(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111462817 B
(45)授权公告日 2023.06.20

(21)申请号 202010221082.8

(22)申请日 2020.03.25

(65)同一申请的已公布的文献号
申请公布号 CN 111462817 A

(43)申请公布日 2020.07.28

(73)专利权人 哈尔滨工业大学(深圳)(哈尔滨工业大学深圳科技创新研究院)
地址 518055 广东省深圳市南山区桃源街道深圳大学城哈尔滨工业大学校区

(72)发明人 廖清 马海轩 杨林 丁烨 王轩 李京竹

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 郭浩辉 麦小婵

(51)Int.Cl.
G16B 25/10(2019.01)
G16B 30/10(2019.01)
G16B 40/00(2019.01)
G06F 18/214(2023.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 109508655 A,2019.03.22
CN 109816002 A,2019.05.28
CN 109961089 A,2019.07.02
CN 109978071 A,2019.07.05

审查员 夏冰

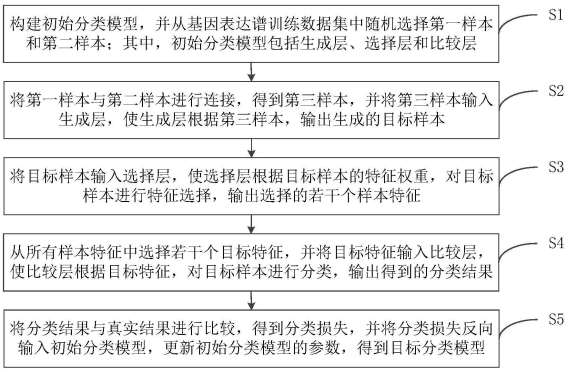
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种分类模型构建方法、装置、分类模型及分类方法

(57)摘要

本发明公开了一种分类模型构建方法、装置、分类模型及分类方法。所述分类模型构建方法通过构建初始分类模型，在初始分类模型中引入生成层、选择层和比较层，并通过在基因表达谱训练数据集中随机选择两个样本，对生成层、选择层和比较层进行训练和更新，得到目标分类模型，使得可利用生成层，根据基因表达谱数据中任意两个样本生成新的样本，利用选择层，根据新的样本各个特征的权重选择若干个样本特征，利用比较层，根据从所有样本特征中选择的若干个目标特征对新的样本进行分类。本发明能够构建一种适用于基因表达谱的分类模型，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。



CN 111462817 B

1.一种适用于基因表达谱的分类模型构建方法,其特征在于,包括:

构建初始分类模型,并从基因表达谱训练数据集中随机选择第一样本和第二样本;其中,所述初始分类模型包括生成层、选择层和比较层;

将所述第一样本与所述第二样本进行连接,得到第三样本,并将所述第三样本输入所述生成层,使所述生成层根据所述第三样本,输出生成的目标样本;

将所述目标样本输入所述选择层,使所述选择层根据所述目标样本的特征权重,对所述目标样本进行特征选择,输出选择的若干个样本特征;

从所有所述样本特征中选择若干个目标特征,并将所述目标特征输入所述比较层,使所述比较层根据所述目标特征,对所述目标样本进行分类,输出得到的分类结果;

将所述分类结果与真实结果进行比较,得到分类损失,并将所述分类损失反向输入所述初始分类模型,更新所述初始分类模型的参数,得到目标分类模型。

2.如权利要求1所述的适用于基因表达谱的分类模型构建方法,其特征在于,所述生成层包括稀疏矩阵,所述选择层包括神经网络。

3.如权利要求1所述的适用于基因表达谱的分类模型构建方法,其特征在于,在所述将所述第三样本输入所述生成层,使所述生成层根据所述第三样本,输出生成的目标样本之后,还包括:

根据所述目标样本,计算所述第一样本与所述第二样本的相似值,并根据所述相似值,计算所述生成层的生成损失,将所述生成损失反向输入所述生成层,更新所述生成层的参数。

4.如权利要求1所述的适用于基因表达谱的分类模型构建方法,其特征在于,所述将所述分类损失反向输入所述初始分类模型,更新所述初始分类模型的参数,得到目标分类模型,具体为:

将所述分类损失分别反向输入所述生成层和所述选择层,更新所述生成层和所述选择层的参数。

5.一种适用于基因表达谱的分类模型构建装置,其特征在于,包括:

初始分类模型构建模块,用于构建初始分类模型,并从基因表达谱训练数据集中随机选择第一样本和第二样本;其中,所述初始分类模型包括生成层、选择层和比较层;

生成层训练模块,用于将所述第一样本与所述第二样本进行连接,得到第三样本,并将所述第三样本输入所述生成层,使所述生成层根据所述第三样本,输出生成的目标样本;

选择层训练模块,用于将所述目标样本输入所述选择层,使所述选择层根据所述目标样本的特征权重,对所述目标样本进行特征选择,输出选择的若干个样本特征;

比较层训练模块,用于从所有所述样本特征中选择若干个目标特征,并将所述目标特征输入所述比较层,使所述比较层根据所述目标特征,对所述目标样本进行分类,输出得到的分类结果;

