

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 6 月 11 日 (11.06.2020)



(10) 国际公布号
WO 2020/113673 A1

(51) 国际专利分类号:
G16B 40/00 (2019.01)

(21) 国际申请号: PCT/CN2018/121838

(22) 国际申请日: 2018 年 12 月 18 日 (18.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811496363.3 2018 年 12 月 7 日 (07.12.2018) CN

(71) 申请人: 深圳先进技术研究院 (SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY) [CN/CN]; 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(72) 发明人: 杨超 (YANG, Chao); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 殷鹏 (YIN, Peng); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。 蒋佳新 (JIANG,

Jiixin); 中国广东省深圳市南山区西丽大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(74) 代理人: 深圳市科进知识产权代理事务所 (普通合伙) (SHENZHEN KEJIN INTELLECTUAL PROPERTY OFFICE); 中国广东省深圳市南山区粤兴三道二号虚拟大学园产业化基地 A701 室, Guangdong 518055 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(54) Title: CANCER SUBTYPE CLASSIFICATION METHOD EMPLOYING MULTIOMICS INTEGRATION

(54) 发明名称: 一种基于多组学集成的癌症亚型分类方法

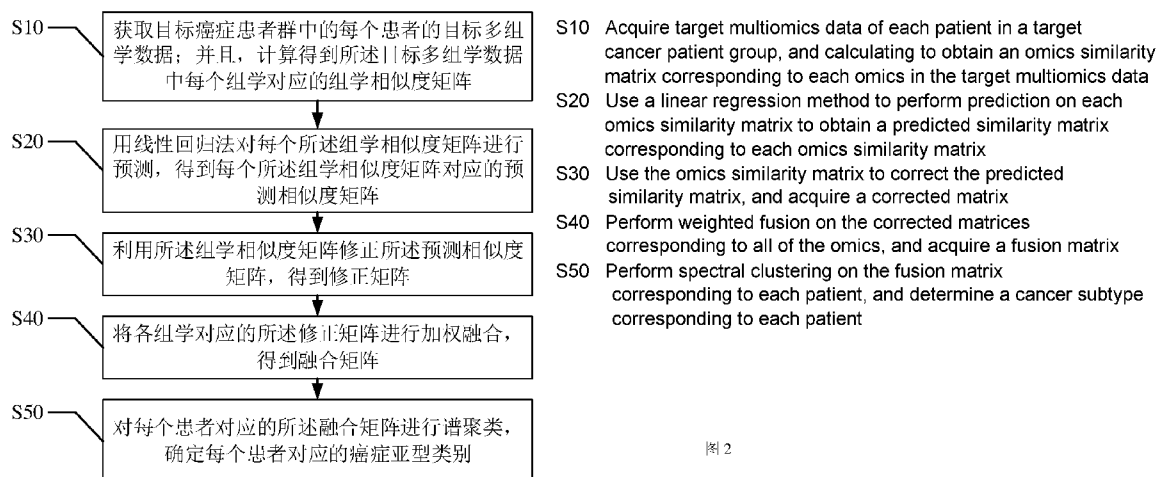


图 2

(57) Abstract: The present application provides a cancer subtype classification method employing multiomics integration. The method comprises: acquiring target multiomics data of each patient in a target cancer patient group; calculating to obtain an omics similarity matrix; performing prediction on each omics similarity matrix to obtain a predicted similarity matrix; using the omics similarity matrix to correct the predicted similarity matrix, and acquiring a corrected matrix; performing weighted fusion to obtain a fusion matrix; and performing spectral clustering on the fusion matrix, and establishing a cancer subtype category label corresponding to the fusion matrix of each patient. The present application improves the accuracy of classification evaluation of cancer subtypes, while also using a more flexible integration method to classify patients, thereby improving the efficiency of data analysis, and facilitating research on cancer subtypes.

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 本申请提供了一种基于多组学集成的癌症亚型分类方法, 包括: 获取目标癌症患者群中的每个患者的目标多组学数据; 计算得到组学相似度矩阵; 对每个所述组学相似度矩阵进行预测, 得到预测相似度矩阵; 利用所述组学相似度矩阵修正所述预测相似度矩阵得到修正矩阵; 进行加权融合, 得到融合矩阵; 对所述融合矩阵进行谱聚类, 并建立与每个患者的所述融合矩阵对应的癌症亚型类别标签。本申请提高了癌症亚型的分类评价的准确性, 并通过更灵活的整合方法实现对于患者进行分类, 提高了数据分析效率, 为对于癌症亚型的研究提供了方便。

一种基于多组学集成的癌症亚型分类方法

相关申请的交叉引用

本申请要求于 2018 年 12 月 07 日提交中国专利局的申请号为 201811496363.3、名称为“一种基于多组学集成的癌症亚型分类方法”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及癌症亚型分类评估技术领域，更具体地说，涉及一种基于多组学集成的癌症亚型分类方法。

背景技术

癌症亚型的鉴定对癌症诊断和治疗至关重要。仅利用单组学信息进行癌症亚型分类存在较失衡的类别划分，往往划分后的癌症亚型具有较大存活率差异。因此，近年来已经提出了许多通过整合目标多组学数据来鉴定癌症亚型的方法。

癌症目标多组学数据集成常用方法包括特征提取、降维和相似度矩阵计算等，其中特征提取与降维方法一般结合使用，如潜变量因子分解。常用的聚类方法有：K-means、均值漂移聚类、基于密度的聚类和谱聚类等。

但是，现有方法中并没有考虑样本之间的相似性偏差和集成中不同组学数据的权重，造成对于患者的癌症亚型分类结果的准确性差，误差较大。

发明内容

有鉴于此，本申请提供一种基于多组学集成的癌症亚型分类方法，包括：

获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵；

用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵；

利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵；

将各组学对应的所述修正矩阵进行加权融合，得到融合矩阵；

对每个患者对应的所述融合矩阵进行谱聚类，确定每个患者对应的癌症亚型类别。

优选地，所述“用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵”包括：

基于线性回归法，分别将每个患者的所述目标多组学数据的其中每个组学对应的组学相似度矩阵作为目标矩阵，利用其他组学对应的组学相似度矩阵对所述目标矩阵进行线性回归预测，分别得到所述目标多组学数据的每个所述目标矩阵中的数据对应的预测值，并得到包含所述预测值的每个所述组学相似度矩阵对应的预测相似度矩阵。

优选地，所述线性回归预测利用如下公式进行：

$$r'_{k,ij} \sim N(\beta_0 + \sum_{t \neq k} \beta_t r_{t,ij}, \sigma_{k,ij}^2);$$

其中， β_0 为超参数， β_t 为线性回归学习模型到的参数； $r'_{k,ij}$ 为预测值。

优选地，所述“利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵”包括：

将所述目标多组学数据中每个组学的所述组学相似度矩阵与对应的所述预测相似度矩阵进行求和平均，分别得到所述目标多组学数据中的每个组学的修正矩阵。

优选地，所述求和平均通过如下公式计算：

$$W_k = \frac{M_k + M'_k}{2};$$

其中， k 为所述目标多组学数据中的组学， W_k 为修正矩阵， M_k 为组学相似度矩阵， M'_k 为所述预测相似度矩阵。

优选地，所述“获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵”包括：

确定所述目标癌症患者群中每个患者的目标多组学数据，并对其中缺失的组学对应的数据进行均值插补；

对所述目标多组学数据进行相似度计算，得到所述目标多组学数据中每个组学对应的所述相似度矩阵；相似度计算公式为：

$$S_{k,ij} = \frac{\sum_{t=1}^m (x_{k,it} - \bar{x}_{k,i})(x_{k,jt} - \bar{x}_{k,j})}{\sqrt{\sum_{t=1}^m (x_{k,it} - \bar{x}_{k,i})^2} \sqrt{\sum_{t=1}^m (x_{k,jt} - \bar{x}_{k,j})^2}};$$

其中， $x_{k,it}$ 为第 k 个组学中，癌症患者 i 对应特征 t 的值； $\bar{x}_{k,i}$ 为第 k 个组学中，癌症患者 i 的平均值。

优选地，所述“对所述目标多组学数据进行相似度计算，得到所述目标多组学数据中每个组学对应的所述相似度矩阵”之后，还包括：

对所述相似度矩阵通过 Fisher 转换进行数据处理，得到处理后的所述相似度矩阵；其中，Fisher 转换的公式为：

$$r_{k,ij} = r_{k,ji} = \frac{1}{2} \ln \left(\frac{1 + M_{k,ij}}{1 - M_{k,ij}} \right);$$

