利用主成分分析法对所述组学特征数据进行降维处理;

S3、对降维处理后的所述组学特征数据进行数据增强,使得所述组学特征数据符合样本量要求;

S4、基于符合样本量要求的所述组学特征数据,采用集成学习算法构建复发转移分析模型;所述复发转移分析模型表示为:

$$log\left(\frac{H(x)}{1-H(x)}\right)=c_0+\sum_{m=1}^{M=3}c_mH_{mT}(x);$$

其中,H(x)是集成分类器;c_0为常数,c_m是每个弱分类器的权值;M是弱分类器的指标,M=1,2,3;H_{mT}是每个弱分类器的权重的同态积分。

2.根据权利要求1所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,当所述多组学数据为为离散数据时,通过Fisher精确检验法或卡方检验法来确定每组组学数据的数据特征与复发转移标签之间的相关性,对所述组学特征数据进行提取。

3.根据权利要求1所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,当所述多组学数据为为连续数据时,根据复发转移标签将所述多组学数据分为第一数据和第二数据,至少结合T检验法、Mann-Whitney U检验法以及方差分析法的假设检验结果的交集,获取所述组学特征数据。

4.根据权利要求1所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,所述组学特征数据至少包括10组。

5.根据权利要求4所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,所述S2中,所述主成分分析法具体为:

$T_L=XW_L$;

其中,W_L为将包含i个变量的原始组学数据X映射到数据集上包含i个不相关变量的新空间T,仅保留前L个主成分进行降维操作。

6.根据权利要求5所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,所述S3具体为:

S31、样本量的估计,计算每个选定的所述组学特征数据相对应于预设统计显著性的最佳样本量n,对所述组学特征数据进行数据扩充;

S32、对所述组学特征数据扩充后的伪数据集进行评估,通过最大Fisher判别比F1验证生成的所述组学特征数据是否可以用于分类,并评估伪数据集数据增强的质量。

7.根据权利要求6所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,所述S31中,所述最佳样本量n表示为:

$$n=\frac{\sigma^2\left(Q_1{}^{-1}+Q_2{}^{-1}\right)\left(\mu_\alpha+\mu_\beta\right)^2}{\delta^2}$$

其中,σ为标准差,μ_α和μ_β是第一类错误率和第二类错误率下μ检验的临界值,Q_1和Q_2为经过二分类后的群体中各部分的比例,δ为两组数据均值的差值。

8.根据权利要求6所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,所述S31中,当所述最佳样本量n不满足最佳样本量的要求时,使用SMOTE算法对所述组学特征数据进行数据增强,生成伪数据集来扩充所述组学特征数据以达到最佳样本量的要求。

9.根据权利要求1所述的基于多组学数据的复发转移分析模型建立方法,其特征在于,在所述S4中,所述复发转移分析模型为基于逻辑回归(LR)、支持向量机(SVM)和朴素贝叶斯(Naive-Bayes)三种分类方法构建获取。

10.一种基于多组学数据的复发转移分析模型建立系统,可用于执行所述权利要求1～9中任一项所述的基于多组学数据的复发转移分析模型建立方法;其特征在于,所述基于多组学数据的复发转移分析模型建立系统包括:

数据采集模块,所述数据采集模块用于获取及存储多组学数据;

数据处理模块,所述数据处理模块用于根据所述多组学数据中每组组学数据的数据特征与复发转移标签进行提取,获取组学特征数据;其中,所述复发转移标签用于标注是否复发;

中央处理器,包括主成分分析模块,根据$T_L=XW_L$对所述组学特征数据进行降维处理;其中,W_L为将包含i个变量的原始组学数据X映射到数据集上包含i个不相关变量的新空间T,仅保留前L个主成分进行降维操作;

所述中央处理器还包括数据增强模块,所述数据增强模块至少用于执行SMOTE算法,对所述组学特征数据进行数据增强,生成伪数据集来扩充所述组学特征数据以达到最佳样本量的要求所述中央处理器用于对所述组学特征数据进行多组学数据进行处理;以及

集成数据模块,所述集成数据模块被配置为根据至少三个弱分类器的模型进行复发转移分析模型的建立;所述复发转移分析模型为:

$$log\left(\frac{H(x)}{1-H(x)}\right)=c_0+\sum_{m=1}^{M=3}c_mH_{mT}(x);$$

其中,H(x)是集成分类器;c_0为常数,c_m是每个弱分类器的权值;M是弱分类器的指标,M=1,2,3;H_{mT}是每个弱分类器的权重的同态积分。

基于多组学数据的复发转移分析模型建立方法及系统

技术领域

[0001] 本发明涉及一种分析模型建立方法，尤其涉及一种基于多组学数据的复发转移分析模型建立方法及应用基于多组学数据的转移分析模型建立方法的系统。

背景技术

[0002] 最新研究中指出，利用多组学信息，即蛋白质组学和磷蛋白组学数据，成功且极有效地区分了是否复发转移。但是目前国内外的研究和模型无法有效对多组学数据进行特征筛选，无法有效利用多维度的数据构建综合的数据分析模型和系统。此外，当前的模型构建还存在分析结果准确率不高，模型构建方法较为单一等问题。