目标分类模型获取模块,用于将所述分类结果与真实结果进行比较,得到分类损失,并将所述分类损失反向输入所述初始分类模型,更新所述初始分类模型的参数,得到目标分类模型。

6.如权利要求5所述的适用于基因表达谱的分类模型构建装置,其特征在于,所述生成层包括稀疏矩阵,所述选择层包括神经网络。

7.如权利要求5所述的适用于基因表达谱的分类模型构建装置，其特征在于，所述生成层训练模块，还用于在所述将所述第三样本输入所述生成层，使所述生成层根据所述第三样本，输出生成的目标样本之后，根据所述目标样本，计算所述第一样本与所述第二样本的相似值，并根据所述相似值，计算所述生成层的生成损失，将所述生成损失反向输入所述生成层，更新所述生成层的参数。

8.如权利要求5所述的适用于基因表达谱的分类模型构建装置，其特征在于，所述将所述分类损失反向输入所述初始分类模型，更新所述初始分类模型的参数，得到目标分类模型，具体为：

将所述分类损失分别反向输入所述生成层和所述选择层，更新所述生成层和所述选择层的参数。

9.一种适用于基因表达谱的分类模型，其特征在于，所述分类模型是应用如权利要求1～4任一项所述的适用于基因表达谱的分类模型构建方法而获得。

10.一种适用于基因表达谱的分类方法，其特征在于，包括：

从获取的基因表达谱数据集中选择第四样本和第五样本，并将所述第四样本和所述第五样本输入如权利要求9所述的适用于基因表达谱的分类模型，得到分类结果。

一种分类模型构建方法、装置、分类模型及分类方法

技术领域

[0001] 本发明涉及基因表达谱分类技术领域，尤其涉及一种分类模型构建方法、装置、分类模型及分类方法。

背景技术

[0002] 基因表达谱包含了大量基因，其中仅有少量基因与特定类型的疾病相关，具有高维度、少样本的特性。通过分类分析基因表达谱数据，对研究人类基因的表达、各种遗传性疾病及由于细胞病变而导致的疾病具有重大意义。在分析基因表达谱数据时，往往先采用特征选择方法筛选基因表达谱数据中的重要特征，再通过机器学习模型等分类器对基因表达谱数据进行分类。

[0003] 现有技术虽然通过筛选基因表达谱数据中的重要特征来缓解高维度特性带来的过拟合问题，但却忽略解决少样本特性带来的欠拟合问题，难以进一步提高基因表达谱数据的分类准确度。

发明内容

[0004] 本发明提供一种分类模型构建方法、装置、分类模型及分类方法，以克服现有技术的缺陷，本发明能够构建一种适用于基因表达谱的分类模型，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。

[0005] 为了解决上述技术问题，第一方面，本发明一实施例提供一种适用于基因表达谱的分类模型构建方法，包括：

[0006] 构建初始分类模型，并从基因表达谱训练数据集中随机选择第一样本和第二样本；其中，所述初始分类模型包括生成层、选择层和比较层；

[0007] 将所述第一样本与所述第二样本进行连接，得到第三样本，并将所述第三样本输入所述生成层，使所述生成层根据所述第三样本，输出生成的目标样本；

[0008] 将所述目标样本输入所述选择层，使所述选择层根据所述目标样本的特征权重，对所述目标样本进行特征选择，输出选择的若干个样本特征；

[0009] 从所有所述样本特征中选择若干个目标特征，并将所述目标特征输入所述比较层，使所述比较层根据所述目标特征，对所述目标样本进行分类，输出得到的分类结果；

[0010] 将所述分类结果与真实结果进行比较，得到分类损失，并将所述分类损失反向输入所述初始分类模型，更新所述初始分类模型的参数，得到目标分类模型。

[0011] 进一步地，所述生成层包括稀疏矩阵，所述选择层包括神经网络。

[0012] 进一步地，在所述将所述第三样本输入所述生成层，使所述生成层根据所述第三样本，输出生成的目标样本之后，还包括：

[0013] 根据所述目标样本，计算所述第一样本与所述第二样本的相似值，并根据所述相似值，计算所述生成层的生成损失，将所述生成损失反向输入所述生成层，更新所述生成层

的参数。

[0014] 进一步地，所述将所述分类损失反向输入所述初始分类模型，更新所述初始分类模型的参数，得到目标分类模型，具体为：

[0015] 将所述分类损失分别反向输入所述生成层和所述选择层，更新所述生成层和所述选择层的参数。

[0016] 第二方面，本发明一实施例提供一种适用于基因表达谱的分类模型构建装置，包括：

[0017] 初始分类模型构建模块，用于构建初始分类模型，并从基因表达谱训练数据集中随机选择第一样本和第二样本；其中，所述初始分类模型包括生成层、选择层和比较层；

[0018] 生成层训练模块，用于将所述第一样本与所述第二样本进行连接，得到第三样本，并将所述第三样本输入所述生成层，使所述生成层根据所述第三样本，输出生成的目标样本；

[0019] 选择层训练模块，用于将所述目标样本输入所述选择层，使所述选择层根据所述目标样本的特征权重，对所述目标样本进行特征选择，输出选择的若干个样本特征；

[0020] 比较层训练模块，用于从所有所述样本特征中选择若干个目标特征，并将所述目标特征输入所述比较层，使所述比较层根据所述目标特征，对所述目标样本进行分类，输出得到的分类结果；