其中， $r_{k,ij}$ 为矩阵变换的相似度矩阵，相似度矩阵 $M_{k,ij}$ 为 $S_{k,ij}$ 构成的矩阵。

优选地，所述“将各组学对应的所述修正矩阵进行加权融合，得到融合矩阵”包括：

利用差分搜索法，确定所述目标多组学数据中每个组学对应的组学权重；

根据每个组学的所述组学权重，对每个患者的各组学对应的所述修正矩阵进行加权融合，得到融合矩阵；其中，加权融合通过如下公式进行：

$$W = \sum_{k=1}^n \omega_k \times W_k \quad s.t. \sum \omega_k = 1;$$

其中， W_k 为本专利修正后的第 k 组学相似度矩阵， ω_k 为 W_k 对应的权重， W 为加权融合后的最终矩阵。

优选地，在所述“对每个患者对应的所述融合矩阵进行谱聚类，确定每个患者对应的癌症亚型类别”之后，还包括：

分别计算所述目标癌症患者群的每个所述癌症亚型类别中所有患者的每个组学的均值，作为亚型均值；

计算所述目标癌症患者群的每个所述癌症亚型类别中的所有所述亚型均值的亚型族群中心点；

获取所述待分析患者群中的患者的待测多组学数据；其中，所述待测多组学数据中的组学类别与所述目标多组学数据的组学类别相同；并且，计算所述待分析患者群中每个患者的待测多组学数据的中心点作为待分析中心点；

基于欧氏距离算法，计算所述待分析患者群中每个患者的待测多组学数据的待分析中心点与每个所述亚型族群中心点的相对距离，作为检测距离值；

选取所述待分析患者群中的每个患者的所有所述检测距离值中，距离最小的所述检测距离值对应的所述癌症亚型类别，作为所述待分析患者群中的该患者的癌症亚型类别。

此外，为解决上述问题，本申请还提供一种基于多组学集成的癌症亚型分类装置，包括：

获取模块，配置成获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵；

预测模块，配置成用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵；

修正模块，配置成利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵；

融合模块，配置成将各组学对应的所述修正矩阵进行加权融合，得到融合矩阵；

聚类模块，配置成对每个患者对应的所述融合矩阵进行谱聚类，确定每个患者对应的癌症亚型类别。

本申请提供一种基于多组学集成的癌症亚型分类方法。本申请通过计算得到目标患者群中每个患者的目标多组学数据中的每个组学对应的组学相似度矩阵，并且利用线性回归法进行预测得到每个所述组学相似度矩阵对应的预测相似度矩阵，进而将组学相似度矩阵和预测相似度矩阵进行组合修正，得到修正矩阵，根据权重进行加权融合，再进行谱聚类，从而为每个患者建立基于预设癌症亚型类别标签的对应的癌症亚型类别编号。本申请在相似性矩阵基础上，提出了一种简单有效的相似性融合模型，配置成整合目标多组学数据以识别癌症亚型。针对每个组学数据中的样本之间的所存在的相似性偏差，并使用线性模型预测样本之间的相似性进行修正，进而权重来整合来自目标多组学数据的校正的修正矩阵，实现将患者样本聚类到不同的亚型组中进行分类。本申请提高了癌症亚型的分类评价的准确性，并通过更灵活的整合方法实现对于患者进行分类，提高了数据分析效率，为对于癌症亚型的研究提供了方便。

附图说明

图1为本申请基于多组学集成的癌症亚型分类方法实施例方案涉及的硬件运行环境的结构示意图；

图2为本申请基于多组学集成的癌症亚型分类方法第一实施例的流程示意图；

图3为本申请基于多组学集成的癌症亚型分类方法第二实施例的流程示意图；

图4为本申请基于多组学集成的癌症亚型分类方法第三实施例的流程示意图；

图5为本申请基于多组学集成的癌症亚型分类方法第三实施例中另一种实施方式的流程示意图；

图6为本申请基于多组学集成的癌症亚型分类方法第四实施例的步骤S50之后的流程示意图；

图7为本申请基于多组学集成的癌症亚型分类方法的胶质母细胞瘤癌症亚型的存活率和存活时间对比关系图；

图8为本申请基于多组学集成的癌症亚型分类装置的功能模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

下面详细描述本申请的实施例，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。

此外，术语“第一”和“第二”仅用于描述目的，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。由此，限定有“第一”和“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本申请的描述中，“多个”的含义是两个或两个以上，除非另有明确具体的限定。

在本申请中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”和“固定”等术语应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或成一体；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通或两个元件的相互作用关系。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本申请中的具体含义。

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

如图 1 所示，是本申请实施例方案涉及的终端的硬件运行环境的结构示意图。

本申请实施例计算机设备可以是 PC，也可以是智能手机、平板电脑、或者具有一定便携计算机等可移动式终端设备。如图 1 所示，该计算机设备可以包括：处理器 1001，例如 CPU，网络接口 1004，用户接口 1003，存储器 1005，通信总线 1002。其中，通信总线 1002 配置成实现这些组件之间的连接通信。用户接口 1003 可以包括显示屏、输入单元比如键盘和遥控器，可选用户接口 1003 还可以包括标准的有线接口和无线接口。网络接口 1004 可选地可以包括标准的有线接口和无线接口（如 WI-FI 接口）。存储器 1005 可以是高速 RAM 存储器，也可以是稳定的存储器，例如磁盘存储器。存储器 1005 可选地还可以是独立于前述处理器 1001 的存储装置。可选地，终端还可以包括 RF（Radio Frequency，射频）电路、音频电路和 WiFi 模块等等。此外，计算机设备还可配置陀螺仪、气压计、湿度计、温度计和红外线传感器等其他传感器，在此不再赘述。

本领域技术人员可以理解，图 1 中示出的计算机设备并不构成对计算机设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。如图 1 所示，作为一种计算机可读存储介质的存储器 1005 中可以包括操作系统、数据接口控制程序、网络连接程序以及基于多组学集成的癌症亚型分类程序。

本申请提供一种基于多组学集成的癌症亚型分类方法。其中，所述方法提高了癌症亚型的分类评价的准确性，并通过更灵活的整合方法实现对于患者进行分类，提高了数据分析效率，为对于癌症亚型的研究提供了方便。

实施例 1:

参照图 2，本申请第一实施例提供一种基于多组学集成的癌症亚型分类方法，包括：

步骤 S10，获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵；

上述，组学相似度矩阵即为 M_k ；其中， M_k 可以通过如下形式表达：

$$M_k = \begin{pmatrix} S_{k,11} & \cdots & \cdots \\ \vdots & \ddots & \cdots \\ \vdots & \cdots & S_{k,ij} \\ & \cdots & \ddots \end{pmatrix}$$

上述，目标癌症患者群为需要进行数据分析，对该群中的所有患者进行批量癌症亚型分类的集合。目标癌症患者群中包含有具有相同类型癌症但具有相同和/或不同情况的患者的病理数据（理化指标数据和生化检验结果等）。

上述，在目标癌症患者群中，包含有多个具有相同类型癌症的患者，其中，每个患者均具有包含有多个组学的目标多组学数据。

上述，所述目标多组学数据，即为在目标癌症患者群中，每个患者均具有的需要进行数据分析的多个组学的组合。

例如，对于肺癌，建立肺癌的目标癌症患者群。群中所有的患者均为肺癌患者，数量为 400 个患者。根据不同组学的重要性和与肺癌的相关性，定义 mRNA、甲基化和基因表达量这 3 个组学进行分析研究，将 mRNA、甲基化和基因表达量这 3 个组学作为每个患者对应的目标多组学数据，而其中 mRNA、甲基化和基因表达量为目标多组学数据中的单个的组学。

步骤 S20，用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵；

步骤 S30，利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵；

上述，现有的癌症亚型的数据分类处理技术中，具有的解决方案一般为如下情况：

- (1)集成目标多组学数据；
- (2)进行聚类；
- (3)聚类结果进行存活率分析；
- (4)评估聚类结果。

由此可见，现有的癌症亚型的数据分类处理方法中并没有考虑样本之间的相似性偏差和集成中不同组学数据的权重，这也是现有的分类方法中的普遍现象，而特征维度过多，特征选择的质量影响聚类结果的质量，大大降低结果的可信度和准确度。

本实施例中考量了这些缺点，利用简单的回归和线性融合集成了不同组学数据的权重，避免了特征选择和降维。本实施例不仅考虑了不同类型数据的患者之间的相似性，还权衡了不同类型占有的权重，最后利用谱聚类方法。模型在相似性矩阵基础上进行改进，可以简单有效地进行癌症亚型分类。提高癌症亚型类别划分的质量，亚型内一致性更强，更有利于保障亚型的后续研究和癌症治疗。

本实施例中，提出在或得到目标癌症患者群中每个患者的多组学相似度矩阵后，通过线性回归法建立每个相似度矩阵对应的预测矩阵，并通过预测矩阵对相似度矩阵进行修正，即将实测值和预测值进行综合校正，从而可以得到一致性更强、准确度更高和数据可信度更高的分类结果。

需要说明的是，线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。本实施例中，通过多元线性回归分析，对组学相似度矩阵中的每个数据进行预测，得到一个与该组学相似度矩阵相对应的预测相似度矩阵。