[0003] 具体的，从数据局限性的角度讲，现有预测模型的数据来源较为单一。大部分研究皆采用单一的影像放射学数据、基因数据和临床相关数据作为研究的数据来源。然而由于癌症等其它疾病的成因和发展复杂，如结直肠癌，相较于多组学数据，仅使用单一数据无法全面完整的阐释患者的状态并进行分析。

[0004] 从特征工程的角度讲，由于现有预测模型的数据来源较为单一，故而缺乏一个针对多组学数据的系统全面的特征选择和降维方法。其临床和基因表达特征选择大多由人工选择或单个统计检验方法完成，例如相关系数检验，卡方检验，T检验或Mann-Whitney U检验等。从而存在特征选择不客观，特征选择方法和结果不相同等等问题。除此之外，利用得到多组学数据的关键特征后，数据可能仍存在特征维数较高的情况，不利于模型构建，故而需要一个系统的针对多组学数据的特征选择和降维方法。

[0005] 进一步的，从模型构建的角度讲，现有的预测模型建模方法单一。在使用机器学习对术后转移预测进行模型构建时，以往的研究采用了比例风险回归模型(Cox模型)，logistics回归，决策树，随机森林等机器学习模型和算法进行分析。这些研究大多只使用了单个模型方法进行构建，没有使用集成学习等方法将各类机器学习的优势进行集成，由于不同模型的表现依赖于数据的选取，这些模型很难迁移或整合多组学数据，并且存在准确率不高的问题。

[0006] 有鉴于此，确有必要提出一种基于多组学数据的复发转移预测模型和系统，以解决上述问题。

发明内容

[0007] 本发明的目的在于提供一种基于多组学数据的复发转移分析模型建立方法及系统，本发明的基于多组学数据的复发转移分析模型建立方法利用多组学数据信息，从多层次对复发转移的数据进行分析。

[0008] 为实现上述发明目的，本发明提供了一种基于多组学数据的复发转移分析模型建立方法，包括以下步骤：

[0009] S1、对来自不同组学的数据源进行归一化处理、比较分析，建立不同组间数据的关系，获取多组学数据，并对所述多组学数据中的组学特征数据进行提取；

[0010] S2、利用主成分分析法对所述组学特征数据进行降维处理；

[0011] S3、对降维处理后的所述组学特征数据进行数据增强，使得所述组学特征数据符合样本量要求；

[0012] S4、基于符合样本量要求的所述组学特征数据，采用集成学习算法构建复发转移分析模型；所述复发转移分析模型表示为：

[0013] $$\log\left(\frac{H(x)}{1-H(x)}\right)=c_0+\sum_{m=1}^{M=3}c_m H_{mT}(x);$$

[0014] 其中，H(x)是集成分类器；c_0为常数，c_m是每个弱分类器的权值；M是弱分类器的指标，M＝1,2,3；H_{mT}是每个弱分类器的权重的同态积分。

[0015] 作为本发明的进一步改进，当所述多组学数据为为离散数据时，通过Fisher精确检验法或卡方检验法来确定每组组学数据的数据特征与复发转移标签之间的相关性，对所述组学特征数据进行提取。

[0016] 作为本发明的进一步改进，当所述多组学数据为为连续数据时，根据复发转移标签将所述多组学数据分为第一数据和第二数据，至少结合T检验法、Mann-Whitney U检验法以及方差分析法的假设检验结果的交集，获取所述组学特征数据。

[0017] 作为本发明的进一步改进，所述组学特征数据至少包括10组。

[0018] 作为本发明的进一步改进，所述S2中，所述主成分分析法具体为：

[0019] $T_L=XW_L$；

[0020] 其中，W_L为将包含i个变量的原始组学数据X映射到数据集上包含i个不相关变量的新空间T，仅保留前L个主成分进行降维操作。

[0021] 作为本发明的进一步改进，所述S3具体为：

[0022] S31、样本量的估计，计算每个选定的所述组学特征数据相对应于预设统计显著性的最佳样本量n，对所述组学特征数据进行数据扩充；

[0023] S32、对所述组学特征数据扩充后的伪数据集进行评估，通过最大Fisher判别比F1验证生成的所述组学特征数据是否可以用于分类，并评估伪数据集数据增强的质量。

[0024] 作为本发明的进一步改进，所述S31中，所述最佳样本量n表示为：

[0025] $$n=\frac{\sigma^2\left(Q_1^{-1}+Q_2^{-1}\right)\left(\mu_\alpha+\mu_\beta\right)^2}{\delta^2};$$

[0026] 其中，σ为标准差，μ_α和μ_β是第一类错误率和第二类错误率下μ检验的临界值，Q_1和Q_2为经过二分类后的群体中各部分的比例，δ为两组数据均值的差值。