[0021] 目标分类模型获取模块，用于将所述分类结果与真实结果进行比较，得到分类损失，并将所述分类损失反向输入所述初始分类模型，更新所述初始分类模型的参数，得到目标分类模型。

[0022] 进一步地，所述生成层包括稀疏矩阵，所述选择层包括神经网络。

[0023] 进一步地，所述生成层训练模块，还用于在所述将所述第三样本输入所述生成层，使所述生成层根据所述第三样本，输出生成的目标样本之后，根据所述目标样本，计算所述第一样本与所述第二样本的相似值，并根据所述相似值，计算所述生成层的生成损失，将所述生成损失反向输入所述生成层，更新所述生成层的参数。

[0024] 进一步地，所述将所述分类损失反向输入所述初始分类模型，更新所述初始分类模型的参数，得到目标分类模型，具体为：

[0025] 将所述分类损失分别反向输入所述生成层和所述选择层，更新所述生成层和所述选择层的参数。

[0026] 第三方面，本发明一实施例提供一种适用于基因表达谱的分类模型，所述分类模型是应用如上所述的适用于基因表达谱的分类模型构建方法而获得。

[0027] 第四方面，本发明一实施例提供一种适用于基因表达谱的分类方法，包括：

[0028] 从获取的基因表达谱数据集中选择第四样本和第五样本，并将所述第四样本和所述第五样本输入如上所述的适用于基因表达谱的分类模型，得到分类结果。

[0029] 相比于现有技术，本发明的实施例，具有如下有益效果：

[0030] 通过构建初始分类模型，在初始分类模型中引入生成层、选择层和比较层，并通过在基因表达谱训练数据集中随机选择两个样本，对生成层、选择层和比较层进行训练和更新，得到目标分类模型，使得可利用生成层，根据基因表达谱数据中任意两个样本生成新的样本，利用选择层，根据新的样本各个特征的权重选择若干个样本特征，利用比较层，根据

从所有样本特征中选择的若干个目标特征对新的样本进行分类。本发明能够构建一种适用于基因表达谱的分类模型，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。

附图说明

[0031] 图1为本发明第一实施例中的一种分类模型构建方法的流程示意图；

[0032] 图2为本发明第一实施例中的初始分类模型的网络结构图；

[0033] 图3为本发明第二实施例中的一种分类模型构建装置的结构示意图。

具体实施方式

[0034] 下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0035] 需要说明的是，文中的步骤编号，仅为了方便具体实施例的解释，不作为限定步骤执行先后顺序的作用。本实施例提供的方法可以由相关的服务器执行，且下文均以服务器作为执行主体为例进行说明。

[0036] 请参阅图1-2。

[0037] 如图1-2所示，第一实施例提供一种适用于基因表达谱的分类模型构建方法，包括步骤S1～S5：

[0038] S1、构建初始分类模型，并从基因表达谱训练数据集中随机选择第一样本和第二样本；其中，初始分类模型包括生成层、选择层和比较层。

[0039] S2、将第一样本与第二样本进行连接，得到第三样本，并将第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本。

[0040] S3、将目标样本输入选择层，使选择层根据目标样本的特征权重，对目标样本进行特征选择，输出选择的若干个样本特征。

[0041] S4、从所有样本特征中选择若干个目标特征，并将目标特征输入比较层，使比较层根据目标特征，对目标样本进行分类，输出得到的分类结果。

[0042] S5、将分类结果与真实结果进行比较，得到分类损失，并将分类损失反向输入初始分类模型，更新初始分类模型的参数，得到目标分类模型。

[0043] 在本实施例的一种优选实施方式当中，生成层包括稀疏矩阵，选择层包括神经网络。

[0044] 基因表达谱数据特征的顺序取决于研究人员选择测试的基因顺序，没有确定的相对位置信息。本实施例采用稀疏矩阵作为生成层，能够抽取第一样本和第二样本特征的差异性生成目标样本，有利于避免引入多余的噪声。

[0045] 首先在对从基因表达谱训练数据集中随机选择的第一样本和第二样本进行连接后，将得到的第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本，然后将目标样本输入选择层，使选择层根据目标样本的特征权重对目标样本进行特征选择，输出选择的若干个样本特征，接着在从所有所述样本特征中选择若干个目标特征后，将目标特

征输入所述比较层，使比较层根据目标特征对目标样本进行分类，输出得到的分类结果，最终将分类结果与真实结果进行比较，得到分类损失，并将分类损失反向输入初始分类模型以更新初始分类模型的参数，得到目标分类模型。

[0046] 例如，以一种癌症基因表达谱数据集X_n作为基因表达谱训练数据集。

[0047] 从癌症基因表达谱数据集X_n中随机选择第一样本x_i和第二样本x_j，并将第一样本x_i与第二样本x_j进行连接，得到第三样本$concat(x_i,x_j)$。其中，concat表示将第一样本x_i与第二样本x_j水平连接在一起，即直接将两个d维特征的样本连接成一个2d维特征的样本。