步骤 S40，将各组学对应的所述修正矩阵进行加权融合，得到融合矩阵；

上述，根据每个组学对应的权重，对目标多组学数据中的多个组学对应的修正矩阵进行加权融合，从而得到了每个患者对应的融合矩阵。

步骤 S50，对每个患者对应的所述融合矩阵进行谱聚类，确定每个患者对应的癌症亚型类别。

上述，预设癌症亚型类别标签为对于目标的癌症类型的分类的标记。例如，通过聚类后，肺癌分类亚型分为三型，一型，二型，三型，分别为肺癌的类别标签。对应地，癌症亚型类别编号，一型为 C1，二型为 C2，三型为 C3，本实施例中，将该类别标号作为对应的类别标签，从而实现基于该标签的分类。

需要说明的是，谱聚类算法建立在谱图理论上，与传统的聚类算法相比，它具有能在任意形状的样本空间上聚类且收敛于全局最优解的优点。该算法首先根据给定的样本数据集定义一个描述成对数据点相似度的亲和矩阵，并且计算矩阵的特征值和特征向量，然后选择合适的特征向量聚类不同的数据

点。

上述，通过谱聚类，对目标癌症进行分类，并在分类后建立不同分类类型对应的癌症亚型类别标签（例如，C1\C2\C3），即实现了患者的融合矩阵与癌症亚型类别标签之间的关系的建立，从而通过癌症亚型类别标签对患有不同亚型癌症的患者进行分类的目的。

本实施例提供的一种基于多组学集成的癌症亚型分类方法，通过计算得到目标患者群中每个患者的目标多组学数据中的每个组学对应的组学相似度矩阵，并且利用线性回归法进行预测得到每个所述组学相似度矩阵对应的预测相似度矩阵，进而将组学相似度矩阵和预测相似度矩阵进行组合修正，得到修正矩阵，根据权重进行加权融合，再进行谱聚类，从而为每个患者建立基于预设癌症亚型类别标签的对应的癌症亚型类别编号。本实施例在相似性矩阵基础上，提出了一种简单有效的相似性融合模型，配置成整合目标多组学数据以识别癌症亚型。针对每个组学数据中的样本之间的所存在的相似性偏差，并使用线性模型预测样本之间的相似性进行修正，进而权重来整合来自目标多组学数据的校正的修正矩阵，实现将患者样本聚类到不同的亚型组中进行分类。本实施例提高了癌症亚型的分类评价的准确性，并通过更灵活的整合方法实现对于患者进行分类，提高了数据分析效率，为对于癌症亚型的研究提供了方便。

实施例 2:

参照图 3，本申请第二实施例提供一种基于多组学集成的癌症亚型分类方法，基于上述图 2 所示的第一实施例，所述步骤 S20，“用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵”包括：

步骤 S21，基于线性回归法，分别将每个患者的所述目标多组学数据的其中每个组学对应的组学相似度矩阵作为目标矩阵，利用其他组学对应的组学相似度矩阵对所述目标矩阵进行线性回归预测，分别得到所述目标多组学数据的每个所述目标矩阵中的数据对应的预测值，并得到包含所述预测值的每个所述组学相似度矩阵对应的预测相似度矩阵；

所述线性回归预测利用如下公式进行：

$$r'_{k,ij} \sim N(\beta_0 + \sum_{t \neq k} \beta_t r_{t,ij}, \sigma_{k,ij}^2);$$

其中， β_0 为超参数， β_t 为线性回归学习模型到的参数； $r'_{k,ij}$ 为预测值。

上述，每个患者的目标多组学数据中，包括多个组学，每个组学对应的通过相似度计算，得到了对应的组学相似度矩阵。然后，利用线性回归方法，对每个组学进行预测，得到预测相似度矩阵。

具体地，目标多组学数据中包含有的每个组学，分别通过其中的一个组学的组学相似度矩阵作为目标矩阵，利用区别于该目标矩阵的其他组学对应的组学相似度矩阵对该目标矩阵进行线性回归预测，可得到该目标矩阵中数据的预测值，即得到了该目标矩阵对应的预测相似度矩阵。然后，利用该方法对区别于目标矩阵的其他矩阵进行预测，从而分别得到了每个组学相似度矩阵对应的预测相似度矩阵。

例如，患者的目标多组学数据中，包含有 M1、M2 和 M3 等 3 个组学。线性回归预测过程为：

利用 M2 和 M3 对 M1 进行线性回归预测，得到 M1'；

利用 M1 和 M3 对 M2 进行线性回归预测，得到 M2'；

利用 M1 和 M2 对 M3 进行线性回归预测，得到 M3'。上述 M1'、M2' 和 M3' 即为通过线性回归预测所得到的分别与 M1、M2 和 M3 对应的预测相似度矩阵。

所述步骤 S30，“利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵”包括：

步骤 S31，将所述目标多组学数据中每个组学的所述组学相似度矩阵与对应的所述预测相似度矩阵进行求和平均，分别得到所述目标多组学数据中的每个组学的修正矩阵；

所述求和平均通过如下公式计算：

$$W_k = \frac{M_k + M'_k}{2};$$

其中， k 为所述目标多组学数据中的组学， W_k 为修正矩阵， M_k 为组学相似度矩阵， M'_k 为所述预测相似度矩阵。

上述，在得到每个患者的组学相似度矩阵后，基于线性回归方法对每个组学相似度矩阵进行预测，得到每个组学相似度矩阵对应的预测相似度矩阵。然后，根据组学相似度矩阵和对应的预测相似度矩阵，进行求和平均，即通过预测值对已经得到的相似度值进行修正，从而将相似度值提高了准确度，这样就在考虑每个组学数据中的样本之间的相似性偏差情况下，使用线性模型预测样本之间的相似性，达到弥补患者间相似性的问题，从而使得到的相似度矩阵的值更加准确，具有可信度。

实施例 3:

参照图 4-5，本申请第三实施例提供一种基于多组学集成的癌症亚型分类方法，基于上述图 2 所示的第一实施例，所述步骤 S10，“获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵”包括：

步骤 S11，确定所述目标癌症患者群中每个患者的目标多组学数据，并对其中缺失的组学对应的数据进行均值插补；

上述，在目标多组学数据中，包含有多个组学，但是由于患者众多，不一定每个患者都完整地进行了每个组学的测试，可能存在缺检的情况，造成部分患者缺少某个组学，无法进行计算的情况，再次需要将缺项的患者所缺少的组学进行利用其它患者的所有该项的均值进行插补，从而补足数据，在不改变数据真实值的情况下保证数据的统计学意义。

步骤 S12，对所述目标多组学数据进行相似度计算，得到所述目标多组学数据中每个组学对应的所述相似度矩阵；相似度计算公式为：

$$S_{k,ij} = \frac{\sum_{t=1}^m (x_{k,it} - \bar{x}_{k,i})(x_{k,jt} - \bar{x}_{k,j})}{\sqrt{\sum_{t=1}^m (x_{k,it} - \bar{x}_{k,i})^2} \sqrt{\sum_{t=1}^m (x_{k,jt} - \bar{x}_{k,j})^2}};$$

其中， $x_{k,it}$ 为第 k 个组学中，癌症患者 i 对应特征 t 的值； $\bar{x}_{k,i}$ 为第 k 个组学中，癌症患者 i 的平均值。

上述，相似度矩阵，即为将所有患者的某个组学的数据进行列表，例如，横坐标为基因表达量，纵坐标为患者名称或编号，在该图标中具有计算每个患者的基因表达量与其他患者的基因表达量的相似性。从而建立一个与患者对应的基因表达量的组学相似度矩阵。

在另一种实施方式下，所述步骤 S10，“对所述目标多组学数据进行相似度计算，得到所述目标多组学数据中每个组学对应的所述相似度矩阵”之后，还包括：

步骤 S60，对所述相似度矩阵通过 Fisher 转换进行数据处理，得到处理后的所述相似度矩阵；其中，Fisher 转换的公式为：

$$r_{k,ij} = r_{k,ji} = \frac{1}{2} \ln \left(\frac{1 + M_{k,ij}}{1 - M_{k,ij}} \right);$$

其中, $r_{k,ij}$ 为矩阵变换的相似度矩阵, 相似度矩阵 $M_{k,ij}$ 为 $S_{k,ij}$ 构成的矩阵。

上述, 在得到每个患者的每个组学对应的组学相似度矩阵后, 通过 Fisher 转换, 对该相似度矩阵进行数据预处理。预处理的过程及对患者的组学相似度矩阵中的数据进行归一化处理, 从而得到的预处理后的组学相似度矩阵, 可在进一步数据处理时更高效地运行。

所述步骤 S40, “将各组学对应的所述修正矩阵进行加权融合, 得到融合矩阵” 包括:

步骤 S41, 利用差分搜索法, 确定所述目标多组学数据中每个组学对应的组学权重;