[0027] 作为本发明的进一步改进，所述S31中，当所述最佳样本量n不满足最佳样本量的要求时，使用SMOTE算法对所述组学特征数据进行数据增强，生成伪数据集来扩充所述组学特征数据以达到最佳样本量的要求。

[0028] 作为本发明的进一步改进，在所述S4中，所述复发转移分析模型为基于逻辑回归(LR)、支持向量机(SVM)和朴素贝叶斯(Naive-Bayes)三种分类方法构建获取。

[0029] 为实现上述发明目的，本发明还提供了一种基于多组学数据的复发转移分析模型建立系统，可用于执行前述的基于多组学数据的复发转移分析模型建立方法；所述基于多组学数据的复发转移分析模型建立系统包括：数据采集模块，所述数据采集模块用于获取及存储多组学数据；数据处理模块，所述数据处理模块用于根据所述多组学数据中每组组

学数据的数据特征与复发转移标签进行提取，获取组学特征数据；其中，所述复发转移标签用于标注是否复发；中央处理器，包括主成分分析模块，根据$T_L=XW_L$对所述组学特征数据进行降维处理；其中，W_L为将包含i个变量的原始组学数据X映射到数据集上包含i个不相关变量的新空间T，仅保留前L个主成分进行降维操作；所述中央处理器还包括数据增强模块，所述数据增强模块至少用于执行SMOTE算法，对所述组学特征数据进行数据增强，生成伪数据集来扩充所述组学特征数据以达到最佳样本量的要求所述中央处理器用于对所述组学特征数据进行多组学数据进行处理；以及集成数据模块，所述集成数据模块被配置为根据至少三个弱分类器的模型进行复发转移分析模型的建立；所述复发转移分析模型为：

[0030] $$\log\left(\frac{H(x)}{1-H(x)}\right)=c_0+\sum_{m=1}^{M=3}c_m H_{mT}(x);$$

[0031] 其中，H(x)是集成分类器；c_0为常数，c_m是每个弱分类器的权值；M是弱分类器的指标，M＝1,2,3；H_{mT}是每个弱分类器的权重的同态积分。

[0032] 本发明的有益效果是：

[0033] 本发明的基于多组学数据的复发转移分析模型建立方法利用多组学数据信息，从多层次对复发转移的数据进行分析，使得对复发转移数据的分析更为全面有效，同时，通过对多组学数据进行系统化的组学特征数据的选择和降维处理，有效利用和筛选了不同组学的数据，对进行复发转移分析模型建立的多组学数据进行了质量控制，最终综合多种经典机器学习模型，提高了复发转移分析模型的准确性。

附图说明

[0034] 图1是本发明基于多组学数据的复发转移分析模型建立方法的流程图；

[0035] 图2是多组学数据为离散数据时，提取组学特征数据流程图；

[0036] 图3是多组学数据为连续数据时，提取组学特征数据流程图；

[0037] 图4是复发转移分析模型的构建流程图；

[0038] 图5是复发转移分析模型与LR、SVM、NB三种分析模型的分类性能比较图；

[0039] 图6是复发转移分析模型与LR、SVM、NB三种分析模型的ROC曲线比较图。

具体实施方式

[0040] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0041] 在此，需要说明的是，为了避免因不必要的细节而模糊了本发明，在附图中仅仅示出了与本发明的方案密切相关的结构和/或处理步骤，而省略了与本发明关系不大的其他细节。

[0042] 另外，还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。

[0043] 请参阅图1所示，为本发明提供的一种基于多组学数据的复发转移分析模型建立方法，其特征在于，包括以下步骤：

[0044] S1、对来自不同组学的数据源进行归一化处理、比较分析，建立不同组间数据的关系，获取多组学数据，并对所述多组学数据中的组学特征数据进行提取；

[0045] S2、利用主成分分析法对所述组学特征数据进行降维处理；

[0046] S3、对降维处理后的所述组学特征数据进行数据增强，使得所述组学特征数据符合样本量要求；

[0047] S4、基于符合样本量要求的所述组学特征数据构件复发转移分析模型；所述复发转移分析模型表示为：

[0048] $$\log\left(\frac{H(x)}{1-H(x)}\right)=c_0+\sum_{m=1}^{M=3}c_m H_{mT}(x);$$

[0049] 其中，H(x)是集成分类器；c_0为常数，c_m是每个弱分类器的权值；M是弱分类器的指标，M＝1，2，3；H_{mT}是每个弱分类器的权重h_t的同态积分。

[0050] 以下说明书部分将针对S1～S2进行详细描述。

[0051] 在S1中，对所述多组学数据中的组学特征数据进行提取为根据所述多组学数据的类型进行提取。请参阅图2所示，当所述多组学数据为离散数据时，通过Fisher精确检验法或卡方检验法来确定每组组学数据中的数据特征与其复发转移标签之间的相关性，对所述组学特征数据进行提取。