[0048] 在生成层中，通过从癌症基因表达谱数据集X_n中任意选择第一样本x_i和第二样本x_j，使得可根据第三样本$concat(x_i,x_j)$来生成目标样本$c_{i,j}$，构建目标样本数据集$C=\{c_{1,1},c_{1,2},\cdots,c_{n2,n}\}$，从而将样本容量为n的癌症基因表达谱数据集$X_n$扩充为样本容量为$n_2$的目标样本数据集C。其中，$C \subset R^{n \times d}$，$c_{i,j}=G_{fea}(x_i,x_j)=concat(x_i,x_j)\times W_G$，$G_{fea}(x_i,x_j)$表示生成层根据第一样本$x_i$和第二样本$x_j$生成目标样本$c_{i,j}$的函数，$W_G$表示稀疏矩阵，稀疏矩阵$W_G$初始化为$W_{i,i}=1$，$W_{i,i+j}=-1$，$i,j\in[0,n]$，其他参数初始化为$W_{i,j}=0$。

[0049] 在对稀疏矩阵进行初始化后，生成层在开始阶段，产生的目标样本$c_{i,j}$的i_{th}特征表示为第一样本x_i和第二样本x_j相同位置特征的差值，即x_i-x_j，反映了第一样本x_i和第二样本x_j指定位置特征的差异性。

[0050] 由于第一样本x_i和第二样本x_j交换位置时，得到的两种目标样本$c_{i,j}$均是反映第一样本x_i和第二样本x_j的差异性，因此采用平方函数替代relu激活函数，则目标样本$c_{i,j}=G_{fea}(x_i,x_j)^2=(concat(x_i,x_j)\times W_G)^2$。

[0051] 在得到目标样本$c_{i,j}$之后，计算第一样本x_i和第二样本x_j的相似值$s_{i,j}$。其中，$s_{i,j}=sigmod(S(c_{i,j}))$，$S(c_{i,j})$表示目标样本$c_{i,j}$的相似性。此处选用最后神经元为1的全连接层，采用sigmod函数计算相似值。生成损失采用二分类损失函数计算，如式：$loss=(1-y_{i,j})\log(1-s_{i,j})+(y_{i,j})\log(s_{i,j})$。其中，$y_{i,j}$表示目标样本$c_{i,j}$的真实类别，$y_{i,j}$在第一样本$x_i$和第二样本$x_j$同类时取1，不同类时取1。

[0052] 在生成层的训练过程中，稀疏矩阵W_G的参数通过梯度$J(\theta)$和学习率a来更新，θ表示生成层的其他参数，更新过程如式：$\mathbf{W_G}=\mathbf{W_G}-\alpha\frac{\partial J(\theta)}{\partial \mathbf{W_G}}$。通过不断更新生成层的参数，目标样本$c_{i,j}$的$i_{th}$特征将反映第一样本$x_i$和第二样本$x_j$同位置特征的差异性和少部分其他特征的差异性。

[0053] 在选择层中，将目标样本$c_{i,j}$各个特征权重W_S初始化为0，后续通过对输入的分类损失进行求导得到的梯度来更新特征权重W_S。

[0054] 在对特征权重W_S进行初始化后，选择层的输出较小，参数学习较为缓慢，因此需要将特征权重W_S的学习率设置为较大值，其他参数的学习率设置为正常值，则输入到选择层的目标样本$c_{i,j}\times W_S$。

[0055] 经一定轮次的训练，选择层的特征权重W_S将在新的目标样本的训练下更新。但因特征权重W_S初始化为0，更新较为缓慢，目标样本$c_{i,j}$各个特征权重W_S均会较小，因此可设置较大的学习率，如式：$w_i=w_i-\alpha_s\frac{\partial J(\theta)}{\partial w_i}$。其中，$w_i$表示$W_S$的取值，$J(\theta)$表示分类损失，$a_s$表示设

置的学习率。

[0056] 通过对w_j取绝对值并进行排序，将前k个变化较大的特征权重W_S对应的特征作为样本特征。此时若要将选择层用于直接预测类别，可将选中的w_i设置为$sign(w_i)$，其他w_i设置为0，固定住w_i的值，然后进行简单的预训练。从而实现只使用从样本特征中选择的目标特征预测分类。

[0057] 选择层将目标样本$c_{i,j}$的所有特征权重W_S初始化为0，在训练过程中筛选变化较大的特征权重W_S，从而将认为的重要特征作为样本特征。

[0058] 在比较层中，通过选用全连接层，结合softmax判断输入的目标样本反映的相似性，可以表示为：$Similarity=sigmod(W_k\times(W_{k-1}\times\cdots(W_0\times c_{i,j}+b_0)+b_{k-1})+b_k)$。其中，$W_k$表示网络中的权重参数，$b_k$表示比较层中的bias值，sigmod函数为神经元激活函数，可使输出属于0和1之间。这些参数均可以使用后向传播来使全连接层学习出一个合适的网络以衡量目标样本$c_{i,j}$的相似性，其过程如式：$\mathbf{W}_k=\mathbf{W}_k-\alpha\frac{\partial J(\theta)}{\partial \mathbf{W}_k}$ 其中，θ表示比较层中的其他参数，$\partial J(\theta)$表示比较层的参数梯度，a表示学习率。