上述, 利用 0.05 步长的差分搜索法, 确定癌症各组学对应最佳权重。

步骤 S42, 根据每个组学的所述组学权重, 对每个患者的各组学对应的所述修正矩阵进行加权融合, 得到融合矩阵; 其中, 加权融合通过如下公式进行:

$$W = \sum_{k=1}^n \omega_k \times W_k \quad s.t \sum \omega_k = 1$$

其中, W_k 为本专利修正后的第 k 组学相似度矩阵, ω_k 为 W_k 对应的权重, W 为加权融合后的最终矩阵。

上述, 根据每个组学对应的最佳权重, 对每个患者所包括的所有组学的修正矩阵进行融合, 从而可得到每个患者的融合矩阵。

例如, 本实施例中, 对加权融合和单一组学进行了对比, 具体情况见表 1:

表 1 单一组学和加权融合在亚型存活分析的 Cox-log P-value 对比表

数据	基因表达	DNA 甲基化	miRNA 表达	加权融合
GBM	2.49×10^{-3}	5.71×10^{-3}	1.50×10^{-3}	2.66×10^{-4}

由此表可见, 加权融合具有更小的 P 值, 亚型分类可靠度更高。所以本实施例中, 采用加权融合方法对多个组学的数据进行融合, 从而可以得到在统计学意义上更加可靠, 更加准确的计算分析结果。

实施例 4:

参照图 6, 本申请第四实施例提供一种基于多组学集成的癌症亚型分类方法, 基于上述图 2 所示的第一实施例, 在所述步骤 S50, “对每个患者对应的所述融合矩阵进行谱聚类, 确定每个患者对应的癌症亚型类别” 之后, 还包括:

步骤 S70, 分别计算所述目标癌症患者群的每个所述癌症亚型类别中所有患者的每个组学的均值, 作为亚型均值;

在确认了目标癌症患者群中每个患者对应的癌症亚型类别后, 可根据确定后的类别, 构建一般规律。根据该一般规律, 作为数据分析模型, 从而对其他的单独患者或多个患者群的病例数据进行分析, 从而实现快速分型的目的。

此外, 作为构建一般规律的数据分析模型, 目标癌症患者群中的患者数量需要达到一定的数量, 数量越大, 则该数据分析模型的作为一般规律的准确性越高, 所以在此可设定一预设阈值, 当目标癌症患者群中的患者的数量达到预设阈值, 才可作为数据分析模型进行对于其他患者的组学数据的癌症分型的分析。例如, 该预设阈值为 300 例, 即目标癌症患者群中患者数量要不小于 300。

上述, 在目标癌症患者群中, 包含有多个患者, 并且, 每个患者对应一个癌症亚型类别, 即为在进行数据分型分析后, 将目标癌症患者群中的所有患者划分为根据不同癌症亚型类别对应的组。

计算所述目标癌症患者群的每个所述癌症亚型类别中所有患者的每个组学的均值, 作为亚型均值;

将所述目标癌症患者群的每个所述癌症亚型类别中所有患者的每个组学的特征数值进行求取均值,

得到亚型均值。其中，亚型均值的个数，与癌症亚型类别的个数相同。

步骤 S80，计算所述目标癌症患者群的每个所述癌症亚型类别中的所有所述亚型均值的亚型族群中心点；

通过步骤 S70 后，可得到目标癌症患者群中每个所述癌症亚型的多个亚型均值，对每个所述癌症亚型的多个亚型均值进行求平均，可得到亚型族群中心点。

步骤 S90，获取所述待分析患者群中的患者的待测多组学数据；其中，所述待测多组学数据中的组学类别与所述目标多组学数据的组学类别相同；并且，计算所述待分析患者群中每个患者的待测多组学数据的中心点作为待分析中心点；

上述，步骤“获取所述待分析患者群中的患者的待测多组学数据；其中，所述待测多组学数据中的组学类别与所述目标多组学数据的组学类别相同”可以在步骤 S70 之前或与步骤 S70 同时进行，只要在进行执行“计算所述待分析患者群中每个患者的待测多组学数据的中心点作为待分析中心点”之前完成即可。

上述，待分析患者群，为区别于目标癌症患者群的患者的组合，该群中可以为一个患者，也可以为多个患者。其中，限定待分析患者群中的每个患者的待测多组学数据的组学类别，要与目标癌症患者群中每个患者的目标多组学数据中的组学类别相一致。例如，目标癌症患者群中的目标多组学数据包括突变、甲基化和 mRNA 等，那对应的待分析患者群中每个患者也要具有突变、甲基化和 mRNA 等这些数据，在多组学数据相一致的情况下，才能进行比较和分析。

上述，通过求取待分析患者群中每个患者的待测多组学数据中所有组学的均值，从而即得到待分析中心点。

步骤 S100，基于欧氏距离算法，计算所述待分析患者群中每个患者的待测多组学数据的待分析中心点与每个所述亚型族群中心点的相对距离，作为检测距离值；

需要说明的是，欧氏距离（Euclid Distance）也称欧几里得度量和欧几里得距离，是一个通常采用的距离定义，它是在 m 维空间中两个点之间的真实距离。在二维空间中的欧氏距离就是两点之间的直线段距离。

通过欧氏距离算法，计算待分析患者群中每个患者的待测多组学数据的待分析中心点与每个亚型族群中心点的欧式距离，作为检测距离值。

所述步骤 S110，选取所述待分析患者群中的每个患者的所有所述检测距离值中，距离最小的所述检测距离值对应的所述癌症亚型类别，作为所述待分析患者群中的该患者的癌症亚型类别。

在得到每个患者的所有的检测距离值后，对其所有的检测距离值进行比较，选取其中数值上最小的检测距离值对应的癌症亚型类别，作为该患者的癌症亚型类别，从而实现了在对目标癌症患者群中所有患者分型分析后，将其作为一般规律的数据分析模型，对其他患者进行迅速分型的目的。

例如，对于新加入的待分析患者群中的单个或多个癌症患者，可以利用原聚类标签数据对单个样本或多个样本群进行归类计算，直接判别其癌症亚型类别。

在目标癌症患者群中有 500 个患者，每个患者的包括 O1, O2, O3 三个组学数据，通过步骤 S10-S50 的方法对此患者群划分出 C1 和 C2 两个亚型。将新加入一批患者（待分析患者群）中的患者设为 n_1 , n_2 , ..., n_k 。具体如下：

1、分别计算目标癌症患者群的每个所述癌症亚型类别（C1、C2）中所有患者的每个组学（O1, O2, O3）的均值，作为亚型均值，亚型均值设为 $X_{1,1}$, $X_{1,2}$, $X_{1,3}$ ，以及 $X_{2,1}$, $X_{2,2}$, $X_{2,3}$ 。其中，X 下标中的逗号前的 1 对应 C1，逗号前的 2 对应 C2，逗号后的 1、2 和 3 分别对应 O1, O2 和 O3。

2、计算所述目标癌症患者群的每个所述癌症亚型类别（C1、C2）中的所有所述亚型均值的亚型族

群中心点:

$X1 = (X_{1,1} + X_{1,2} + X_{1,3}) / 3$; 对应 C1;

$X2 = (X_{2,1} + X_{2,2} + X_{2,3}) / 3$; 对应 C2。

3、获取所述待分析患者群中的患者的待测多组学数据; 其中, 所述待测多组学数据中的组学类别与所述目标多组学数据的组学类别相同; 并且, 计算所述待分析患者群中每个患者的待测多组学数据的中心点作为待分析中心点;

分别计算新样本 $n1, n2, \dots, nk$ 的中心点:

$new1 = (n1,1 + n1,2 + n1,3) / 3$;

$new2 = (n2,1 + n2,2 + n2,3) / 3$;

...

$newk = (n_{k,1} + n_{k,2} + n_{k,3}) / 3$;

其中, $n_{k,1}, n_{k,2}, n_{k,3}$ 分别是新样本第 k 个患者 (本例中为多个患者) 在组学 O1, O2, O3 的值。

4、新样本亚型归类: 选取所述待分析患者群中的每个患者的所有所述检测距离值中, 距离最小的所述检测距离值对应的所述癌症亚型类别, 作为所述待分析患者群中的该患者的癌症亚型类别。

$$d_{i,j} = \sqrt{\sum_{j=1}^k [X_i - new(j)]^2}$$

利用欧式距离算法公式:

;

进行求取所述每个患者的所有所述检测距离值; 其中, i 为亚型类别个数, 计算新样本 k 与各亚型族中心的检测距离值 $d_{1,k}$ 和 $d_{2,k}$ (本实施例中确定的癌症亚型类别为 C1 和 C2 两个, 所以对应的要计算得到两个检测距离值)。

若 $d_{1,k} < d_{2,k}$, 则新样本 k 属于亚型 C1; 若 $d_{1,k} > d_{2,k}$, 属于亚型 C2。此外, 如果是多个癌症亚型类别的话, 例如 5 个, 则可选取其中最小的检测距离值对应的癌症亚型类别作为该患者的癌症亚型类别即可。

本实施例中, 通过对于新加入的待分析患者群中的单个或多个癌症患者, 可以利用原聚类标签数据对单个样本或多个样本群进行归类计算, 直接判别其癌症亚型类别, 从而可根据癌症亚型分类方法建立一般规律作为数据分析模型, 实现对于其他患者的数据分析, 从而可在临床研究中, 为目标患者或患者群进行快速分型和数据分析提供了方便。此外, 每个后加入的其他患者的分型的数据, 也可加入到该模型中, 从而不断地修正和提高模型分析的准确度, 可统计学上的可信度。