[0052] 具体的，所述Fisher精确检验法具体为：根据所述多组学数据构建列联表(如下表1所示)，并根据选取合适的阈值p确定是否选取该数据特征。

[0053] 表1列联表

[0054]

	第一数据A+	第一数据A-	总计
第二数据B+	a	b	a＋b
第二数据B-	c	d	c＋d
总计	a＋c	b＋d	n

[0055] 其中，第一数据为复发转移标签数据；第二数据为数据特征；a、b、c、d、n均为构建列联表时的统计获取的数据。

[0056] 进一步的，在本发明的一较佳实施例中，数据特征共包含四类，分别是临床特征、体细胞突变特征、蛋白质组学特征以及磷酸化蛋白质组学特征，当然本发明的其它实施例中数据特征还可为其它特征。

[0057] 在本方法中，阈值p表示为：

[0058] $$p=\frac{C_{a+b}^{a}C_{c+d}^{c}}{C_{n}^{a+c}}=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!};$$

[0059] 优选的，阈值p为0.05或0.01，当然在本发明的其他实施例中，所述阈值p还可为其它数值。

[0060] 所述卡方检验法具体为：

[0061] $$\chi^2=\sum_{i=1}^{k}\frac{(x_i-np_i)^2}{np_i};$$

[0062] 其中，n是观察次数，k是不同类别的数量，x_i是观察值，p_i是第i类出现的概率。

[0063] 请参阅图3所示，当所述多组学数据为为连续数据时，根据标签将所述多组学数据

分为第一数据和第二数据，至少结合T检验法、Mann-Whitney U检验法以及方差分析法的假设检验结果的交集，获取所述组学特征数据。

[0064] 具体的，T检验法具体为：

[0065] $$t=\frac{\overline{X}_1-\overline{X}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}};$$

[0066] $$s_p=\sqrt{\frac{(n_1-1)s^2{}_{X_1}+(n_2-1)s^2{}_{X_2}}{n_1+n_2-2}};$$

[0067] 其中，$s^2{}_{X_1}$和$s^2{}_{X_2}$是第一数据和第二数据两个集合的方差，n是第一数据和第二数据两个集合的大小。

[0068] Mann-WhitneyU检验法具体为：

[0069] $$U=\sum_{i=1}^{n}\sum_{j=1}^{m}S(X_i,Y_j);$$

[0070] $$S(X_i,Y_j)=\begin{cases}1 & Y<X\\ 0.5 & Y=X\\ 0 & Y>X\end{cases};$$

[0071] 其中，n表示第一数据的数量；m表示第二数据的数量；X_i表示第i个第一数据；Y_j表示第i个第二数据。

[0072] 方差分析法(ANOVA)具体为：

[0073] $SS_{total}=SS_{treatment}+SS_{error}$

[0074] $DF_{total}=DF_{treatment}+DF_{error}$

[0075] $MS_{treatment}=SS_{treatment}/DF_{treatment}$

[0076] $MS_{error}=SS_{error}/DF_{error}$

[0077] $$F=\frac{MS_{treatment}}{MS_{error}}=\frac{SS_{treatment}/DF_{treatment}}{SS_{error}/DF_{error}}$$

[0078] 其中，SS表示平方和，DF表示自由度，MS表示均方；Treatment表示不同组组学数据；Error表示同一组组学数据。

[0079] 需要说明的是，在本发明中，获取的所述组学特征数据至少包括10组。

[0080] 所述S2为利用主成分分析法对所述组学特征数据进行降维处理。具体的，S1中获取的组学特征数据为高维的组学特征，所述主成分分析法具体为：

[0081] $T_L=XW_L$；

[0082] 其中，W_L为将包含i个变量的原始数据X映射到数据集上包含i个不相关变量的新空间T，仅保留前L个主成分进行降维操作。

[0083] 如此设置，可以将高维的组学特征缩减到新的低维特征上，从而便于所述组学特征数据进一步分析和后续复发转移分析模型的建立。

[0084] 所述S3具体为：

[0085] S31、样本量的估计，计算每个选定的所述组学特征数据相对应于预设统计显著性的最佳样本量n，对所述组学特征数据进行数据扩充；

[0086] S32、对所述组学特征数据扩充后的伪数据集进行评估，通过最大Fisher判别比或F1验证生成的所述所述组学特征数据是否可以用于分类，并评估伪数据集数据增强的质量。

[0087] 具体的，所述S31中，所述最佳样本量n表示为：

[0088] $$n=\frac{\sigma^2(Q_1^{-1}+Q_2^{-1})(\mu_\alpha+\mu_\beta)^2}{\delta^2};$$

[0089] 其中，σ为标准差，μ_α和μ_β是第一类错误率和第二类错误率下μ检验的临界值，Q_1和Q_2为经过二分类后的群体中各部分的比例，δ为两组数据均值的差值。