[0059] 通过由生成层生成较多可反映样本对差异性的新样本，以增加样本数量来缓解基因表达谱数据少样本特性带来的欠拟合问题。同时，通过由选择层根据新样本进行特征选择，以筛选重要特征来缓解基因表达谱数据高维度特性带来的过拟合问题。通过由比较层不断学习判断样本对相似性，以提供损失函数梯度更新生成层和选择层的参数。

[0060] 整个过程如式：

[0061] $C_{comp}(x_i,x_j)=C_{comp}(F(G_{fea}(x_i,x_j)))$

[0062] $=C_{comp}(concat(x_i,x_j)\times W_G\times W_S)$。

[0063] 其中，x_i和x_j为两个样本，通过水平连接后得到$concat(x_i,x_j)$，然后输入到生成层网络G_{fea}，生成新样本，再将新样本输入到选择层F。

[0064] 分类损失如式：

[0065] $$loss=\sum_{i=1}^{n}y_i\times log\bar{y}_i+(1-y_i)\times log(1-\bar{y}_i)$$

。

[0066] 其中，n为样本个数，y_i为样本的真实类别，$\bar{y}_i$为样本的预测类别。

[0067] 本实施例通过构建初始分类模型，在初始分类模型中引入生成层、选择层和比较层，并通过在基因表达谱训练数据集中随机选择两个样本，对生成层、选择层和比较层进行训练和更新，得到目标分类模型，使得可利用生成层，根据基因表达谱数据中任意两个样本生成新的样本，利用选择层，根据新的样本各个特征的权重选择若干个样本特征，利用比较层，根据从所有样本特征中选择的若干个目标特征对新的样本进行分类。

[0068] 本实施例能够构建一种适用于基因表达谱的分类模型，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。

[0069] 在优选的实施例当中，步骤S2在将第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本之后，还包括：根据目标样本，计算第一样本与第二样本的相似值，并

根据相似值，计算生成层的生成损失，将生成损失反向输入生成层，更新生成层的参数。

[0070] 本实施例通过计算第一样本与第二样本的相似值，进而计算生成层的生成损失，以将生成损失反向输入生成层，更新生成层的参数，使得生成层能够生成更能反映两个样本之间的差异性的目标样本，从而进一步提高基因表达谱的分类准确度。

[0071] 在优选的实施例当中，所述将分类损失反向输入初始分类模型，更新初始分类模型的参数，得到目标分类模型，具体为：将分类损失分别反向输入生成层和选择层，更新生成层和选择层的参数。

[0072] 本实施例通过将分类损失分别反向输入生成层和选择层，以更新生成层和选择层的参数，使得生成层能够生成更能反映两个样本之间的差异性的目标样本，选择层能够筛选出目标样本中更重要的特征，从而进一步提高基因表达谱的分类准确度。

[0073] 如图3所示，第二实施例提供一种适用于基因表达谱的分类模型构建装置，包括：初始分类模型构建模块21，用于构建初始分类模型，并从基因表达谱训练数据集中随机选择第一样本和第二样本；其中，初始分类模型包括生成层、选择层和比较层；生成层训练模块22，用于将第一样本与第二样本进行连接，得到第三样本，并将第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本；选择层训练模块23，用于将目标样本输入选择层，使选择层根据目标样本的特征权重，对目标样本进行特征选择，输出选择的若干个样本特征；比较层训练模块24，用于从所有样本特征中选择若干个目标特征，并将目标特征输入比较层，使比较层根据目标特征，对目标样本进行分类，输出得到的分类结果；目标分类模型获取模块25，用于将分类结果与真实结果进行比较，得到分类损失，并将分类损失反向输入初始分类模型，更新初始分类模型的参数，得到目标分类模型。

[0074] 在本实施例的一种优选的实施方式当中，生成层包括稀疏矩阵，选择层包括神经网络。

[0075] 基因表达谱数据特征的顺序取决于研究人员选择测试的基因顺序，没有确定的相对位置信息。本实施例采用稀疏矩阵作为生成层，能够抽取第一样本和第二样本特征的差异性生成目标样本，有利于避免引入多余的噪声。

[0076] 在通过初始分类模型构建模块21，构建初始分类模型后，首先通过生成层训练模块22，在对从基因表达谱训练数据集中随机选择的第一样本和第二样本进行连接后，将得到的第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本，然后通过选择层训练模块23，将目标样本输入选择层，使选择层根据目标样本的特征权重对目标样本进行特征选择，输出选择的若干个样本特征，接着通过比较层训练模块24，在从所有所述样本特征中选择若干个目标特征后，将目标特征输入所述比较层，使比较层根据目标特征对目标样本进行分类，输出得到的分类结果，最终通过目标分类模型获取模块25，将分类结果与真实结果进行比较，得到分类损失，并将分类损失反向输入初始分类模型以更新初始分类模型的参数，得到目标分类模型。