基于胶质母细胞瘤癌症的统计学应用实验:

为更好地说明本申请中所提供的基于多组学集成的癌症亚型分类方法, 分别进行应用对比实验。

首先, 针对于胶质母细胞瘤癌症患者, 包含有 215 个病例, 对上述 215 个病例的患者, 分别通过基于多组学集成的癌症亚型分类方法进行分类。从而得到分类结果 (如表 2)。由表 2 中可见, 通过聚类后得到的分类结果进行统计, 统计三种亚型患者的年龄、性别和生存时间, 可分析得出: C1 亚型与 C2 亚型存在显著不同的发病机制, 后续研究分析可根据临床药物等对亚型治疗效果进行实验对比, 研究各亚型患者对应的治疗药物和治疗方法。

进一步地, 对所得到的分类结果与对应的存活率进行绘图, 其中, 包含的 Subtype1、Subtype2 和 Subtype3 等三种亚型对应 C1、C2 和 C3。根据分析结果, 建立上述胶质母细胞瘤癌症患者群中三种亚型存活率与对应存活时间的比较, 获得结果如图 7, 由图 7 可见三种亚型之间的存活率存在显著性差异, 证明本实施例中所提供的基于多组学集成的癌症亚型分类方法准确有效, 并且具有数据的统计学意义,

且具有可信度。

表 2 胶质母细胞瘤的临床特征对比表

子类型 ID	C1 (N=42)	C1 (N=112)	C1 (N=61)
患者 (男性: 女性)	(24:18)	(69:43)	(41:20)
平均年龄 (岁)	46.4	58.8	54.8
平均生存时间 (天)	931.9	402.5	504.9

此外, 参考图 8, 本申请还提供一种基于多组学集成的癌症亚型分类装置, 包括:

获取模块 10, 配置成获取目标癌症患者群中的每个患者的目标多组学数据; 并且, 计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵;

预测模块 20, 配置成用线性回归法对每个所述组学相似度矩阵进行预测, 得到每个所述组学相似度矩阵对应的预测相似度矩阵;

修正模块 30, 配置成利用所述组学相似度矩阵修正所述预测相似度矩阵, 得到修正矩阵;

融合模块 40, 配置成将各组学对应的所述修正矩阵进行加权融合, 得到融合矩阵;

聚类模块 50, 配置成对每个患者对应的所述融合矩阵进行谱聚类, 确定每个患者对应的癌症亚型类别。

此外, 本申请还提供一种计算机设备, 所述计算机设备包括存储器以及处理器, 所述存储器配置成存储基于多组学集成的癌症亚型分类程序, 所述处理器运行所述基于多组学集成的癌症亚型分类程序以使所述移动终端执行如上述所述基于多组学集成的癌症亚型分类方法。

此外, 本申请还提供一种计算机可读存储介质, 所述计算机可读存储介质上存储有基于多组学集成的癌症亚型分类程序, 所述基于多组学集成的癌症亚型分类程序被处理器执行时实现如上述所述基于多组学集成的癌症亚型分类方法。

上述本申请实施例序号仅仅为了描述, 不代表实施例的优劣。

通过以上的实施方式的描述, 本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现, 当然也可以通过硬件, 但很多情况下前者是更佳的实施方式。基于这样的理解, 本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来, 该计算机软件产品存储在如上所述的一个存储介质 (如 ROM/RAM、磁碟和光盘) 中, 包括若干指令用以使得一台终端设备 (可以是手机, 计算机, 服务器, 或者网络设备等) 执行本申请各个实施例所述的方法。以上仅为本申请的优选实施例, 并非因此限制本申请的专利范围, 凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换, 或直接或间接运用在其他相关的技术领域, 均同理包括在本申请的专利保护范围内。

权利要求书

1、一种基于多组学集成的癌症亚型分类方法，其特征在于，包括：

获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵；

用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵；

利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵；

将各组学对应的所述修正矩阵进行加权融合，得到融合矩阵；

对每个患者对应的所述融合矩阵进行谱聚类，确定每个患者对应的癌症亚型类别。

2、如权利要求 1 所述基于多组学集成的癌症亚型分类方法，其特征在于，所述“用线性回归法对每个所述组学相似度矩阵进行预测，得到每个所述组学相似度矩阵对应的预测相似度矩阵”包括：

基于线性回归法，分别将每个患者的所述目标多组学数据的其中每个组学对应的组学相似度矩阵作为目标矩阵，利用其他组学对应的组学相似度矩阵对所述目标矩阵进行线性回归预测，分别得到所述目标多组学数据的每个所述目标矩阵中的数据对应的预测值，并得到包含所述预测值的每个所述组学相似度矩阵对应的预测相似度矩阵。

3、如权利要求 2 所述基于多组学集成的癌症亚型分类方法，其特征在于，所述线性回归预测利用如下公式进行：

$$r'_{k,ij} \sim N(\beta_0 + \sum_{t \neq k} \beta_t r_{t,ij}, \sigma_{k,ij}^2);$$

其中， β_0 为超参数， β_t 为线性回归学习模型到的参数； $r'_{k,ij}$ 为预测值。

4、如权利要求 1 所述基于多组学集成的癌症亚型分类方法，其特征在于，所述“利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵”包括：

将所述目标多组学数据中每个组学的所述组学相似度矩阵与对应的所述预测相似度矩阵进行求和平均，分别得到所述目标多组学数据中的每个组学的修正矩阵。

5、如权利要求 4 所述基于多组学集成的癌症亚型分类方法，其特征在于，所述求和平均通过如下公式计算：

$$W_k = \frac{M_k + M'_k}{2};$$

其中， k 为所述目标多组学数据中的组学， W_k 为修正矩阵， M_k 为组学相似度矩阵， M'_k 为所述预测相似度矩阵。

6、如权利要求 1 所述基于多组学集成的癌症亚型分类方法，其特征在于，所述“获取目标癌症患者群中的每个患者的目标多组学数据；并且，计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵”包括：

确定所述目标癌症患者群中每个患者的目标多组学数据，并对其中缺失的组学对应的数据进行均值插补；

对所述目标多组学数据进行相似度计算，得到所述目标多组学数据中每个组学对应的所述相似度矩阵；相似度 $S_{k,ij}$ 的计算公式为：

$$S_{k,ij} = \frac{\sum_{t=1}^m (x_{k,it} - \bar{x}_{k,i})(x_{k,jt} - \bar{x}_{k,j})}{\sqrt{\sum_{t=1}^m (x_{k,it} - \bar{x}_{k,i})^2} \sqrt{\sum_{t=1}^m (x_{k,jt} - \bar{x}_{k,j})^2}};$$

其中, $x_{k,it}$ 为第 k 个组学中, 癌症患者 i 对应特征 t 的值; $\bar{x}_{k,i}$ 为第 k 个组学中, 癌症患者 i 的平均值。

7、如权利要求 6 所述基于多组学集成的癌症亚型分类方法, 其特征在于, 所述“对所述目标多组学数据进行相似度计算, 得到所述目标多组学数据中每个组学对应的所述相似度矩阵”之后, 还包括:

对所述相似度矩阵通过 Fisher 转换进行数据处理, 得到处理后的所述相似度矩阵; 其中, Fisher 转换的公式为:

$$r_{k,ij} = r_{k,ji} = \frac{1}{2} \ln \left(\frac{1 + M_{k,ij}}{1 - M_{k,ij}} \right);$$

其中, $r_{k,ij}$ 为矩阵变换的相似度矩阵, 相似度矩阵 $M_{k,ij}$ 为 $S_{k,ij}$ 构成的矩阵。

8、如权利要求 1 所述基于多组学集成的癌症亚型分类方法, 其特征在于, 所述“将各组学对应的所述修正矩阵进行加权融合, 得到融合矩阵”包括:

利用差分搜索法, 确定所述目标多组学数据中每个组学对应的组学权重;

根据每个组学的所述组学权重, 对每个患者的各组学对应的所述修正矩阵进行加权融合, 得到融合矩阵; 其中, 加权融合通过如下公式进行:

$$W = \sum_{k=1}^n \omega_k \times W_k \quad s.t. \sum \omega_k = 1;$$

其中, W_k 为本专利修正后的第 k 组学相似度矩阵, ω_k 为 W_k 对应的权重, W 为加权融合后的最终矩阵。

9、如权利要求 1 所述基于多组学集成的癌症亚型分类方法, 其特征在于, 在所述“对每个患者对应的所述融合矩阵进行谱聚类, 确定每个患者对应的癌症亚型类别”之后, 还包括:

分别计算所述目标癌症患者群的每个所述癌症亚型类别中所有患者的每个组学的均值, 作为亚型均值;

计算所述目标癌症患者群的每个所述癌症亚型类别中的所有所述亚型均值的亚型族群中心点;