[0090] 进一步的，所述S31中，当所述最佳样本量n不满足最佳样本量的要求时，使用SMOTE算法对所述组学特征数据进行数据增强，生成伪数据集来扩充所述组学特征数据以达到最佳样本量的要求。

[0091] 具体的，S31为使用SMOTE算法对所述组学特征数据进行过采样，SMOTE算法具体包括：

[0092] 定义组学特征数据集T；组学特征数据集T表示为：

[0093] $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，

[0094] 其中，x_i为样本i对应的所述组学特征数据，y_i为样本i对应的复发转移标签；n为样本数量；

[0095] 进一步的，采用最近邻算法对组学特征数据集T中的每个数据(x_i,y_i)选择具有相同标签的最近邻居，且所述最近邻居的选择数量为k，且k个最近邻居所构成的最近邻居集为K，最近邻居集K为：

[0096] $K=\{(x_{i1},y_i),(x_{i2},y_i),\cdots,(x_{ik},y_i)\}$；

[0097] 在最近邻居集K中随机选取m个邻居，m个邻居构成随机邻居集M，随机邻居集M为：

[0098] $M=\{(x_{i1},y_i),(x_{i2},y_i),\cdots,(x_{im},y_i)\}$；

[0099] 进一步的，根据随机邻居集M中的每个数据(x_{ij},y_{ij})生成新数据：

[0100] $x_{new}=x_i+rand(0,1)*(x_{ij}-x_i)$；

[0101] $y_{new}=y_i$；

[0102] 根据随机邻居集M中的每个数据(x_{ij},y_{ij})整理获取伪数据集G；伪数据集G表示为：

[0103] $G=\{(x_1,y_i),(x_2,y_i),\cdots,(x_{n*m},y_i)\}$；

[0104] 其中，y_i为标签。

[0105] 所述S32具体为对所述组学特征数据扩充后的伪数据集进行评估，通过最大Fisher判别比或F1验证生成的所述所述组学特征数据是否可以用于分类，并评估伪数据集数据增强的质量。

[0106] 在本发明中，对所述组学特征数据扩充后的伪数据集进行评估为通过最大Fisher判别比F1进行验证，以确保扩充后的伪数据集G中每个数据均足以用于分类，以对扩充的伪数据集数据的质量进行增强。

[0107] 具体的，F1值表示伪数据集G中伪数据的重叠程度；F1值越大，说明伪数据集G中的重叠程度越低，越适合分类；进一步的，重叠程度值F1为通过每个特征的重叠程度值f_i计算获取；

[0108] $$f_i = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2};$$

[0109] $F_1 = \max(f_i)$；

[0110] 其中，μ_1，μ_2，σ_1和σ_2分别是第一数据和第二数据的均值和标准差。

[0111] 如此，可通过分别计算初始数据集的F1值和伪数据集的F1值对伪数据集G的质量进行评估；且若伪数据集的F1值比初始数据集的F1值大或者相近，则认为构造的伪数据集G质量较好，否则构造的伪数据集G可能影响后续的复发转移分析模型的准确性。

[0112] 请参阅图4所示，S4为基于符合样本量要求的所述组学特征数据构件复发转移分析模型；在本申请中，所述复发转移分析模型为基于逻辑回归(LR)、支持向量机(SVM)和朴素贝叶斯(Naive-Bayes)三种分类方法构建获取。

[0113] 具体的，复发转移分析模型的建立主要通过以下方程依次计算获取；首先，获取所述组学特征数据的权重分布$D_t(i)$，其中，所述组学特征数据为获取的最佳样本量n符合最佳样本量要求的原始组学特征数据；也可为经扩充后补充伪数据集G的组学特征数据。

[0114] $$D_t(i) = \frac{1}{n};$$

[0115] 其中，i为样本指标，n为样本个数；

[0116] 获取各弱分类器的错误率ε_t；错误率ε_t表示为：

[0117] $$\varepsilon_t = \sum_{h_t(x_i)\neq y_i}^{n} D_t(i);$$

[0118] 其中，h_t为弱分类器；

[0119] 获取各弱分类器的权重α_t，权重α_t表示为：

[0120] $$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right);$$

[0121] 进一步的，对权重分布$D_t(i)$进行更新，获取更新后的权重分布$D_{t+1}(i)$；权重分布$D_{t+1}(i)$表示为：

[0122] $$D_{t+1}(i) = \frac{D_t}{sum(D_{t+1})}\begin{cases} e^{-\alpha_t} & h_t(x_i) = y_i \\ e^{\alpha_t} & h_t(x_i) \neq y_i \end{cases};$$

[0123] 其中，样本集$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i)\}$，为符合最佳样本量n的组合特征数据样本集；$x_i$为样本集S中的第i个样本，$y_i\in\{0,1\}$表示第n个样本的复发转移标签；$y_i=0$表示第i个样本$x_i$不是复发转移患者，$y_i=1$表示第i个样本$x_i$是复发转移患者