[0077] 例如，以一种癌症基因表达谱数据集X_n作为基因表达谱训练数据集。

[0078] 从癌症基因表达谱数据集X_n中随机选择第一样本x_i和第二样本x_j，并将第一样本x_i与第二样本x_j进行连接，得到第三样本concat(x_i,x_j)。其中，concat表示将第一样本x_i与第二样本x_j水平连接在一起，即直接将两个d维特征的样本连接成一个2d维特征的样本。

[0079] 在生成层中，通过从癌症基因表达谱数据集X_n中任意选择第一样本x_i和第二样本

x_j，使得可根据第三样本concat(x_i,x_j)来生成目标样本$c_{i,j}$，构建目标样本数据集C＝{$c_{1,1}$，$c_{1,2}$，…，$c_{n,n}$}，从而将样本容量为n的癌症基因表达谱数据集X_n扩充为样本容量为n_2的目标样本数据集C。其中，$C \subset R^{n \times d}$，$c_{i,j}=G_{fea}(x_i,x_j)=concat(x_i,x_j)\times W_G$，$G_{fea}(x_i,x_j)$表示生成层根据第一样本$x_i$和第二样本$x_j$生成目标样本$c_{i,j}$的函数，$W_G$表示稀疏矩阵，稀疏矩阵$W_G$初始化为$W_{i,i}=1$，$W_{i,i+j}=-1$，i，j∈[0,n]，其他参数初始化为$W_{i,j}=0$。

[0080] 在对稀疏矩阵进行初始化后，生成层在开始阶段，产生的目标样本$c_{i,j}$的i_{th}特征表示为第一样本x_i和第二样本x_j相同位置特征的差值，即x_i-x_j，反映了第一样本x_i和第二样本x_j指定位置特征的差异性。

[0081] 由于第一样本x_i和第二样本x_j交换位置时，得到的两种目标样本$c_{i,j}$均是反映第一样本x_i和第二样本x_j的差异性，因此采用平方函数替代relu激活函数，则目标样本$c_{i,j}=G_{fea}(x_i,x_j)^2=(concat(x_i,x_j)\times W_G)^2$。

[0082] 在得到目标样本$c_{i,j}$之后，计算第一样本x_i和第二样本x_j的相似值$s_{i,j}$。其中，$s_{i,j}$=sigmod(S($c_{i,j}$))，S($c_{i,j}$)表示目标样本$c_{i,j}$的相似性。此处选用最后神经元为1的全连接层，采用sigmod函数计算相似值。生成损失采用二分类损失函数计算，如式：$loss=(1-y_{i,j})\log(1-s_{i,j})+(y_{i,j})\log(s_{i,j})$。其中，$y_{i,j}$表示目标样本$c_{i,j}$的真实类别，$y_{i,j}$在第一样本$x_i$和第二样本$x_j$同类时取1，不同类时取1。

[0083] 在生成层的训练过程中，稀疏矩阵W_G的参数通过梯度J(θ)和学习率a来更新，θ表示生成层的其他参数，更新过程如式：$\mathbf{W_G}=\mathbf{W_G}-\alpha\frac{\partial J(\theta)}{\partial \mathbf{W_G}}$。通过不断更新生成层的参数，目标样本$c_{i,j}$的$i_{th}$特征将反映第一样本$x_i$和第二样本$x_j$同位置特征的差异性和少部分其他特征的差异性。

[0084] 在选择层中，将目标样本$c_{i,j}$各个特征权重W_S初始化为0，后续通过对输入的分类损失进行求导得到的梯度来更新特征权重W_S。

[0085] 在对特征权重W_S进行初始化后，选择层的输出较小，参数学习较为缓慢，因此需要将特征权重W_S的学习率设置为较大值，其他参数的学习率设置为正常值，则输入到选择层的目标样本$c_{i,j}\times W_S$。

[0086] 经一定轮次的训练，选择层的特征权重W_S将在新的目标样本的训练下更新。但因特征权重W_S初始化为0，更新较为缓慢，目标样本$c_{i,j}$各个特征权重W_S均会较小，因此可设置较大的学习率，如式：$w_i=w_i-\alpha_S\frac{\partial J(\theta)}{\partial w_i}$。其中，$w_i$表示$W_S$的取值，J(θ)表示分类损失，$a_s$表示设置的学习率。

[0087] 通过对w_i取绝对值并进行排序，将前k个变化较大的特征权重W_S对应的特征作为样本特征。此时若要将选择层用于直接预测类别，可将选中的w_i设置为sign(w_i)，其他w_i设置为0，固定住w_i的值，然后进行简单的预训练。从而实现只使用从样本特征中选择的目标特征预测分类。

[0088] 选择层将目标样本$c_{i,j}$的所有特征权重W_S初始化为0，在训练过程中筛选变化较大的特征权重W_S，从而将认为的重要特征作为样本特征。

[0089] 在比较层中，通过选用全连接层，结合softmax判断输入的目标样本反映的相似

性，可以表示为：$Similarity=sigmod(W_k\times(W_{k-1}\times\cdots(W_0\times c_{i,j}+b_0)+b_{k-1})+b_k)$。其中，$W_k$表示网络中的权重参数，$b_k$表示比较层中的bias值，sigmod函数为神经元激活函数，可使输出属于0和1之间。这些参数均可以使用后向传播来使全连接层学习出一个合适的网络以衡量目标样本$c_{i,j}$的相似性，其过程如式：$\mathbf{W}_k=\mathbf{W}_k-\alpha\frac{\partial J(\theta)}{\partial \mathbf{W}_k}$ 其中，θ表示比较层中的其他参数，$\partial J(\theta)$表示比较层的参数梯度，a表示学习率。