获取待分析患者群中的患者的待测多组学数据; 其中, 所述待测多组学数据中的组学类别与所述目标多组学数据的组学类别相同; 并且, 计算所述待分析患者群中每个患者的待测多组学数据的中心点作为待分析中心点;

基于欧氏距离算法, 计算所述待分析患者群中每个患者的待测多组学数据的待分析中心点与每个所述亚型族群中心点的相对距离, 作为检测距离值;

选取所述待分析患者群中的每个患者的所有所述检测距离值中, 距离最小的所述检测距离值对应的所述癌症亚型类别, 作为所述待分析患者群中的该患者的癌症亚型类别。

10、一种基于多组学集成的癌症亚型分类装置, 其特征在于, 包括:

获取模块, 配置成获取目标癌症患者群中的每个患者的目标多组学数据; 并且, 计算得到所述目标多组学数据中每个组学对应的组学相似度矩阵;

预测模块, 配置成用线性回归法对每个所述组学相似度矩阵进行预测, 得到每个所述组学相似度矩

阵对应的预测相似度矩阵；

修正模块，配置成利用所述组学相似度矩阵修正所述预测相似度矩阵，得到修正矩阵；

融合模块，配置成将各组学对应的所述修正矩阵进行加权融合，得到融合矩阵；

聚类模块，配置成对每个患者对应的所述融合矩阵进行谱聚类，确定每个患者对应的癌症亚型类别。

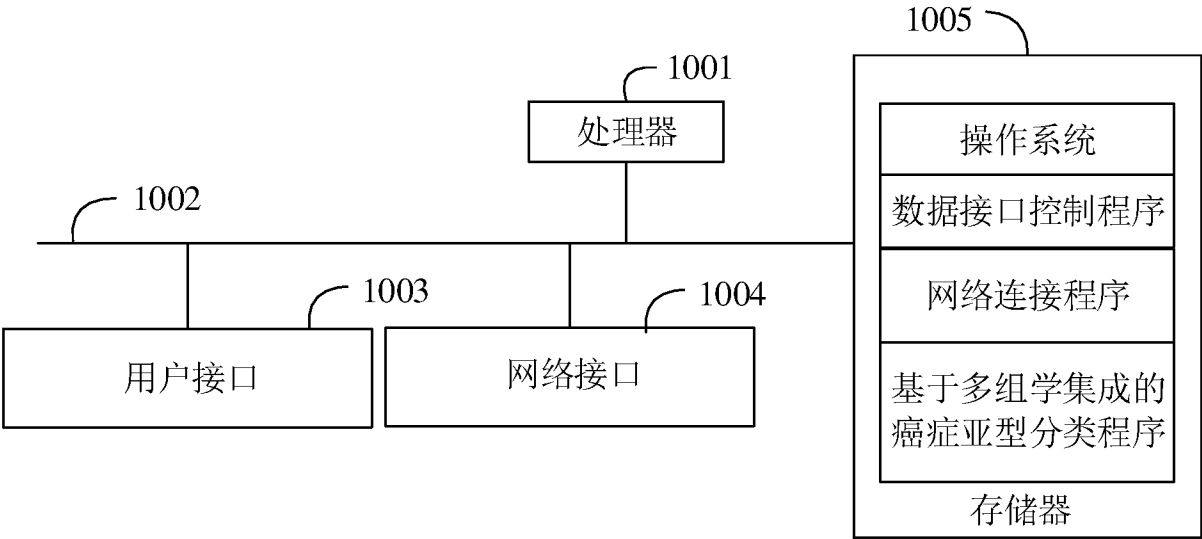


图 1

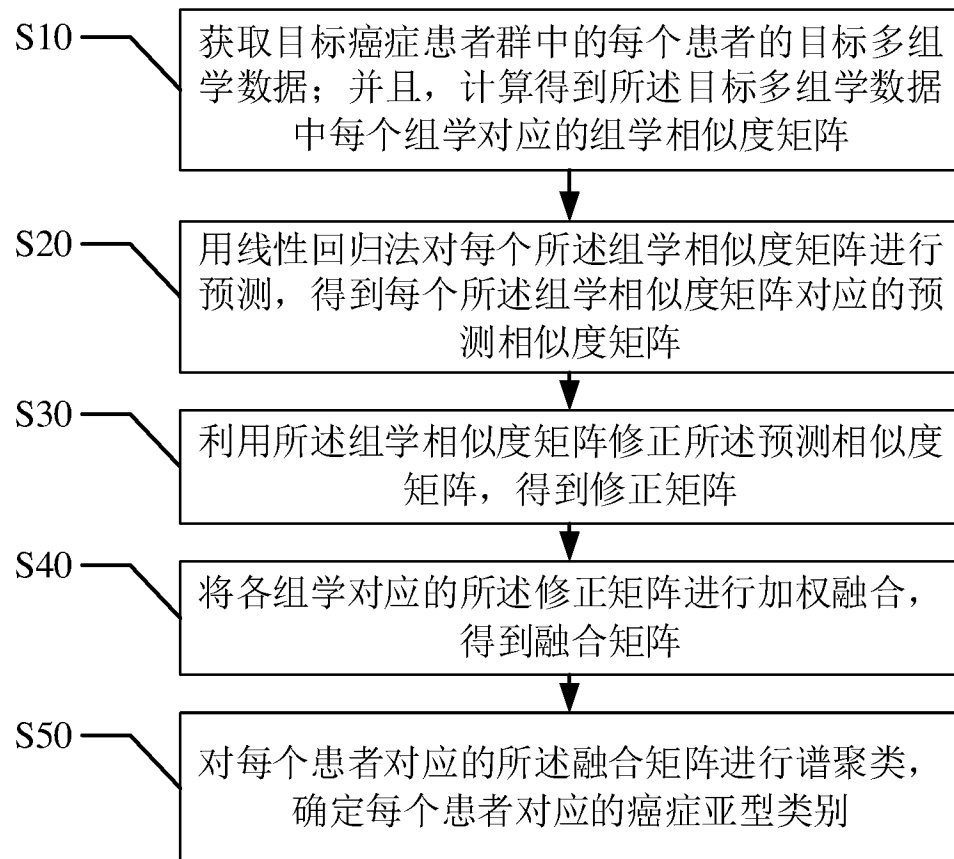


图 2

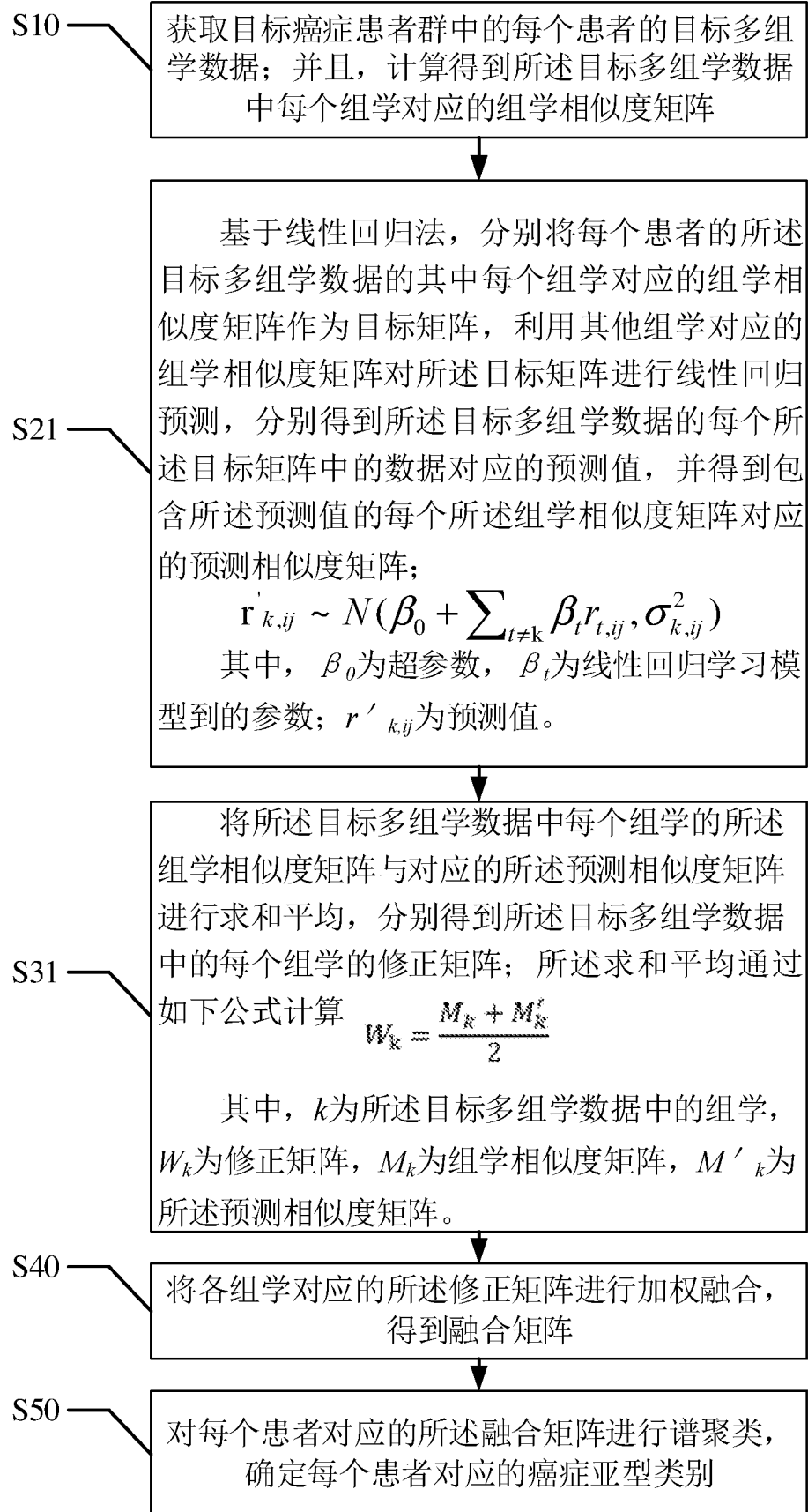


图 3

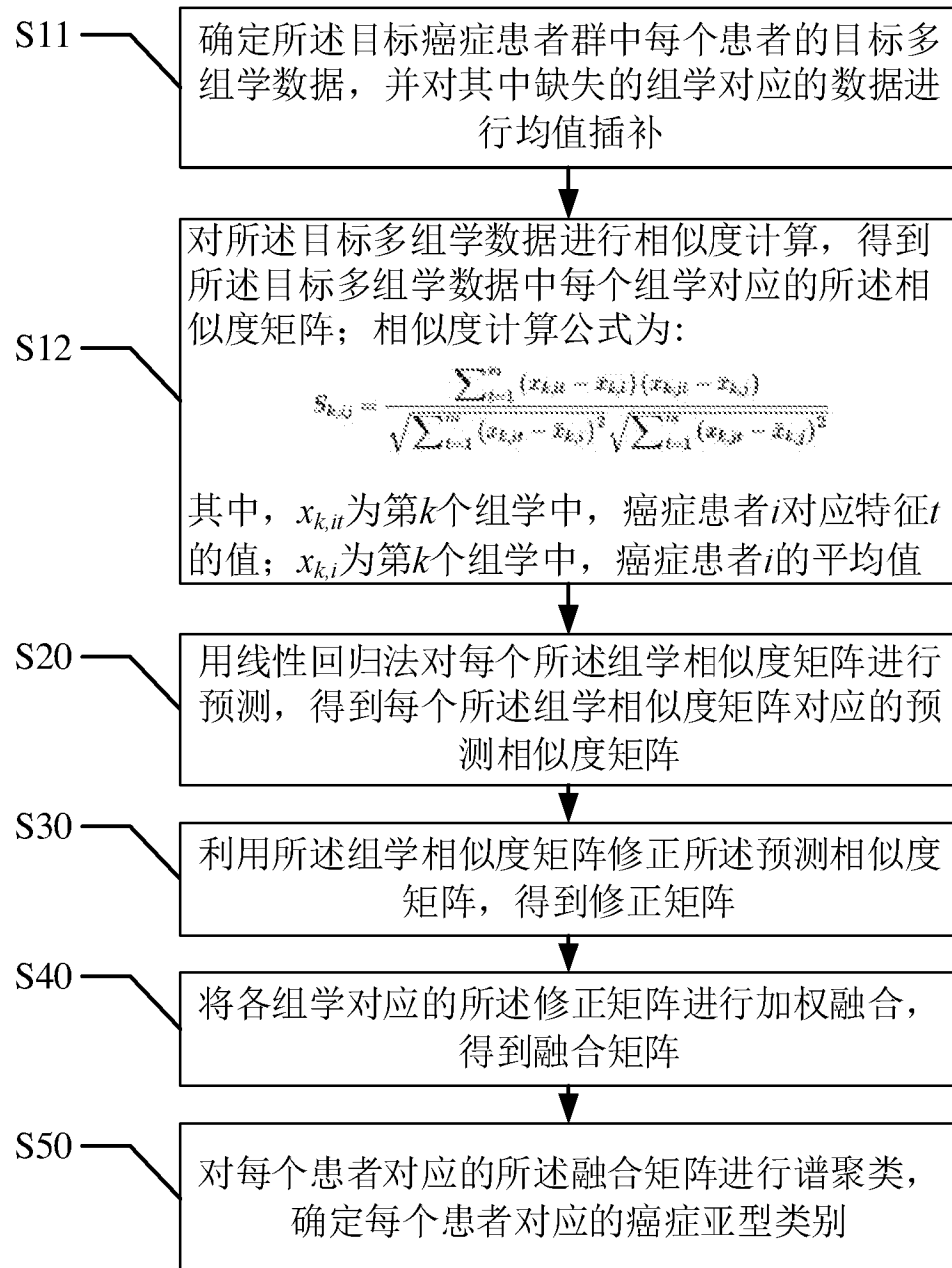


图 4

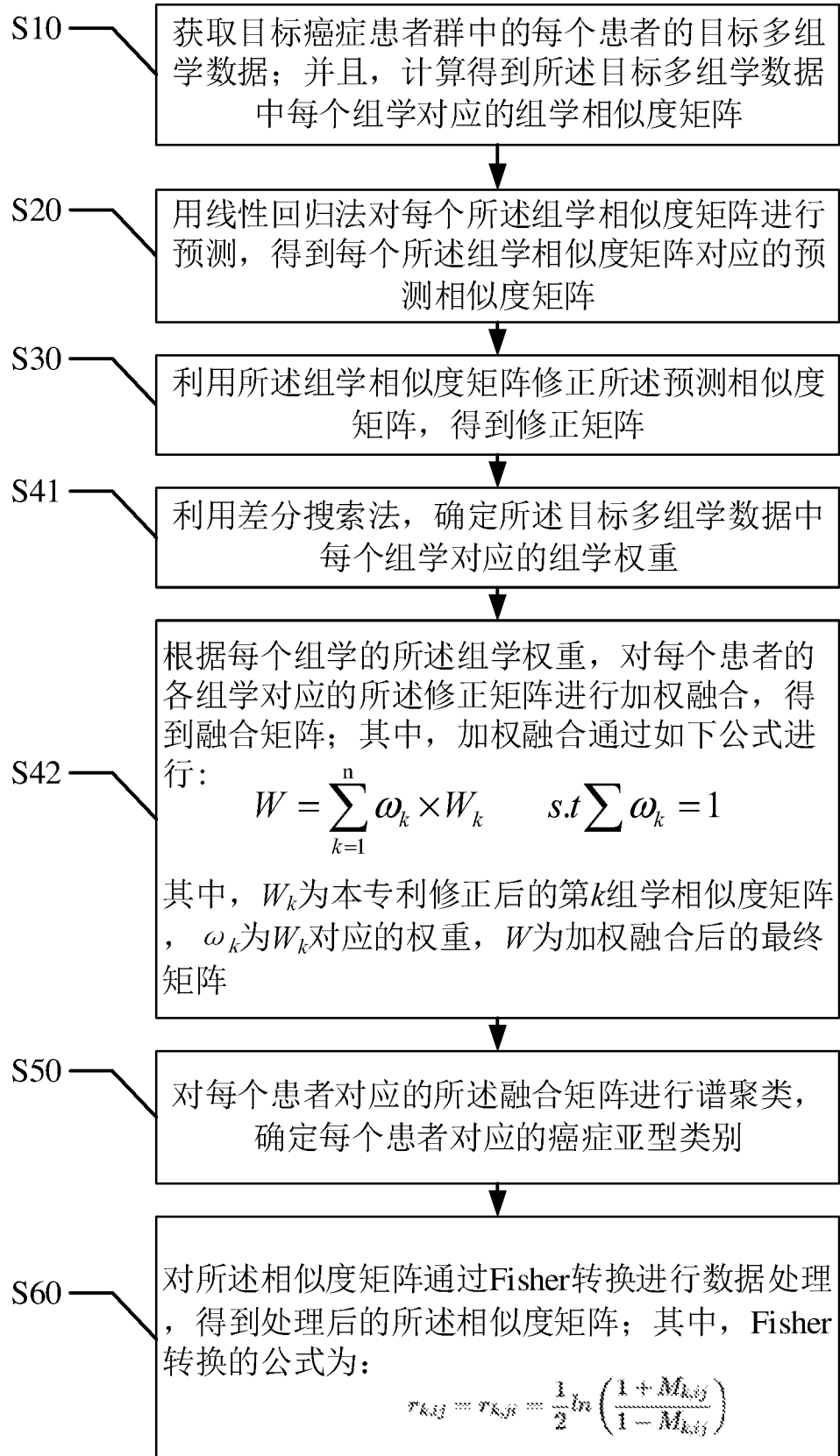


图5

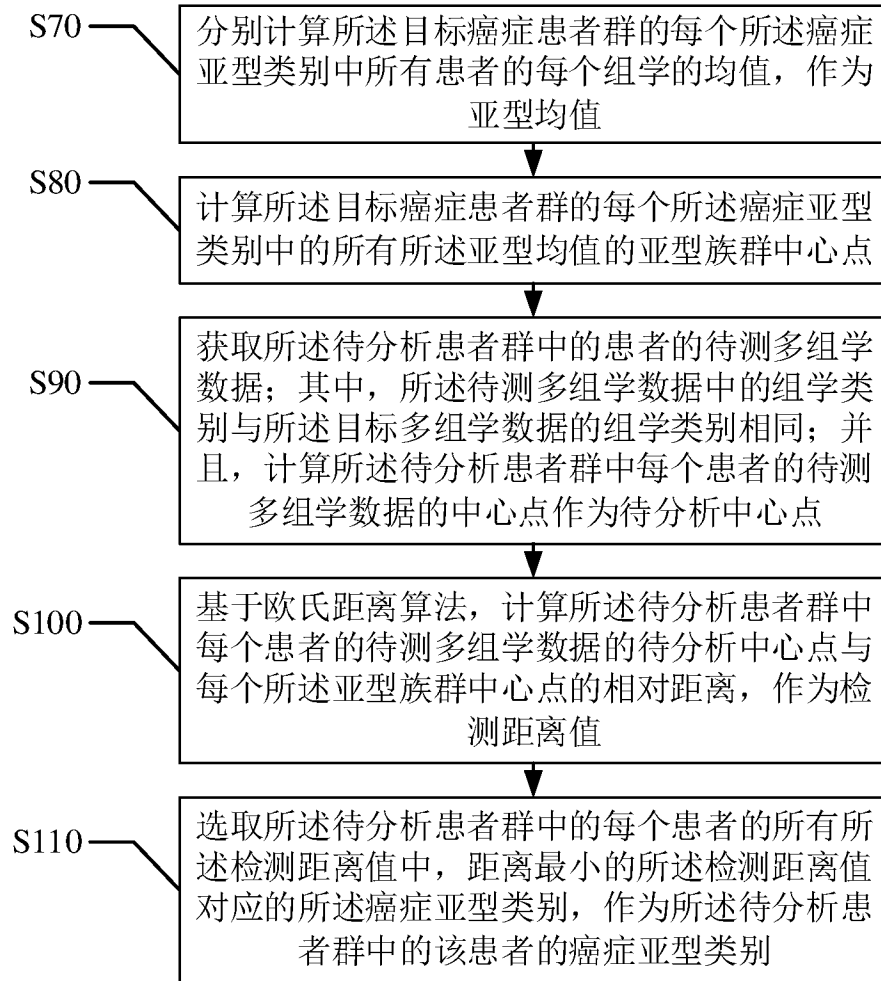


图 6

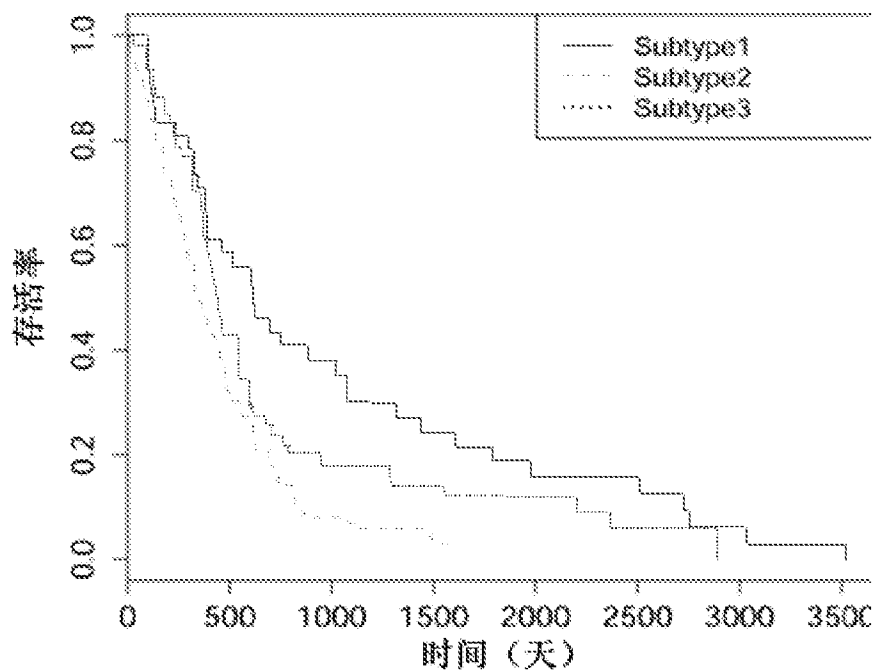


图 7

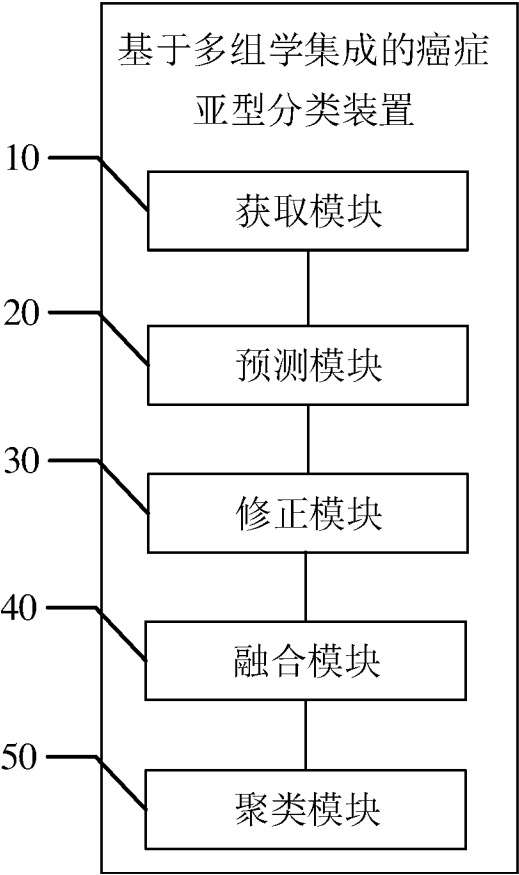


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/121838

A. CLASSIFICATION OF SUBJECT MATTER