[0124] 获取每个弱分类器h_t的同态积分H_{mT}；同态积分H_{mT}表示为：

[0125] $$H_{mT}(x) = \sum_{t=1}^{T}\alpha_t h_t(x_i);$$

[0126] 其中，T为迭代时间的阈值。

[0127] 进一步的，通过上述方程，拟合获取复发转移分析模型，复发转移分析模型表示为：

[0128] $$\log\left(\frac{H(x)}{1-H(x)}\right) = c_0 + \sum_{m=1}^{M=3} c_m H_{mT}(x);$$

[0129] 其中，M是弱分类器的指标，M=1,2,3；H(x)是集成分类器；c_m是每个弱分类器h_t的

权值。

[0130] 需要说明的是，本发明中优选弱分类器为3个，并分别为LR、Naive-Bayes、SVM分类模型，当然在本发明的其他实施例中，弱分类器还可以设置为其它数量。

[0131] 进一步的，参阅下表2所示，为本发明复发转移分析模型与LR、Naive-Bayes、SVM分类模型的性能比较表。

[0132] 表2本发明与LR、Naive-Bayes、SVM分类模型的性能比较

[0133]

	SVM	LR	Naive Bayes	复发转移分析模型
Accuracy	0.9102±0.0348	0.9095±0.0345	0.8474±0.0885	0.9368±0.0285
Precision	0.9050±0.0612	0.8976±0.0627	0.7591±0.1159	0.9315±0.0511
Sensitivity	0.8589±0.0746	0.8669±0.0714	0.9385±0.0493	0.9043±0.0680
Specificity	0.9437±0.0317	0.9379±0.0357	0.7899±0.1569	0.9586±0.0290

[0134] 图5为本发明复发转移分析模型与LR、Naive-Bayes、SVM分类模型的分类性能图，可见，相较于传统的LR、Naive-Bayes、SVM分类模型，本发明的复发转移分析模型的分类性能明显优于其他三种模型。

[0135] 进一步的，从图6可以看出，通过构造ROC曲线综合考虑灵敏度和特异性，本发明复发转移分析模型的ROC曲线优于LR、Naive-Bayes和SVM模型。

[0136] 本发明还提供了一种基于多组学数据的复发转移分析模型建立系统，可用于执行所述基于多组学数据的复发转移分析模型建立方法；所述基于多组学数据的复发转移分析模型建立系统包括：数据采集模块，所述数据采集模块用于获取及存储多组学数据；数据处理模块，所述数据处理模块用于根据所述多组学数据中每组组学数据的数据特征与复发转移标签进行提取，获取组学特征数据；其中，所述复发转移标签用于标注是否复发。

[0137] 中央处理器，包括主成分分析模块，根据$T_L = XW_L$对所述组学特征数据进行降维处理；其中，W_L为将包含i个变量的原始组学数据X映射到数据集上包含i个不相关变量的新空间T，仅保留前L个主成分进行降维操作。

[0138] 所述中央处理器还包括数据增强模块，所述数据增强模块至少用于执行SMOTE算法，对所述组学特征数据进行数据增强，生成伪数据集来扩充所述组学特征数据以达到最佳样本量的要求所述中央处理器用于对所述组学特征数据进行多组学数据进行处理。

[0139] 所述中央处理器进一步还包括集成数据模块，所述集成数据模块被配置为根据至少三个弱分类器的模型进行复发转移分析模型的建立；所述复发转移分析模型为：

[0140] $$log\left(\frac{H(x)}{1-H(x)}\right) = c_0 + \sum_{m=1}^{M=3} c_m H_{mT}(x);$$

[0141] 其中，H(x)是集成分类器；c_0为常数，c_m是每个弱分类器的权值；M是弱分类器的指标，M＝1,2,3；H_{mT}是每个弱分类器的权重的同态积分

[0142] 所述中央处理器还用于基于所述多组学数据获取所述组学特征数据，并基于修正后的所述组学特征数据构件复发转移分析模型。

[0143] 综上所述，本发明的基于多组学数据的复发转移分析模型建立方法利用多组学数据信息，从多层次对复发转移的数据进行分析，使得对复发转移数据的分析更为全面有效，

同时，通过对多组学数据进行系统化的组学特征数据的选择和降维处理，有效利用和筛选了不同组学的数据，对进行复发转移分析模型建立的多组学数据进行了质量控制，最终综合多种经典机器学习模型，提高了复发转移分析模型的准确性。

[0144] 以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或等同替换，而不脱离本发明技术方案的精神和范围。