[0090] 通过由生成层生成较多可反映样本对差异性的新样本，以增加样本数量来缓解基因表达谱数据少样本特性带来的欠拟合问题。同时，通过由选择层根据新样本进行特征选择，以筛选重要特征来缓解基因表达谱数据高维度特性带来的过拟合问题。通过由比较层不断学习判断样本对相似性，以提供损失函数梯度更新生成层、选择层和比较层的参数。

[0091] 整个过程如式：

[0092] $C_{comp}(x_i,x_j)=C_{comp}(F(G_{fea}(x_i,x_j)))$

[0093] $=C_{comp}(concat(x_i,x_j)\times W_G\times W_S)$。

[0094] 其中，x_i和x_j为两个样本，通过水平连接后得到$concat(x_i,x_j)$，然后输入到生成层网络G_{fea}，生成新样本，再将新样本输入到选择层F。

[0095] 分类损失如式：

[0096] $$loss=\sum_{i=1}^{n}y_i\times log\bar{y}_i+(1-y_i)\times log(1-\bar{y}_i)$$

。

[0097] 其中，n为样本个数，y_i为样本的真实类别，$\bar{y}_i$为样本的预测类别。

[0098] 本实施例通过构建初始分类模型，在初始分类模型中引入生成层、选择层和比较层，并通过在基因表达谱训练数据集中随机选择两个样本，对生成层、选择层和比较层进行训练和更新，得到目标分类模型，使得可利用生成层，根据基因表达谱数据中任意两个样本生成新的样本，利用选择层，根据新的样本各个特征的权重选择若干个样本特征，利用比较层，根据从所有样本特征中选择的若干个目标特征对新的样本进行分类。

[0099] 本实施例能够构建一种适用于基因表达谱的分类模型，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。

[0100] 在优选的实施例当中，生成层训练模块22，还用于在将第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本之后，根据目标样本，计算第一样本与第二样本的相似值，并根据相似值，计算生成层的生成损失，将生成损失反向输入生成层，更新生成层的参数。

[0101] 本实施例通过生成层训练模块22，计算第一样本与第二样本的相似值，进而计算生成层的生成损失，以将生成损失反向输入生成层，更新生成层的参数，使得生成层能够生成更能反映两个样本之间的差异性的目标样本，从而进一步提高基因表达谱的分类准确度。

[0102] 在优选的实施例当中，所述将分类损失反向输入初始分类模型，更新初始分类模型的参数，得到目标分类模型，具体为：将分类损失分别反向输入生成层和选择层，更新生成层和选择层的参数。

[0103] 本实施例通过将分类损失分别反向输入生成层和选择层，以更新生成层和选择层的参数，使得生成层能够生成更能反映两个样本之间的差异性的目标样本，选择层能够筛选出目标样本中更重要的特征，从而进一步提高基因表达谱的分类准确度。

[0104] 第三实施例提供一种适用于基因表达谱的分类模型，所述分类模型是应用如第一实施例所述的适用于基因表达谱的分类模型构建方法而获得，且能达到与之相同的有益效果。

[0105] 第四实施例提供一种适用于基因表达谱的分类方法，包括：从获取的基因表达谱数据集中选择第四样本和第五样本，并将第四样本和第五样本输入如第三实施例所述的适用于基因表达谱的分类模型，得到分类结果。

[0106] 本实施例利用第三实施例所述的适用于基因表达谱的分类模型，对基因表达谱数据进行分类，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。

[0107] 综上所述，实施本发明的实施例，具有如下有益效果：

[0108] 通过构建初始分类模型，在初始分类模型中引入生成层、选择层和比较层，并通过在基因表达谱训练数据集中随机选择两个样本，对生成层、选择层和比较层进行训练和更新，得到目标分类模型，使得可利用生成层，根据基因表达谱数据中任意两个样本生成新的样本，利用选择层，根据新的样本各个特征的权重选择若干个样本特征，利用比较层，根据从所有样本特征中选择的若干个目标特征对新的样本进行分类。本实施例能够构建一种适用于基因表达谱的分类模型，实现增加基因表达谱数据的样本数量，缓解少样本特性带来的欠拟合问题，从而进一步提高基因表达谱数据的分类准确度。

[0109] 以上所述是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也视为本发明的保护范围。

[0110] 本领域普通技术人员可以理解实现上述实施例中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory，ROM)或随机存储记忆体(Random Access Memory，RAM)等。