G16B 40/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G16B, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNTXT, SIPOABS, VEN, CNKI, IEEE: 分类, 类别, 划分, 相似度, 矩阵, 预测, 加权, 谱聚类, 修正, 平均, class+, divid+, similarity, matrix, predict+, weight+, spectral, cluster+, correct+, average

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106529165 A (HEFEI UNIVERSITY OF TECHNOLOGY) 22 March 2017 (2017-03-22) entire document	1-10
A	CN 108563660 A (WENZHOU UNIVERSITY) 21 September 2018 (2018-09-21) entire document	1-10
A	CN 107506617 A (HANGZHOU DIANZI UNIVERSITY) 22 December 2017 (2017-12-22) entire document	1-10
A	CN 108171012 A (HENAN NORMAL UNIVERSITY) 15 June 2018 (2018-06-15) entire document	1-10
A	CN 105243300 A (HEFEI UNIVERSITY OF TECHNOLOGY) 13 January 2016 (2016-01-13) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

09 April 2019

Date of mailing of the international search report

14 May 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/121838

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	106529165	A	22 March 2017	None	
CN	108563660	A	21 September 2018	None	
CN	107506617	A	22 December 2017	None	
CN	108171012	A	15 June 2018	None	
CN	105243300	A	13 January 2016	None	

国际检索报告

国际申请号

PCT/CN2018/121838