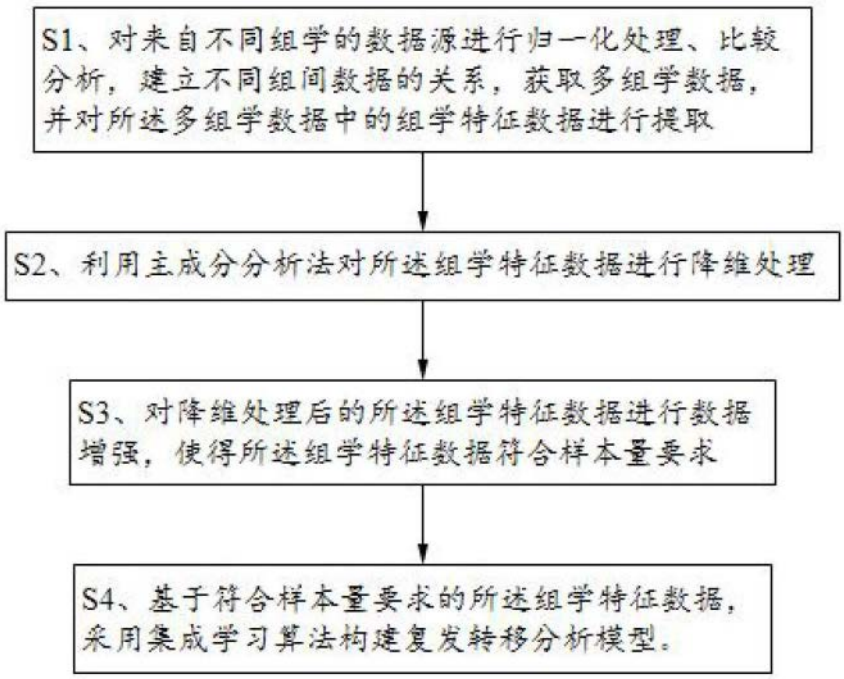

S1、对来自不同组学的数据源进行归一化处理、比较分析，建立不同组间数据的关系，获取多组学数据，并对所述多组学数据中的组学特征数据进行提取
S2、利用主成分分析法对所述组学特征数据进行降维处理
S3、对降维处理后的所述组学特征数据进行数据增强，使得所述组学特征数据符合样本量要求
S4、基于符合样本量要求的所述组学特征数据，采用集成学习算法构建复发转移分析模型。

图1

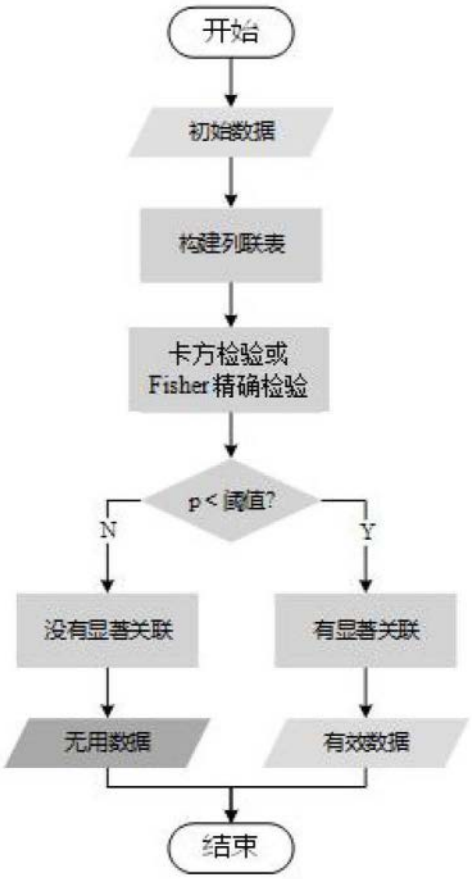

开始
初始数据
构建列联表
卡方检验或
Fisher精确检验
p < 阈值?
N
Y
没有显著关联
有显著关联
无用数据
有效数据
结束

图2

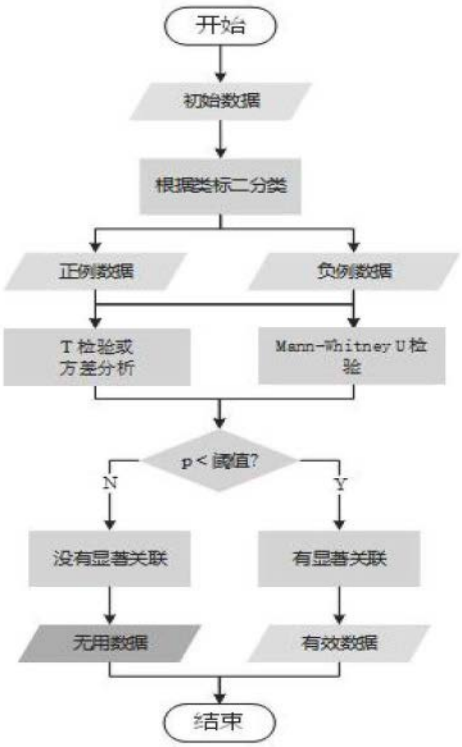

开始
初始数据
根据类标二分类
正例数据
负例数据
T检验或
方差分析
Mann-Whitney U检
验
p<阈值?
N
Y
没有显著关联
有显著关联
无用数据
有效数据
结束