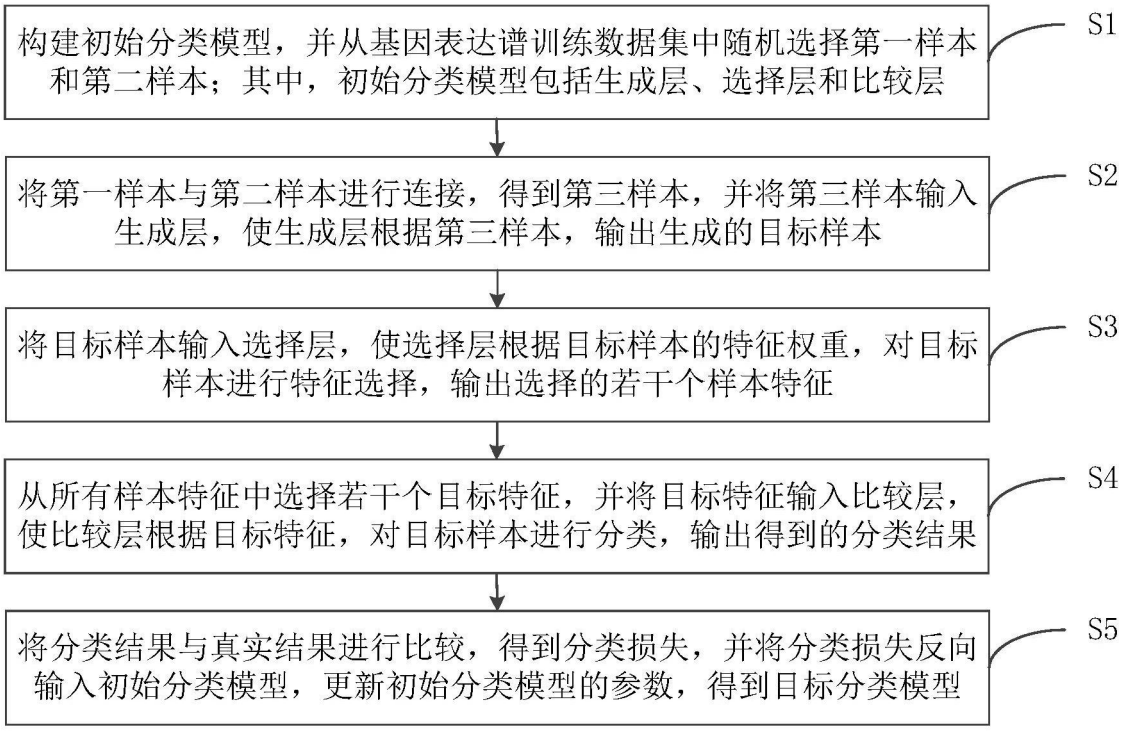

构建初始分类模型，并从基因表达谱训练数据集中随机选择第一样本和第二样本；其中，初始分类模型包括生成层、选择层和比较层
S1
将第一样本与第二样本进行连接，得到第三样本，并将第三样本输入生成层，使生成层根据第三样本，输出生成的目标样本
S2
将目标样本输入选择层，使选择层根据目标样本的特征权重，对目标样本进行特征选择，输出选择的若干个样本特征
S3
从所有样本特征中选择若干个目标特征，并将目标特征输入比较层，使比较层根据目标特征，对目标样本进行分类，输出得到的分类结果
S4
将分类结果与真实结果进行比较，得到分类损失，并将分类损失反向输入初始分类模型，更新初始分类模型的参数，得到目标分类模型
S5

图1

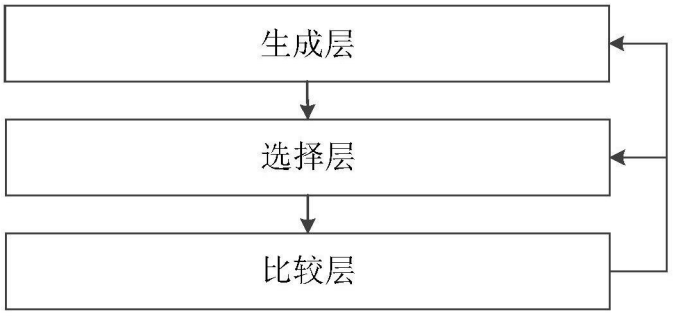

生成层
选择层
比较层

图2

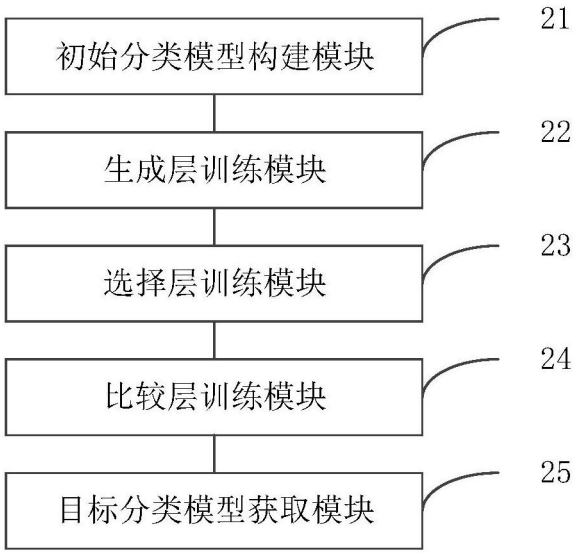

初始分类模型构建模块
21
生成层训练模块
22
选择层训练模块
23
比较层训练模块
24
目标分类模型获取模块
25

图3