A. 主题的分类 G16B 40/00 (2019.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G16B, G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNABS, CNTXT, SIPOABS, VEN, CNKI, IEEE: 分类, 类别, 划分, 相似度, 矩阵, 预测, 加权, 谱聚类, 修正, 平均, class+, divid+, similarity, matrix, predict+, weight+, spectral, cluster+, correct+, average		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 106529165 A (合肥工业大学) 2017年 3月 22日 (2017 - 03 - 22) 全文	1-10
A	CN 108563660 A (温州大学) 2018年 9月 21日 (2018 - 09 - 21) 全文	1-10
A	CN 107506617 A (杭州电子科技大学) 2017年 12月 22日 (2017 - 12 - 22) 全文	1-10
A	CN 108171012 A (河南师范大学) 2018年 6月 15日 (2018 - 06 - 15) 全文	1-10
A	CN 105243300 A (合肥工业大学) 2016年 1月 13日 (2016 - 01 - 13) 全文	1-10
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期		国际检索报告邮寄日期
2019年 4月 9日		2019年 5月 14日
ISA/CN的名称和邮寄地址		受权官员
中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		刘欢 电话号码 86-010-62411658

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/121838

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	106529165	A	2017年 3月 22日	无	
CN	108563660	A	2018年 9月 21日	无	
CN	107506617	A	2017年 12月 22日	无	
CN	108171012	A	2018年 6月 15日	无	
CN	105243300	A	2016年 1月 13日	无	