图3

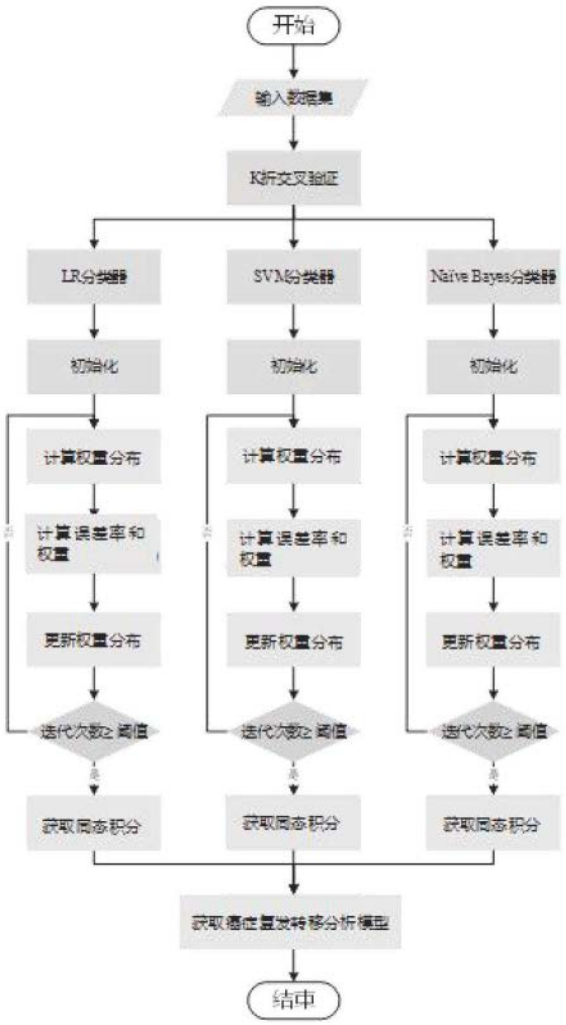

开始
输入数据集
K折交叉验证
LR分类器
SVM分类器
Naive Bayes分类器
初始化
初始化
初始化
计算权重分布
计算权重分布
计算权重分布
计算误差率和权重
计算误差率和权重
计算误差率和权重
更新权重分布
更新权重分布
更新权重分布
迭代次数≥阈值
迭代次数≥阈值
迭代次数≥阈值
是
是
是
获取同态积分
获取同态积分
获取同态积分
获取德应置发转移分析模型
结束

图4

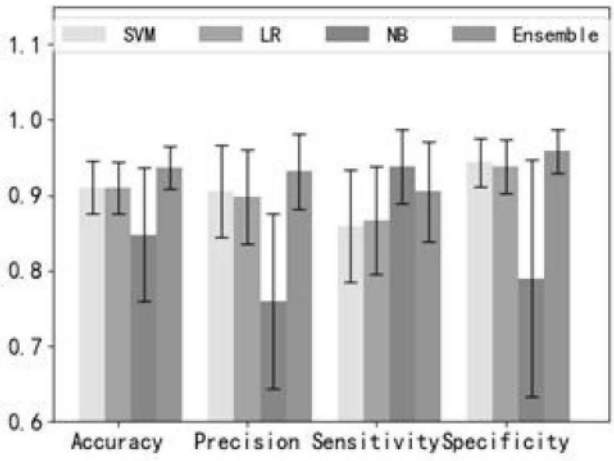

SVM
LR
NB
Ensemble
1.1
1.0
0.9
0.8
0.7
0.6
Accuracy
Precision
Sensitivity
Specificity

图5

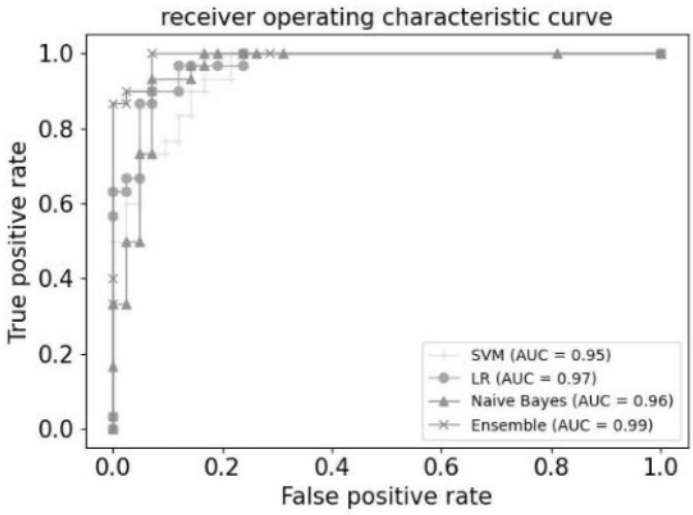

receiver operating characteristic curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
SVM (AUC = 0.95)
LR (AUC = 0.97)
Naive Bayes (AUC = 0.96)
Ensemble (AUC = 0.99)

图6